Water supply forecast combining LightGBM model with distribution estimation approach

Abstract

Water supply forecasts require an effective and reliable tool. Hence, machine learning algorithm seems to be the most promising solution to help with tackling the problem. LightGBM model is proposed as a main predictor for seasonal water supply. In response to the fact that forecasts have to be issued in various stages of the water year, creating different models for each month is introduced, to account for different data available for different months. As it additionally requires forecasts for different quantiles, distribution estimates from data are added to achieve more robust results, as they act as a smoothing effect to LightGBM models. Due to the fact that there is not much data available for given water reservoir's volumes, the solution intends to not be too complicated, using no more than 10 explainable features for month, while still achieving decent performance.

1 Technical Approach

1.1 Algorithm and Architecture Selection

The solution includes 2 main methods to make predictions – LightGBM models and quantiles estimation from historical data distribution for different sites. Predictions are made separately for each month. Issue dates include 7 months of predictions, so 7 different prediction blocks are created for each of those months. It is created this way to take into account new information obtained through different months. For comparison, models from June can use data from already known monthly naturalized flow from some previous months of this water year that are a part of the water volume to predict but for January, other features must be taken into consideration, as such data isn't yet available for that period.

Figure 1 presents a schema of a building block for one month for the first 4 months of predictions (Jan-Apr). For calculating Q0.5 (quantile 0.5) predictions, only LightGBM model is used, whereas for Q0.1 and Q0.9, a weighed average of LightGBM model result and estimation from distribution data are combined together.

Then, clipping methods are introduced to adjust obtained values. Min/max clipping is used to clip values below/over minimum/maximum value of given site id's volume. If predictions exceed those values, they are replaced with minimum/maximum values. Historical data is used as a range of possible values that results could take to make them more robust.

Another clipping technique is used to deal with quantile crossing problem. It could occur that lower quantile is given higher value than higher quantile, though it isn't logical (Schmidt & Zhu, 2016). To deal with this issue, the results are evened out. If Q0.1 is greater than Q0.5, its value is changed to the Q0.5 value. If Q0.9 prediction is less than Q0.5, it gets replaced with Q0.5 volume. There aren't any changes made in this regard to Q0.5 predictions, they stay the same.

Finally, Q0.1 and Q0.9 results are calculated as a weighted average of LightGBM and distribution estimates from previous steps. Each month uses its own distribution estimation weight. Thanks to that, more weight could be given to distribution estimates for early months, when LightGBM couldn't be fully trusted, as it bases on early predictors of future water flow volumes.

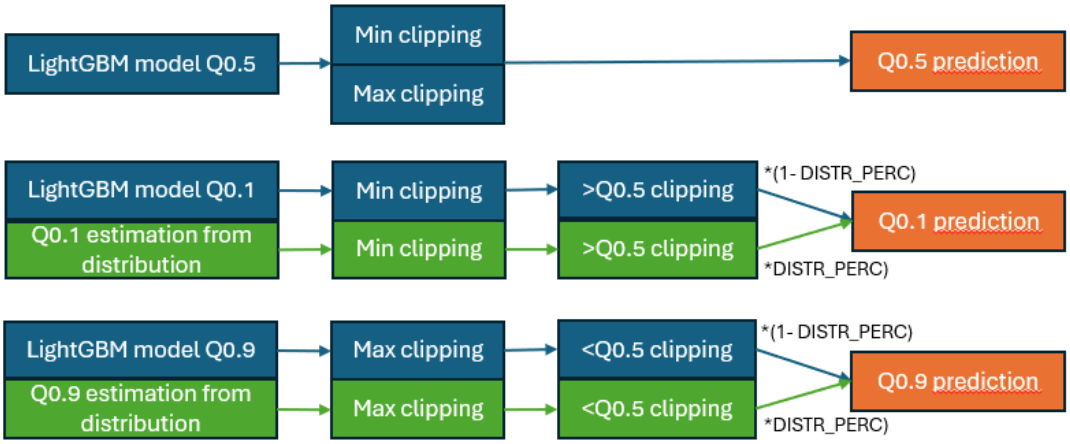

Figure 1: Single block for the first 4 months (January-April)

Figure 2 presents a second type of model block that is intended for later months (May - July). It is built upon first type. For each month, there are created 2 separate not overlapping LightGBM models: one for 23 site ids that contains data from previous month's naturalized water flow from this water year and the second model for the remaining 3 site ids without information on previous water flow volume (american river folsom lake, san_joaquin_river_millerton_reservoir, merced_river_yosemite_at_pohono_bridge). The second model uses data from all 26 site ids to learn patterns from all available data but only 3 mentioned site ids are evaluated.

