Probabilistic forecast model for the 0.10, 0.50, and 0.90 quantiles of seasonal water supply

Mikhail Sarafanov (Dreamlone) mik.sarafanov@gmail.com
Espoo, Finland

Abstract

The report discusses a quantile regression algorithm that uses The Snow Telemetry (SNOTEL) network data, Climate teleconnection indices and Palmer Drought Severity Index (PDSI) to predict seasonal water supply. In the process of experiments with training the most efficient model at different stages, the following data were included in various versions of the model: USGS streamflow data, SNOTEL, SNOw Data Assimilation System (SNODAS) data, PDSI, Climate teleconnection indices. It is proposed to use the algorithm of meteorological parameters aggregation for n days as the basis of the method. Approaches to iterative improvement of the model are considered, as well as applied ways to optimize hyperparameters.

Keywords: hydrology, forecasting, quantile regression, water supply forecast rodeo

1. Introduction

Hydrology is an essential science which is supporting human activity. Hydrological forecasts are used in agriculture [1], construction [2], etc. to prevent material and human losses and to improve operational processes. To improve the prediction of hydrological processes, machine learning methods have been actively used in recent decades [3]. However, classical probabilistic and regression models show satisfactory performance as well [4].

In this study, different methods and data sources were used for probabilistic forecast model for the 0.10, 0.50, and 0.90 quantiles of seasonal water supply at

26 different hydrologic sites in the USA to find out which modeling method is the most effective.

2. Technical Approach

The material for current report, as well as the source code, is available in the WASU repository via link: `https://github.com/Dreamlone/wasu`.

The algorithm is designed to generate individual models for each site. The visualization of the basins of the 26 sites is shown in Figure 1.

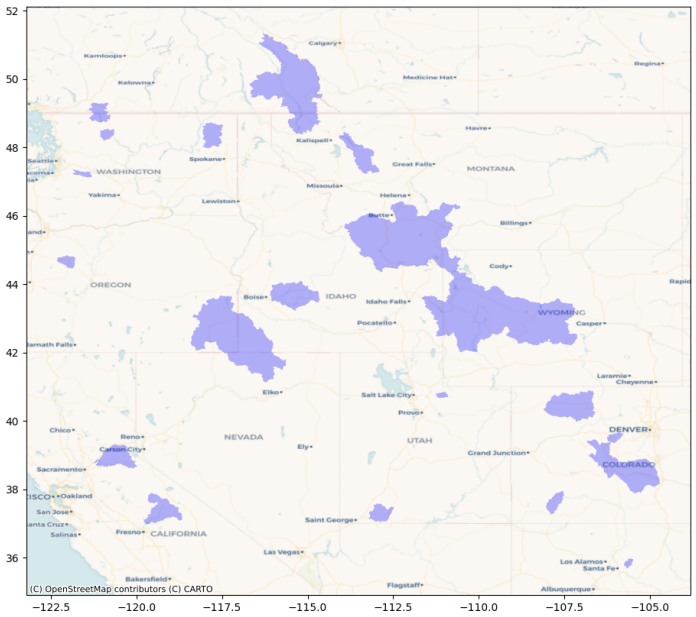

Figure 1: Spatial polygons for river basins

Details of the modelling quality assessment process are given in section 2.4.

Table 1: Validation metrics for simple and advanced repeating algorithms. Baseline approach for forecasting

Metric name	**Simple repeating**	**Advanced repeating**
MAE	396.65	386.02
MAPE	56.50	59.98
Symmetric MAPE	61.37	52.97
Quantile loss (0.5 quantile)	396.65	386.02
Average mean Quantile loss	367.66	275.82

2.1. Algorithm and Architecture Selection

The first algorithm that was used to generate the predictions: simple repeating. For provided above validation years the algorithm takes the value from 2015 and assigns it to each subsequent year (Figure 2). Upper and lower bounds of the forecast were calculated using a buffer of 10 per cent of the predicted value.

