Water Supply Forecast Rodeo - Final Model Report

Author: Rasyid Ridha (rasyidstat)
Model ID: `ens_all`
Model date: 2024-03-17

Summary

We use an ensemble of LightGBM models with Tweedie loss for point forecast and quantile loss for 0.10 and 0.90 quantile forecast. We mainly use lagged data of SNOTEL/CDEC SWE and cumulative precipitation averaged within and near the basins from top $K = 9$ SNOTEL/CDEC sites as the input of the model. In addition, for different ensemble members, we incorporate various gridded data products averaged within the basins, including gridMET PDSI, UA/SWANN SWE, ERA5-Land, and seasonal forecast products from ECMWF SEAS51. To mitigate small sample size problems, We use synthetic data generation and increase the training sample size by 5x, which significantly improves forecast skill, prediction interval reliability, and generalizability. We also incorporate daily USGS and USBR observed flow from sites with minimal impairment. Compared to the model in the Hindcast Stage, forecast skill has improved by ~5 KAF and it is attributed to additional data sources used and better estimates of snowpack in some regions, especially in Sierra Nevada and Cascades region.

Keywords: LightGBM, Tweedie, ensemble, SNOTEL, CDEC, gridMET PDSI, UA/SWANN, ERA5-Land, SEAS51, synthetic data

Algorithm and Architecture Selection

LightGBM

We use LightGBM as the base model for this challenge. It is a fast, high-performance, flexible, and SOTA model for tabular or regression tasks. It also can capture non-linearity which is common in hydrological processes. Furthermore, LightGBM's efficient handling of categorical data allows us to train a single global model, rather than training multiple local models for each site and issue date. This global training approach is preferred for its simplicity, scalability, and increased forecast skill benefit from cross-learning capabilities (Montero-Manso & Hyndman, 2021).[5] In more detail, our LightGBM base model consists of this architecture and parameters:

- Loss function: Tweedie loss (for point forecast), Quantile loss (for quantile forecast)
  The distribution of target variables is quite skewed for high flow especially if we take all basins. LightGBM with Tweedie loss yields a slightly better forecast skill compared to L1 (MAE) or L2 (RMSE) loss. In addition, it enforces non-negative prediction which is more physically plausible compared to standard regression using L1/L2 loss.
- Minimal model parameter tuning
  We did not focus on parameter tuning as it is not recommended considering we have a small sample size for each basin.[1] Instead, we focus on experimenting with new data sources and feature engineering. We only changed `min_data_in_leaf` and `num_leaves`

parameters as it is recommended to have smaller values of these parameters for a small sample size.

- No early stopping applied during model training
  It appears that there is a best score for a single validation year with different training iterations but it is shown to be inconsistent overall. Instead, we select a fixed number of training iterations which is generalized to any water year even though it is near-optimal on individual water years.

Synthetic data generation

On this challenge, we face the issue of small sample sizes for each site based on number of training years. There are some sites with historical data more than 100 years ago but it cannot be used since the features input is not available. Small sample sizes can cause overfitting, especially if we include more features and use more complex models. To mitigate this problem, we use the synthetic data generation approach applied in the study by Herbert et al, 2021.[2] For each training year, we generate $S = 4$ synthetic years, increasing the training sample size by 5x. For each synthetic year, we multiplied the target and feature value with a scalar that was randomly selected between 0.5-1.5. This approach significantly improves forecast skill, prediction interval reliability, and generalizability, with estimated 1-3 KAF improvement.

Ensemble

For each quantile, we combine forecast output from multiple variants of a single LightGBM model (Table 1) by aggregating them with averaging for point forecast and percentile 10% and 90% for quantile forecast. Unlike the Hindcast Stage, the ensemble mechanism does not involve inner fold ensemble averaging. It still can be done by using a nested LOOCV approach but the training iteration time will increase 19 times. Without inner fold ensemble averaging, we sacrifice ~1 KAF improvement. Despite that, the ensemble's performance gain is better than our Hindcast Stage model because it uses multiple variants and benefits from the strengths of different data sources. In addition, we design the model variants and ensemble to be operationally robust in case some of the data sources are not available. Rather than relying on a model with all features and tightly dependent, we develop multiple models ranging from those with minimal dependencies to more complex ones, which can be combined using ensemble and still yield better performance compared to a single model.

Training with supplemental NRCS sites

We experimented with model training using supplemental NRCS sites, incorporating only ERA5-Land data for the initial trial. There is an improvement in forecast skill, even without using synthetic data generation, with estimated 1-2 KAF improvement (Table 8). This improvement might be attributed to larger samples across basins and the effect of cross-learning from similar basins (Kratzert et al., 2019; Montero-Manso & Hyndman, 2021).[3,5] However, we did not experiment with all model variants as the training time is more than four times longer. In addition, we use basin polygons for supplemental sites from NLDI which are not confirmed as data sources that need extra approval or not. Another alternative is to derive basin polygons

from Copernicus DEM GLO-90 or use BasinATLAS Basin Attributes from approved data sources list.