Response variable used in the competition is the volume of naturalized flow between April and July (for most sites). However, for the first type of model, volume residuals are used as labels (with past naturalized flow added at the end of the prediction pipeline), as naturalized water flow from at least 1 month is already known and it is easier to predict the remaining volume. Additionally, naturalized flow clipping is added, to avoid a situation when, after averaging over LightGBM model results and distribution estimates, the predicted volume is less than already known naturalized flow from previous months. Moreover, multipliers and thresholds are introduced. Threshold indicates what data will be affected by nat flow clipping, Only predictions that are less than prediction * threshold will be corrected. Multipliers indicate by what amount predictions will be multiplied. The values weren't excessively optimized to not overfit to the data. The logic for parameters choosing was that if results are less than naturalized flow from April, it should have a higher multiplier than cumulative values between April and June. Also, quantile 0.9 is corrected more than quantile 0.1. The chosen parameters are presented in Table 1.

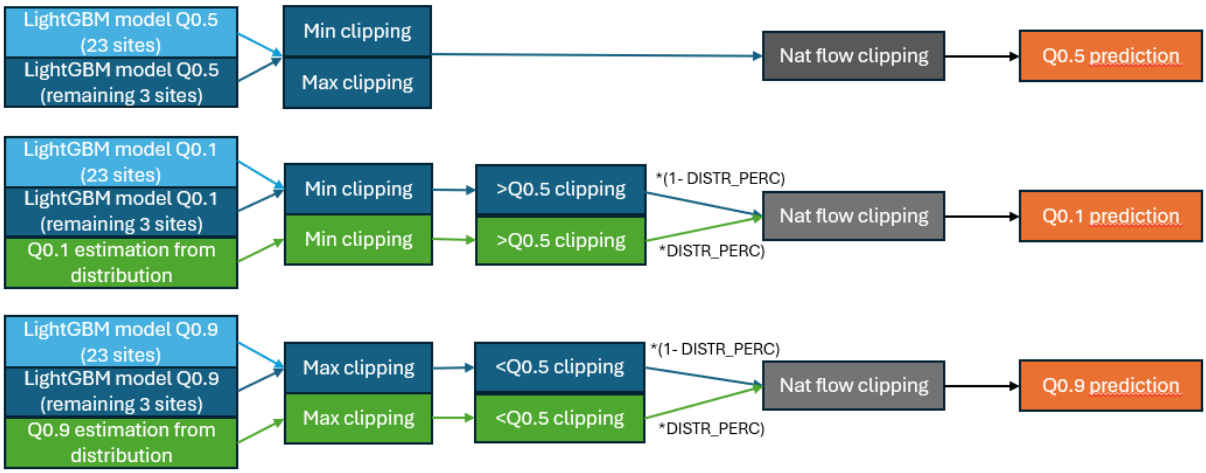

Figure 2: Single block for the last 3 months (May-July)

Month	May	June	July
Q0.1	1.3/1.2	1.2/1.1	1.1/1.0
Q0.5	1.35/1.25	1.25/1.15	1.15/1.05
Q0.9	1.4/1.25	1.3/1.15	1.2/1.05

Table 1: Multipliers/thresholds for known naturalized flow values clipping

1.1.1 LightGBM models

LightGBM model is used as the main predictor. Tree-based ensemble algorithms are regarded as one of the most powerful tools for tabular data prediction. LightGBM is established as one of the tools of choice for winning solutions in machine learning competitions (Makridakis et al., 2022). It is a very convenient framework for a feature engineering and data-centric approach to quickly validate new ideas by just adding a feature and examining change in performance. It also doesn't require data standardization as a tree-based algorithm. LightGBM additionally handles missing values, so there's no need to process them thoroughly.

LightGBM has a quantile loss function. Unfortunately, it doesn't have a method for using all desired quantiles at once. Instead of that, 3 different models must be built for different quantiles (0.1, 0.5 and 0.9). The models are created for each of 7 months. In total, there are 30 different models (12 models for Jan-Apr and 18 models for May-Jul). Though LightGBM results are decent, there are cases where Q0.1 and Q0.9 are very similar to Q0.5, resulting in not smooth results.