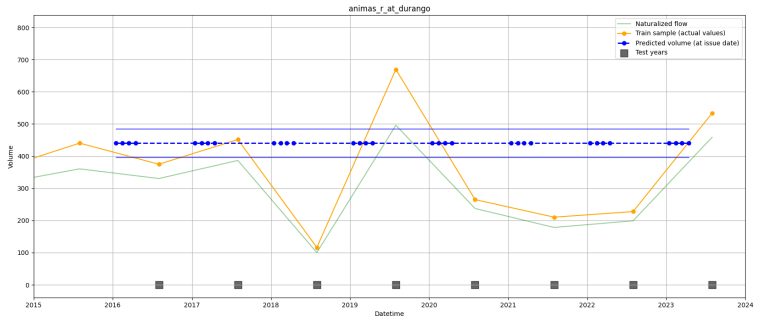

Figure 2: Forecasts for tests years for site 'animas_r_at_durango' using simple repeating since 2015 (forecast values (blue) are given as of the issue date)

More advanced approach for repeating predictions uses values from the previous year for this site to be used as a forecast. That is, for 2005 the year 2004 will be used, for 2007 the year 2006 will be used, etc (Figure 3).

The metrics for the repeating algorithms are shown in Table 1.

Figure 3: Forecasts for tests years for site 'animas_r_at_durango' using advanced repeating repeating since 2015 (forecast values (blue) are given as of the issue date)

Further modifications of the algorithm were made in the following order:

- USGS streamflow-based - uses flow values aggregated over a specific period (for example 40, 80 or 120 days before the forecast issue date) to generate a forecast into the future;
- SNOTEL-based predictions - aggregation of statistics on precipitation, average daily temperature, max temperature, min temperature, snow water equivalent for a given period of time preceding the issue date is used;
- Ensembling of previous predictions - combination of forecasts from SNOTEL model and USGS streamflow model with predictions averaging;
- Ensembling of previous predictions (with smoothing) - previous ensemble with moving average;
- SNODAS-based predictions - aggregation of spatio-temporal statistics of SNODAS data;
- Complex model ver 1 - uses features from SNOTEL, PDSI and Climate indices data sources and lags for prediction - **Model on Hindcast Stage**;
- Complex model ver 2 - previous version with excluded Climate indices;

- Complex model (optimized) - previous version with optimized hyperparameters (aggregation lags) - **Model on Forecast Stage**.

The Quantile linear regression algorithm implemented through the scikit-learn [5] package was used as the core of each of the models because it's proven to be effective in hydrology [6]. Nonlinear models such as LightGBM were also tried during the experiments, but linear quantile regression proved to be more robust. Details of feature preprocessing for above mentioned models are given in section 2.2.

Complex model ver 1 aggregate PDSI data with lag 180 days, SNOTEL data with lag 110 days and Climate indices with offset 150 days for all sites.

The difference between Complex model ver 2 and Complex model (optimized) is that the data sources for the optimized one the experiments were conducted to find optimal hyperparameters (aggregation lags) for each site - see Figure 4. Thus, for each site the optimal set of lags for each data source was found.