Consideration of using other models

We considered using LSTM and framed the problem as monthly time series forecasting with 1-7 months of the forecast horizon. However, the sample size is limited after considering available features in different water years. In addition, bottom-up forecasting settings might accumulate the error and uncertainty bigger for a longer forecast horizon period. LSTM might be a better option if we are interested in a more granular scale (hourly, daily, weekly) which can capture a more detailed hydrology process rather than a monthly or seasonal scale.

Model simplification

Compared to the model in the Hindcast Stage, we have decreased the number of ensemble members by excluding different random seed iterations. To further simplify the model, we suggest excluding model variants that exhibit lower forecast skills based on validation scores or adjusting them as needed according to data availability in operational settings.

Data Sources and Feature Engineering

Data and features selection

There are a lot of data sources to be used for this challenge. Hence, it's important to have criteria and prioritization on which data we want to experiment with. Our data source selection criteria are defined as follows:

1. **Historical data availability**: data is available with enough history and lookback period (statistical rule of thumb is at least 30 sample size = 30 training years per basin)
2. **Data recency**: the gap between the forecast issue date and the latest data date is minimal to make sure that we get the latest condition and information available (especially for the Forecast Stage)

In addition, some technical feasibility is considered such as data size is not too big, preprocessing should be easy and fast so we can iterate more experiments. For feature selection, we define features to be included based on:

1. Correlation with the target variable
2. Cross-validation result

It's possible that there is a significant correlation and signal with the target variable but it does not reflect well on the validation score. Hence, we discard this kind of feature as it is uncertain and can lead to overfitting.

Data sources

We use multiple data sources that meet the criteria outlined above to forecast water supply volume, focusing on key parameters like snowpack, precipitation, soil moisture and temperature. Below are the details of data sources used (U), experimented (E) and considered (C).

Data type	Data sources	Status	Comment
Antecedent streamflow	NRCS and RFC monthly naturalized flow	U	Antecedent streamflow is used for target preprocessing to reduce error for issue dates within the season forecast months. However, we do not use it as a direct feature since there is no significant improvement in overall forecast skill. In addition, not all sites have available observed and natural flow data.
Antecedent streamflow	USGS daily streamflow	U	
Antecedent streamflow	USBR daily reservoir inflow	U	
Snowpack	NRCS SNOTEL daily SWE and precipitation	U	Snowpack measurement is very crucial in determining seasonal water supply volume. NRCS SNOTEL/CDEC SWE and precipitation in-situ measurement include long history data and have been used mainly in the operational setting. In addition, there is a strong multi-year variability and consistency which is very important for generalizability.
Snowpack	CDEC daily SWE and precipitation	U	
Snowpack	UA/SWANN SWE and precipitation	U	UA/SWANN data is used to complement SNOTEL/CDEC data with additional assimilation to better reflect spatial variability of snowpack measurement within the basin.
Drought condition	gridMET PDSI	U	gridMET PDSI is used to better reflect drought conditions of the basin. Incorporating PDSI improves forecast skills in dry periods.
Weather	ERA5-Land	U	ERA5-Land global weather product includes historical data dating back to 1940. It contains various weather and meteorological variables, including SWE. It can be used to complement basins with limited or no SNOTEL/CDEC stations, such as Skagit Ross Reservoir and Libby Reservoir Inflow.
Weather forecast	Seasonal meteorological forecasts from Copernicus ECMWF SEAS51	U	ECMWF SEAS51 seasonal forecast product includes re-forecast dating back to 1981. We only use re-forecast to a maximum lead time of one month because longer lead times have limited forecast skill.
Teleconnection index	ONI, Nino SST, SOI, MJO, PNA, PDO	E	There is a potential to improve long lead time forecasts using multi-year teleconnection indices. For issue dates in January, we saw slight improvement based on median skill but not with average skill. In the end, we discarded these features since there is bigger uncertainty and inconsistency which leads to worse overall forecast skill.
Weather	RCC-ACIS PRISM precipitation and temperature	E	PRISM gridded data includes historical data dating back to 1981. However, there is no significant improvement in forecast skill using only cumulative precipitation and temperature. Additional preprocessing might be done to extract snowpack estimates using SNOW-17 model to further complement SNOTEL/CDEC data.
Snowpack	SNODAS	C	Even though we can use more sample size in the

Snowpack	MODIS Snow Cover	C	Overall Stage, we still have limited training water years to be used with these data sources. Based on empirical results, training with 23-28 years of data results in poorer forecast skills. Hence, we decide to not further experiment with these data sources.
Vegetation	MODIS Vegetation Indices	C	
Weather forecast	CPC Seasonal Outlooks	C	

Feature engineering

Lag features

We use lag features before the issue date which ensures that there is no future data leak. For some cases where recent data is not available, we prepared backup models with longer lag periods, deployed in the Forecast Stage. For most of the data sources, we use a lumped approach where parameters are aggregated using mean within basin polygon. In addition, we use HUC (hydrologic unit code) and link USGS ID based on the mapping from USGS Water Services. Below are the details of parameters and preprocessing from each data source.