1.1.2 Distribution estimate

To address the issue of Q0.1 and Q0.9 miscalculation, distribution estimate is used. For each site id, 97 distributions are fitted to historical volume for available years. So many distributions were used to maximally optimize the results. When distributions are fitted, they are additionally examined in case that the best distribution doesn't seem to represent the data well. 9 next best fit distributions for site id were taken into consideration in such case and one of them was manually chosen as a replacement. The same amendments were chosen for each LOOCV year.

Having fitted distributions for each site id, the next step was to calculate Q0.1 and Q0.9, provided LightGBM model calculation for Q0.5. Simply getting Q0.1 and Q0.9 from fitted distributions yielded poor results. The new idea was to use a conditional estimation, treating Q0.5 forecast from LightGBM model as our data knowledge and making an approximation of Q0.1 and Q0.5 in a new distribution that is a part of the original distribution.

1.1.3 Other considerations

1.1.3.1 Architecture changes

Using one model for all months – it was rejected due to the assumption that it is beneficial to create different features for different months, as better quality data is available for later months.

Using different models for all issue dates based on month and day - it didn't give a significant boost to performance. It also would require much more models to create (4 times more).

Blending different machine learning models – it is a common practice to use different models and average their results. It usually boosts overall forecasting performance (Gao & Balyan, 2022). However, it reduces transparency and the objective of the proposed solution was to keep the models simple. The final solution includes only 2 factors for calculating result for a given site id-issue date-quantile combination - a suitable LightGBM model and distribution estimate.

1.1.3.2 Possible simplifications

Use only LightGBM models – that would simplify model structure by using only one method of prediction. On the other hand, the predictions would have lost their smoothness.

Use the same LightGBM architecture for all months – to reduce a number of created LightGBM models, only the standard architecture for all 26 site ids could be used. It would however decrease May-Jul performance.

Use only more recent data – it would require careful examination of performance downgrade. There could be some patterns that happen only every few decades and they would be missed.

1.2 Data Sources and Feature Engineering

1.2.1 Data sources

Table 2 shows data sources used in the proposed solution with explanation why the datasets were used and their time delay (days of delay for merging the dataset with the main training set).

Data source	About the database	Delay time
Train (from Data download)	Source of data labels.	Not applicable
Antecedent monthly naturalized flow (from Data download)	Naturalized flow, part of the label volume. It contains monthly data up to June, Due to the fact that it contains part of the final water flow volume, it is especially helpful for May, June and July issue dates.	1 day
USGS streamflow	Contains information on actual observed flow of water. It isn't exactly the same as the label volume but could still be helpful to estimate general flow of water, especially as it is published daily.	1 day
NRCS SNOTEL	Contains daily environmental features, including Snow Water Equivalent and accumulated precipitation. Results of different stations associated with a given site id were averaged to get features from this dataset.	1 day
PDSI (Palmer Drought Severity Index (PDSI) from gridMET)	Contains gridded index of drought, indicating "severity of the departure from normal conditions (National Integrated Drought Information System, U. S., n.d.). Data belonging to a site id was averaged, only pixels whose center is within the polygon was used.	5 days (contains latest data from 5 days before, so 10 days in total)
ERA5-Land monthly average	Gridded monthly reanalysis dataset with "a consistent view of the evolution of land variables over several decades" (Muñoz Sabater, 2019), containing various atmospheric variables. It provides ERA5-T-Land with 5 days of delay which is then corrected with ERA5-Land data within 2-3 months. However, historical data was treated as if it was ERA5-Land-T data. Such an assumption could have been made, as "in the event that serious flaws are detected in ERA5-Land-T, the latter could be different to the final consolidated ERA5-Land data [...] and the expectation is that the latter occur only on rare occasions." (Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2014), so not many changes between versions are expected.	5 days after the end of he month (ERA5-Land-T lag was adopted)

	Data from different grids from a site id was averaged, using pixels whose center is within the polygon.	
Seasonal meteorological forecasts from Copernicus	Contains forecasts for most features from ERA5-Land. ECMWF data from System 51 was used. For this system, only forecasts issued on the 1st day of the month are available. Forecasts from the next 90, 120, 150 and 180 days (for January issue date) were downloaded (similarly for different issue months for forecasts for the same months as for January). Data from different grids belonging to a site id was averaged, all pixels touched by geometries were used due to lower granularity (data is available with step 1 for coordinates, whereas it's 0.1 for ERA5-Land). 5 requests for Dec, Jan, Feb, Mar, Apr months were made.	Published on the 6th day of the month (for used ECMWF source)
Metadata (from Data download)	Used to locate positions of different sites.	Not applicable
submission_format (from Data download)	Used to derive submission format, including issue dates.	Not applicable