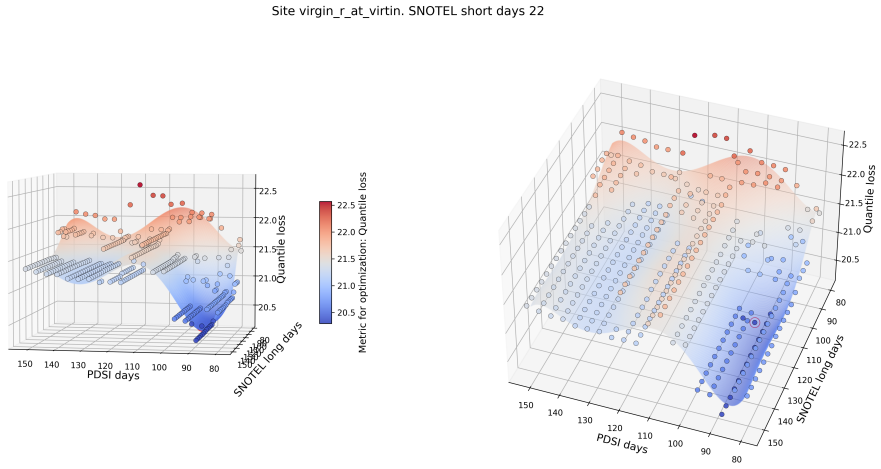

Figure 4: Exploration of Average mean Quantile loss landscape for common model for "virgin r at virtin" site with constant days SNOTEL short=22 parameter. Optimal configuration for this site: days SNOTEL short=22, days SNOTEL long=108, days PDSI=92

The values of the "Average mean Quantile loss" metric on the validation sample for all the above models are shown in Figure 5.

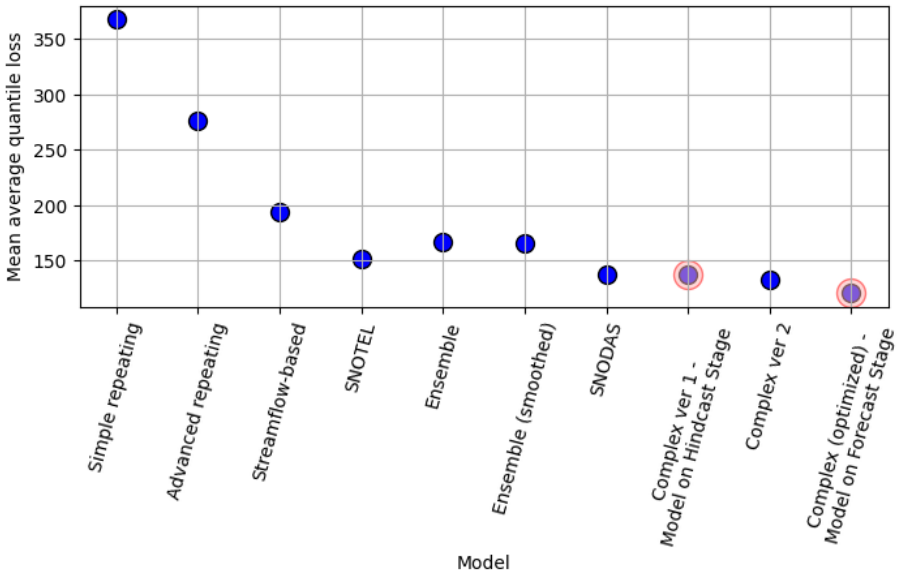

Figure 5: Average mean Quantile loss for implemented models on validation sample

2.2. Data Sources and Feature Engineering

This section discusses the concept of data preparation with aggregation by selected days using SNOTEL data as an example (Figure 6).

This approach was decided to be used because it proved to be more effective than other classical methods of predicting hydrological characteristics [3]. Moreover, it is data source-agnostic, scalable and and highly customizable.

2.3. Uncertainty Quantification

The quantile linear regression approach [7] was applied to make predictions for the 0.10 and 0.90 quantiles.