Data sources	Parameters	Lags	Comment
SNOTEL/CDEC	swe, prec_cml	$t-1$ $t-8$ $t-15$	All parameters are aggregated using mean based on the highest coefficient of determination with target variable from top $K=9$ SNOTEL/CDEC sites for each basin. Pair between basins and SNOTEL/CDEC sites is determined based on four approaches: within basin polygon, 10 km near the basin polygon, same HUC 6-digit code, and 200 km near the basin sites location. Before that, scaling using 0-max for each SNOTEL/CDEC site is applied so the parameters can be aggregated even though scale is different and there's partial missing data.
gridMET PDSI	pdsi	$t-5$	PDSI is aggregated using mean for each site within basin polygon.
UA/SWANN	uaswe, uaprec_cml	$t-1$	We use gridded data and aggregated HUC 6-digit and 8-digit code. For gridded data, all parameters are aggregated using mean for each site within basin polygon.
ERA5-Land	t2m, lai_hv, lai_lv, asn, snowc, sd, swvl1, swvl2, swvl3, swvl4	$t-1$	Due to storage constraints, we only retrieve lag 1 days data from the issue date. With this limitation, we only include instantaneous parameters like snow, temperature, soil water, and vegetation index which are aggregated using mean for each site within basin polygon. Later, we noticed that it's better to use lag 5 days to mimic the operational setting because the real-time version, ERA5-Land-T, is available 5 days later from the current date. However, download queue time can take several weeks for all training years so we use the existing one for simplicity.
Seasonal meteorological forecasts from Copernicus ECMWF SEAS51	sd, tprate, t2m	N/A	We use re-forecast up to 1 month lead-time from the issue date starting at day 8 of each month for parameters such as snow, total precipitation and temperature. All parameters are aggregated using mean for each site and ensemble member within basin polygon. Final features are aggregated using mean, percentile 10% and 90% from 26 ensemble members.

Static and supplemental features

We use static features from the given metadata such as elevation, longitude, latitude, region, drainage basin size, season target months, and number of season months. For missing drainage basin size, we estimate the value using simple linear regression. We also include supplemental features such as month, day of month, day of year, and extra flagging whether target differencing is applied or not. In general, the forecast skill improvement and importance with static features is minimal since we only have 26 basins with varying characteristics in the training dataset.

Target preprocessing

We apply two types of target preprocessing such as:

- Scaling using 0-max to better benefit from cross-learning, also used for extension of training with all supplemental NRCS sites
- Differencing the ground truth with the latest known volume for issue dates within the seasonal month target. Latest known volume is derived from monthly NRCS naturalized flow from 23 sites and daily USGS/USBR observed flow from 14 sites with minimal impairment determined based on the ratio of historical observed flow and naturalized flow near 1 (Table 9)

The preprocessed target is used as the final target when training the model and we do the inverse transformation for inferencing by rescaling and summing back the forecast with known volume (example forecast result: Figure 7 with NRCS data only, and Figure 8 with NRCS/USGS/USBR data).

Missing data gaps handling

We use the previous value (`pandas.DataFrame.ffill`) to handle missing data gaps. This approach is preferred because there are limited missing data gaps and it is simple to use. A better approach is to use the interpolation method (`pandas.DataFrame.interpolate`) but it has the risk of future data leakage since we designed the dataset to be unified for feeding into a single model rather than separate models for each issue date. In an operational setting, we expect that there will be more missing data gaps and using the interpolation method is preferred with no risk of future data leakage.

Post-processing

Since we use a model-based approach to estimate the prediction interval, there is a possibility of incoherent prediction coming from each model with different losses. This happens because there is a small variance and narrow prediction interval, usually in later issue dates. To handle this, we apply post-processing by rearranging the prediction to match its quantile orders.

Physical explanation and features intuition

Our model mainly used known historical snowpack and precipitation as these two are major drivers for total seasonal water supply volume (Figure 9). For snow-dominated basins, snowpack is the best proxy for future seasonal water supply volume as it stores water during the winter season and releases water during the melting season, reflecting total water supply volume in the future. For basins with no or less snow, prediction might be more difficult as we

rely mostly on historical precipitation and have many unknowns on future weather. In addition to snowpack and precipitation, we also include other features in the Overall Stage to better describe the initial condition of the basin, such as temperature which can be used to describe many hydrological phenomena like snow accumulation, snow melting and evapotranspiration, as well as soil water and PDSI which is used for drought indicator. It would be interesting to see more detailed explanations in the Model Explainability part where we can see the difference and contribution of each feature based on different locations, climate conditions, lead time and quantiles.