Table 2: Data sources used in the processing

1.2.2 Features

The assumption for the solution was to use no more than 10 features per month. It ensures that LightGBM models aren't too complicated, don't overfit to too many features, with a limited number of rows and are easily interpretable. Separate models are created for different months, so different features could be used for each month. Thus, variables that improve accuracy only for early months, having too weak predictive power for later months, were added only for the early months. On the other hand, there are also features available only for the later months. Features based solely on the same water year are used. They were carefully prepared to don't use future data not available at the time of prediction. Only features that significantly boost model performance are used (i.e. features that increased performance by 0.1-0.2 KAF weren't used). The features are listed in Table 3. Months marked with * stand for models for 3 site ids.

Feature	**Data source**	**Description**	**Months used**
site_id	Train	Site id. Used in all months, as volume is different sits can vary greatly. It is the only categorical variable.	All
issue_date_no_year	Submission format	Month-day combination from issue date. Acts as a clarification about more exact issue time. It is used in numeric format, so it is easy for LightGBM to create a rule for 2 or 3 neighbouring issue dates.	All
nat_flow_prev	Monthly naturalized flow	It gives information on naturalized flow from previous month.	3, 7
nat_flow_Apr_mean	Monthly naturalized flow	Gives more specific information about already known volume, an averaged value since April.	7
WTEQ_DAILY_prev	NRCS SNOTEL	Most up-to-date value of Snow water equivalent. It is a very powerful variable, as it informs about how much	1, 2, 3, 4, 5, 6;

		water could be released after snowpack melting. USDA Climate Hubs (n.d.) indicates that "Between 60 and 70% of water supplies come from snowmelt".	6*
WTEQ_DAILY_Apr_mean	NRCS SNOTEL	Snow Water Equivalent daily average since April. It could help estimate how much snowpack was left during volume predictions period.	5, 7; 5*, 6*, 7*
WTEQ_DAILY_Jun_prev_diff	NRCS SNOTEL	Snow Water Equivalent difference between previous value and the value from May 30. It could be used as a current month progress in snow melting.	6; 6*
WTEQ_DAILY_Jul_prev_diff	NRCS SNOTEL	Snow Water Equivalent difference between previous value and the value from June 29.	7
PREC_DAILY_Apr_mean	NRCS SNOTEL	Daily average since April of precipitation accumulation from this water year.	5, 5*
PREC_DAILY_Apr_prev_diff	NRCS SNOTEL	Difference between latest available precipitation accumulation value and the value from Mar 30. It could be interpreted as precipitation since the beginning of forecasting period.	5, 6
discharge_cfs_mean_Apr_mean	USGS streamflow	Average actual observed flow of water since April. It can be perceived as an up-to-date complement to naturalized flow data that is only added monthly, whereas streamflow is updated daily.	7*
discharge_cfs_mean_since_Oct_std	USGS streamflow	Standard deviation of actual observed flow of water. Acts as an indicator of fluctuations in water flow, a measure of variability in the model that uses mostly features based on averages. It uses data from the whole water year up to a day before issue date.	2, 3, 4, 5, 6; 5*, 6*
pdsi_prev	PDSI	Latest available value of PDSI index.	1, 2, 4, 5, 6, 7; 5*, 6*,7*
pdsi_prev_to_last_month_diff	PDSI	Difference between latest available PDSI value and the value 30 days before. Could be interpreted as last month's change in soil wetness.	4; 7*
sd_prev	ERA5-Land	Latest available value of monthly snow depth water equivalent.	5, 6, 7; 5*, 6*
sd_forecasts	Copernicus forecasts	Averaged monthly forecasts over snow depth water equivalent since end of March up to the end of May. The feature is used only for early months, before snow depth values from later months are known.	1, 2
sd_forecasts_with_jun	Copernicus forecasts	Averaged monthly forecasts over snow depth water equivalent since end of March up to end of June. June predictions seem to be of higher quality since March.	3, 4
longitude	Metadata	Longitude serves as and indicator of site's location. It seems enough as a more general site location, as using latitude didn't help with model performance.	2, 3, 4, 5, 6, 7; 5*, 6*,7*

Table 3: Features used to train models

1.2.3 Other considered datasets

Most of the other explored data sources didn't improve the results enough, so they weren't taken into account, to not make the models too complicated. Some of the considered datasets are:

ERA5-Land hourly data from 1950 to present (Muñoz Sabater, J. (2019)) – hourly Copernicus data didn't bring much more information, compared to monthly data used in the models.