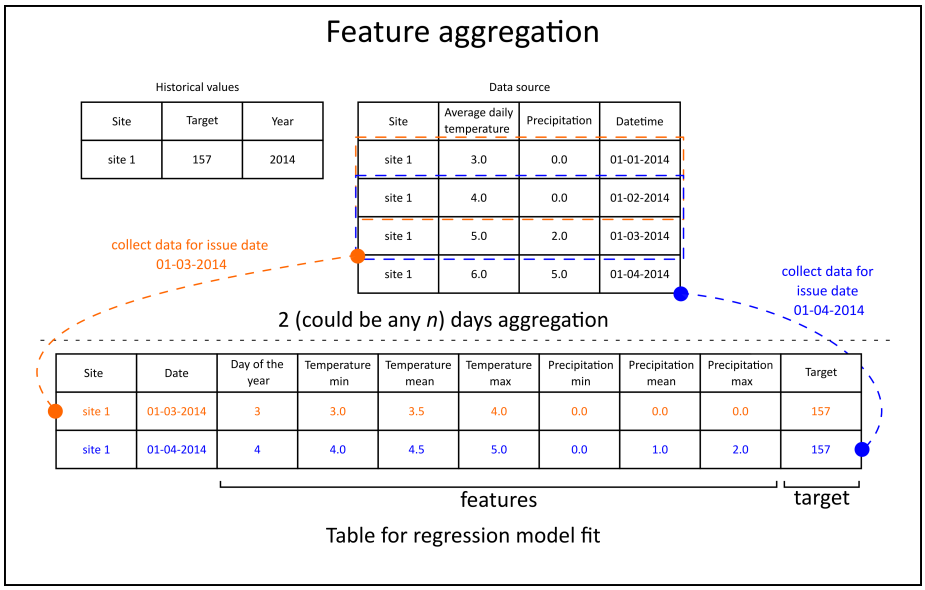

Figure 6: Example of generating features for a model using aggregation with defined lag of 2 days

2.4. Training and Evaluation Process

The model was trained on all available data excluding validation sample. The best model based on the results of experiments was fitted on the entire data set. The modelling quality was assessed on actually known data for the years 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023 [1] using the following metrics:

- Mean absolute error (MAE) for 0.5 quantile predictions;
- Mean absolute percentage error (MAPE) for 0.5 quantile predictions;
- Symmetric Mean absolute percentage error metric for 0.5 quantile predictions;
- Quantile loss metric for 0.5 quantile predictions;

[1] the validation on the given years was done after the organisers provided the data for the declared dates

- Average mean Quantile loss metric.

The calculation of the last two metrics was implemented according to the equation given by the organiser of the competition. [2]

3. Conclusion

As a result of these experiments, an algorithm for identifying predictive models for each of the 26 sites was prepared. The Mean average quantile loss metric was improved by a factor of three relative to baseline.

4. Machine Specifications

The models presented in this report are not computationally intensive. The fitting time for complex model was 4 minutes for all 26 sites including data preparation on a laptop (16GB RAM, AMD Ryzen 5 5000 series). GPU is not required.

References

[1] D. H. Wilber, R. E. Tighe, L. J. O'Neil, Associations between changes in agriculture and hydrology in the cache river basin, arkansas, usa, Wetlands 16 (1996) 366–378.

[2] J. McEnery, J. Ingram, Q. Duan, T. Adams, L. Anderson, Noaa's advanced hydrologic prediction service: building pathways for better science in water forecasting, Bulletin of the American Meteorological Society 86 (3) (2005) 375–386.

[2] https://www.drivendata.org/competitions/257/reclamation-water-supply-forecast-hindcast/page/807/#code-submission-performance-metric

[3] M. Sarafanov, Y. Borisova, M. Maslyaev, I. Revin, G. Maximov, N. O. Nikitin, Short-term river flood forecasting using composite models and automated machine learning: The case study of lena river, Water 13 (24) (2021) 3482.

[4] J. Hardy, J. J. Gourley, P.-E. Kirstetter, Y. Hong, F. Kong, Z. L. Flamig, A method for probabilistic flash flood forecasting, Journal of Hydrology 541 (2016) 480–494.

[5] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[6] A. Weerts, H. Winsemius, J. Verkade, Estimation of predictive hydrological uncertainty using quantile regression: examples from the national flood forecasting system (england and wales), Hydrology and Earth System Sciences 15 (1) (2011) 255–265.

[7] R. Koenker, V. Chernozhukov, X. He, L. Peng, Handbook of quantile regression.