Climate teleconnection indices – they didn't seem correlated much with water supply. ONI (Oceanic Niño Index) improved only January predictions by 0.5 points, only when using its December values for all January rows. However, Climate Prediction Center (n.d.) says that "the index values are updated around the 10th of each month". After taking that into account, January improvement was insignificant and ultimately ONI wasn't used in the models.

1.2.4 Data preprocessing

1.2.4.1 Issue dates

Data was processed based on different issue dates. Predictions are issued in general for 7 months (Jan-Jul), 4 dates for each month (on 1st, 8th, 15th and 22nd day). Data was transformed to create rows for each issue date, with the same label for given year-site id combination.

1.2.4.2 Years constraint

Only data since 1965 was used, as it improved results. Possibly, older measuring devices could have provided lower quality data. Also, there is a possibility of changes in reservoirs over years.

1.2.4.3 Data leakage prevention

To prevent data leakage within LOOCV years, min/max calculations for historical site id volumes were performed 20 times, excluding each time the data from this fold's year. It was done similarly for distribution estimates - distribution fitting was created 20 times, without LOOCV year. Thanks to that, data from a given year wasn't used as a factor for the output parameters.

1.2.4.4 Removing outliers

Outliers removal was done using Z-score. It is usually used for data standardization but it is also recognized as a method for dealing with outliers (Anuradha et al., 2019). Though 3.0 absolute value of Z-score is an established threshold, 2.5 value was used in this case, in order to be more restrictive, due to quite small dataset. The operation was conducted only for fitting distributions to data and finding min/max values for a given site id (excluding data from a given LOOCV year).

1.2.4.5 Handling site ids with non-standard forecast season

Detroit_lake_inflow is the only site id with end of forecast season in June. Still, it was processed the same way as other sites but excluding it from July issue dates. Similar situation was with pecos_r_nr_pecos, which starts in March but used the same features and processing as other sites. Still, accumulated naturalized flow for pecos_r_nr_pecos included March values.

1.3 Uncertainty Quantification

LightGBM model is used as a main predictor. Unfortunately, its 0.1 and 0.9 predictions are sometimes very similar to Q0.5 predictions. To tackle this problem, distribution estimation was introduced, which has an effect of smoothing LightGBM results.

1.3.1 Distribution estimate calculation

After fitting best distributions for different site ids, distribution estimate logic is explained below:

1. Use LightGBM Q0.5 prediction as a base. It acts as a data knowledge. Let LightGBM Q0.5 prediction for one observation be "x".
2. Calculate x's position on CDF (cumulative distribution function) for given site id. This point becomes the median of the new distribution (which is a part of the site id's fitted distribution).

    a. If LightGBM model returned 3600 volume and it is on 0.64 quantile, then we assume that q(0.5) = 0.64.
3. Calculate a point on CDF function where maximum possible historical data for site_id ("max(site_id)") appears on. That point becomes "qmax". It is used to assess maximum value for the new distribution.
    a. If 5000 value (maximum for site id from historical data) is on quantile 0.98, then qmax = 0.98.
4. If x is greater than max(site_id), set it to max(site_id), as prediction can't be greater than max value.
5. Calculate a difference on CDF between qmax and q0.5
6. This difference becomes a base for calculating quantile 0.9 ("q0.9"). The desired quantile is then calculated based on proportions.
7. The formula for calculating q0.9 is presented below. Number in brackets shows which previous step is performed for a given formula.
    a. (5) $qmax - q0.5 = value$
    b. (6) $q0.9 = q0.5 + \frac{4}{5} * value$
    c. Multiplier 4/5 is used for Q0.9 due to the assumption that distance between quantiles is proportional; it would be 3/5 for q0.8 and 2/5 for q0.7)
8. Finding Q0.1 is similar to Q0.9 with some changes made to work with low quantile.
    a. Calculate a point on CDF function where min possible historical data for site id ("min(site_id)") appears on. That point becomes "qmin".
    b. Set x to min(site_id) if x is less than min(site_id).
    c. Calculate a difference on CDF between q0.5 and qmin.
    d. This difference becomes a base for calculating quantile 0.1 (q0.1). The quantile is calculated based on proportions.
        i. $q0.5 - qmin = value$
        ii. $q0.1 = q0.5 - \frac{4}{5} * value$

1.3.2 Combining LightGBM with distribution estimate

The final result is calculated as a weighted average of LightGBM models and distribution forecasts. Distribution estimate percentage is different for each month: 60% for Jan, 50% for Feb, 45% for Mar, 30% for Apr, 25% for May, 15% for Jun and 5% for July (23 site ids) and 10% (3 site ids). Later months require less information from distribution, as data at that time is more credible. The chosen percentages ensure that joint interval coverage is close to 0.8.

Distribution estimate acts as a smoothing effect. Increasing distribution percentage results in deteriorating performance but better interval coverage. Similarly, decreasing it increases overall performance but worsens smoothness. For some months, there exists an inflection point – if distribution percentage is below some level, decreasing it results in a worse performing model. Finding inflection points for different months wasn't used in the solution. Instead of that, the objective was to find the best trade-off between model performance and interval coverage where overall interval coverage is close to 0.8 and the models still perform well.

1.4 Training and Evaluation Process

LightGBM models were created separately for each month, with different hyperparameters. Distribution estimates were also created separately from LightGBM model training.

1.4.1 Model validation

Models were validated using a 20-year LOOCV (Leave-one-out cross-validation) for 2004-2024 years. Due to the fact that different years are regarded as independent observations, all data

was used for training, except for the year from a given LOOCV fold. Early stopping with 10 iterations step was used to stop training if model doesn't improve for 30 iterations.

Though LOOCV of the Averaged Mean-Quantile-Loss is a final metric in the competition, it was also used for hyperparameters tuning. It already covers 20 independent water years, so it is hard for the models to overfit to the data. Additionally, RMS (root mean square) of yearly results was introduced. Thanks to that, years with higher error were penalized more in the optimization (Deepchecks Community Blog, 2024), preventing overfitting to years that are easier to train.

Jan-Apr months and May-Jul models for 3 site ids used volume as a label, whereas May-Jul models for 23 site ids used volume residuals (converted back to volume in the training pipeline).

1.4.2 Hyperparameters tuning

Optuna framework (Akiba et al., 2019) was used for hyperparameters tuning as a powerful tool that uses Bayesian optimization. Different number of optimization iterations was used for different months: 150 for Jan and Feb, 130 for Mar and Apr, 60 for volume residuals for 23 site ids and 40 with volume prediction for 3 site ids without historical naturalized flow for May-Jul. Wider range of hyperparameters was used for Apr-Jul (as those months used higher quality data). There was done an additional optimization for July with 50 iterations, as the trained hyperparameters weren't optimal, taking into account best hyperparameters from the first optimization.

Each optimization iteration included no more than 2000 iterations of LightGBM to speed up the optimization process. Different months were optimized separately from each other, as well as models for 3 and 23 site ids. Distribution estimates were included in the optimization pipeline, so LightGBM hyperparameters were optimized based on both LightGBM and distribution estimation influence, including also clipping methods in the training pipeline.

2 Discussion of Performance

The presented solution achieves **86.7809** Averaged Mean-Quantile-Loss (MQL) with 0.7897 average interval coverage over 20 LOOCV years. Below are presented more in-depth results.

2.1 By issue months

Table 4 shows results divided into issue months. With better quality data each month, MQL values get better. Also, interval coverage improves nearly every month. Jan-Apr predictions don't contain data on partial monthly naturalized flow from the predicted period, so their predictive power is worse. For later issue dates, April starts to accumulate data on variables containing information from the initial stage of the predicted Apr-Jul period. In later months, even more information from the predicted period is known. July result is nearly 2 times better than June's.

Month	MQL Result	Interval coverage
Jan	136.5	0.772
Feb	122.6	0.773
Mar	111.3	0.784
Apr	88.1	0.794
May	69.0	0.799
Jun	51.3	0.796
Jul	26.4	0.812

Table 4: Monthly results

2.2 By years

Table 5 presents results by year with interval coverage, volume and sum of precipitation from SNOTEL over different site ids. As the table shows, MQL isn't consistent over all the years. It occurs especially for years with higher volumes and precipitation but the third worst result is for year 2015 with one of the lowest volumes. High errors correlate also with poorly calibrated interval coverage. 2011 and 2015 years are outliers with significantly worse interval coverage below 0.55. The next lowest value is 0.69. On the other hand, years with best results have interval coverage over 0.85. It seems that more attention in training should have been given to interval coverage from individual years, instead of average over years.

Still, 15 out of 20 years achieved MQL results below 90, with 11 years achieving MQL results below 80, which is a satisfactory result.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
MQL Result	74	89	89	66	79	76	79	166	127	65
Interval coverage	0.84	0.82	0.80	0.87	0.87	0.87	0.90	0.54	0.69	0.85
Volume	16579	22537	28335	18036	24405	23015	21715	37294	24813	18908
SNOTEL Precipitation	831	890	960	867	957	923	908	1201	903	858
Year	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
MQL Result	69	107	59	107	76	92	67	72	80	87
Interval coverage	0.87	0.53	0.93	0.69	0.81	0.69	0.90	0.69	0.83	0.80
Volume	24262	15880	20578	34293	25875	25429	22083	15547	21005	26966
SNOTEL Precipitation	853	826	969	1176	869	939	816	764	937	950

Table 5: Yearly results with accumulated water flow volume

2.3 By site id

Table 6 presents results divided into different site ids. Average volume per site id and average SNOTEL Snow Water Equivalent over different sites are additionally shown, to seek correlation between results and those conditions. Results should be interpreted mainly based on MQL ratio to average site volume, as raw MQL is influenced by site id average volumes. It seems that the solution performs especially well for sites with bigger snowpack (stehekin and skagit sites). Interval coverage seems quite solid for different site ids. Apart from detroit site with poor value of 0.66 and the highest 0.92 value for boise, other sites oscillate over 0.72-0.88. The worst ratio of MQL to volume is for sites with the lowest volume (it seems that if the maximum value is very limited, volumes tend to deviate more) and owyhee with a very small Snow Water Equivalent.

Site id	MQL Result	Interval coverage	Volume	Ratio MQL to volume	SNOTEL avg Snow Water Equivalent
pecos_r_nr_pecos	7	0.79	44	0.15	3
sweetwater_r_nr_alcova	10	0.74	53	0.19	6
virgin_r_at_virtin	9	0.83	58	0.16	5
taylor_park_reservoir_inflow	9	0.77	89	0.11	6
weber_r_nr_oakley	11	0.80	107	0.11	7
colville_r_at_kettle_falls	20	0.74	123	0.16	14
ruedi_reservoir_inflow	12	0.77	129	0.10	7

dillon_reservoir_inflow	16	0.76	156	0.10	6
green_r_bl_howard_a_hanson_dam	25	0.83	249	0.10	15
owyhee_r_bl_owyhee_dam	68	0.77	315	0.22	5
pueblo_reservoir_inflow	42	0.87	321	0.13	6
animas_r_at_durango	37	0.88	370	0.10	7
merced_river_yosemite_at_pohono_bridge	46	0.72	384	0.12	10
detroit_lake_inflow	64	0.66	495	0.13	9
boysen_reservoir_inflow	114	0.75	704	0.16	6
stehekin_r_at_stehekin	45	0.74	705	0.06	17
fontenelle_reservoir_inflow	89	0.81	733	0.12	7
yampa_r_nr_maybell	110	0.78	905	0.12	9
american_river_folsom_lake	184	0.86	1200	0.15	8
san_joaquin_river_millerton_reservoir	192	0.73	1205	0.16	11
boise_r_nr_boise	115	0.92	1209	0.10	9
skagit_ross_reservoir	97	0.77	1306	0.07	21
missouri_r_at_toston	195	0.83	1789	0.11	8
hungry_horse_reservoir_inflow	153	0.78	2016	0.08	9
snake_r_nr_heise	213	0.83	3248	0.07	9
libby_reservoir_inflow	367	0.79	5454	0.07	12

Table 6: Site id results

2.4 By site ids from the Bonus Prizes

Table 7 presents results for particular site ids and early lead time predictions.

Subcategory	Subcategory details	MQL Result	Interval coverage
Cascades	skagit_ross_reservoir, stehekin_r_at_stehekin, green_r_bl_howard_a_hanson_dam, detroit_lake_inflow	57.6319	0.752
Sierra Nevada	san_joaquin_river_millerton_reservoir, merced_river_yosemite_at_pohono_bridge, american_river_folsom_lake	140.5409	0.770
Colorado Headwaters	ruedi_reservoir_inflow, dillon_reservoir_inflow taylor_park_reservoir_inflow, animas_r_at_durango	18.6099	0.796
Chalenging basins	owyhee_r_bl_owyhee_dam, virgin_r_at_virtin, pecos_r_nr_pecos	27.9865	0.796
Early lead time	Issue dates up to March 15	125.0725	0.774

Table 7: Bonus prize site id results

3 Changes between stages

3.1 Changes between Hindcast Stage and Forecast Stage

- Changed year from which data was kept from 1930 to 1965,
- Added 20-fold LOOCV with one changing year,
- Added different distribution percentage for each month,

- Added odd years since 2005 for training, appended them also for distribution fitting,
- Optimized hyperparameters for the new settings.

3.2 Changes between Forecast Stage and Final Stage

- Added May-Jul models with volume residuals and previous naturalized flow clipping,
- Added distribution fitting separately for all LOOCV years,
- Added minimum/maximum values for site id without volume from LOOCV year,
- Added root mean square for LOOCV optimization,
- Removed outliers removal from full data processing (kept it for min/max volume calculations and distribution estimates),
- Added PDSI data, snow depth water equivalent from Seasonal meteorological forecasts from Copernicus and ERA5-Land monthly averaged data,
- Added new hyperparameters, features and distribution percentage,
- Interval coverage was taken more into consideration in optimization of results.

4 Machine Specifications

4.1 Hardware specifications

CPU: Intel Core i9-13900KF, RAM: 64GB, GPU: NVIDIA GeForce RTX 3090 Ti 24GB,
OS: Windows 11 Pro

4.2 Execution time

Table 8 shows execution time for a training pipeline - data download, processing, feature engineering and LOOCV training. Keep in mind that data download could take longer. It depends on CDS data availability. Occasionally, there is a queue that could last even a few hours.

Task	**Elapsed time**
Data download	4h+ (40m for runtime repository, 20m for ERA5-Land, 3h for 5 Copernicus forecast requests with included 2h of queue; could be more)
Data preprocessing and feature engineering for the first time (must be done only once)	40m (mostly due to SNOTEL and PDSI processing)
Data preprocessing and feature engineering after the first time was already executed	30s
LOOCV pipeline with results	36m (7m Jan-Apr, 19m May-Jul residuals, 10m 3 site ids)

Table 8: Execution time of training pipeline

Table 9 shows execution time of optional tasks. The tasks were already run and their fixed values are available to use. Distribution estimation was performed 20 times to exclude LOOCV years. For real-time prediction pipeline, distribution estimates would have to be executed only once (approximate time of one iteration of distribution estimation provided in brackets).

Task	**Elapsed time**
Distribution estimation	~5h (~20m)
Hyperparameters tuning	45h

Table 9: Execution time of calculating parameters for model training

5 References

Akiba T., Sano S., Yanase T., Ohta T. & Koyama M (2019). Optuna: A Next-generation Hyperparameter Optimization Framework

Anuradha, C., Murty, P.S.R.C., Kiran, C.S. (2019): Detecting outliers in high dimensional data sets using Z-score methodology. Int. J. Innov. Technol. Explor. Eng. IJITEE 9(1), 48–53. https://doi.org/10.35940/ijitee.a3910.119119

Climate Prediction Center (n.d.). Frequently Asked Questions Regarding CPC's Current Monthly Atmospheric and SST Index Values. https://www.cpc.ncep.noaa.gov/data/indices/Readme.index.shtml

Copernicus Climate Change Service, Climate Data Store (2018): Seasonal forecast daily and subdaily data on single levels. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.181d637e (Accessed on 22-Mar-2024)

Copernicus Climate Change Service, Climate Data Store (2024). ERA5-Land: data documentation. https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation#ERA5Land:datadocumentation-Dataupdatefrequency

Deepchecks Community Blog (2024). The Role of Root Mean Square in Data Accuracy. https://deepchecks.com/the-role-of-root-mean-square-in-data-accuracy/

Gao, B. & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. Journal of Intelligent Systems, 31(1), 767-779. https://doi.org/10.1515/jisys-2022-0036

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting.

National Integrated Drought Information System, U. S. (n.d.) U.S. Gridded Palmer Drought Severity Index (PDSI) from gridMET. https://www.drought.gov/data-maps-tools/us-gridded-palmer-drought-severity-index-pdsi-gridmet

Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.e2161bac (Accessed on 24-Mar-2024)

Muñoz Sabater, J. (2019): ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.68d2bb30 (Accessed on 22-Mar-2024)

Schmidt, L. D. & Y. Zhu (2016). Quantile spacings: A simple method for the joint estimation of multiple quantiles without crossing.

USDA Climate Hubs (n.d.). Snow Water Equivalent (SWE) - Its Importance in the Northwest. https://www.climatehubs.usda.gov/hubs/northwest/topic/snow-water-equivalent-swe-its-importance-northwest