Uncertainty Quantification

We use quantile loss in LightGBM for uncertainty quantification for the 0.10 and 0.90 quantiles as mentioned in Algorithm and Architecture Selection.

Training and Evaluation Process

Evaluation scheme

Since the beginning of the challenge, we used year-wise LOOCV (leave-one-out cross-validation) as the evaluation scheme. This approach is preferred due to the limited number of training years that can be used. Nevertheless, there's still a risk of overfitting, especially if we do extensive hyperparameter tuning and always validate using the same dataset in years 2004-2023 for the Overall Stage final submission. It's recommended to use nested LOOCV where inner cross-validation is used solely for hyperparameter tuning or any decision for model selection and recalibration, and outer cross-validation for model performance estimation. However, it comes with the cost of higher training time which is 19 times longer. In addition, since we used fixed model hyperparameters and did not change model architecture from the beginning, we ended up using existing LOOCV for efficiency.

In our case, the risk of overfitting might arise from the selection of ensemble members. Therefore, before submitting the final submission, we extend the validation including years 1991-2003 to ensure that our ensemble is robust. Generally, we achieve consistent performance where the gap between the best single model and ensemble model is 4-6 KAF, similar to 2004-2023. Regarding forecast skills, years 1991-2003 have a lower score, ~73 KAF, compared to years 2004-2023. This gap indicates that recent years 2004-2023 are more challenging to forecast as there are more extreme years, notably in 2011 and 2015. By excluding these years from evaluation, we get a similar score, ~74 KAF, suggesting that our ensemble model is robust at different sets of years and not overfitting in years 2004-2023.

Training process

For each validation year, we train 9 variants of single LightGBM models (Table 1) for all basins and issue dates. Years of the training dataset start from 1981, in which all data sources are available. Extreme years in 2011 (wet) and 2015 (dry) are excluded from the training dataset used for model development in the Forecast Stage and Overall Stage because these

anomalous years can increase uncertainty and worsen forecast skills. Unless there are antecedent parameters that can be used to describe the anomaly, we may include these extreme years in the training dataset, especially to overcome the challenge of extreme years in the future.

The training dataset is formed by excluding a single validation year of interest. The computation below only uses the training dataset to avoid data leakage and improve the variety of the datasets:

- Aggregation statistics for scaling SNOTEL/CDEC features and target variable for each site
- Selection of pair between SNOTEL/CDEC sites and basins based on the top $K = 9$ highest coefficient of determination between SWE and target variable
- Synthetic data generation applied only to the training dataset

Using this approach, we can have different pairs between SNOTEL sites and basins, and also different synthetic data generation scale factor configurations for each validation year. This approach will benefit in more variety of training datasets formed and improve generalizability. In addition, we also cached aggregation statistics and pairs between SNOTEL sites and basins, to be used later for inference. Finally, we generate all quantile forecasts from all model variants and apply an ensemble to get the final forecast.

Model Performance

Relative error metric and baseline

Given that the primary metric for this challenge relies on absolute metrics, we also incorporate relative metrics to get a better view of model performance across sites with different scales. We extend MPL (averaged mean pinball/quantile loss) with the formulas below:

- Normalized: $NMPL = MPL/\bar{y}$
- Skill score: $MPLSS = 1 - MPL/MPL_{climatological}$

For the skill score calculation, we use a climatological baseline which is calculated based on the median, percentile 10%, and 90% of historical seasonal water supply volume before 2004. Compared to the Hindcast Stage, we also enhance this baseline with target differencing within season months so model comparison will be more aligned.

Overall performance

Our model has a satisfying performance of 79.49 KAF for MPL and 0.42 for MPLSS (Table 1). Compared to our model in the Hindcast Stage (Table 2), forecast skill has improved by ~5 KAF and it is attributed to additional data sources used and better estimates of snowpack in some regions, especially in Sierra Nevada and Cascades region. Single models used for the ensemble have performance ranging from 85-90 KAF, with the best model being `pdsi_swe_era5` based on MPL and `pdsi_swe_era5_s51` based on MAE. Generally, models with more features have the lowest MAE but they have bigger uncertainty leading to worse MPL.

Performance across conditions

Forecast skill is limited at long lead time as the snowpack accumulation is not fully complete (Table 3). As we get closer to the seasonal month target, it should be expected that forecast skill gets better (Table 4, Figure 1). In particular, forecast skill improves significantly in April as the snowpack usually has already fully accumulated and begins to melt.

To assess model performance in different weather conditions, we use yearly average water supply and classify years into wet, normal, and dry. While weather conditions may vary across regions annually, for simplicity, we use this classification. Overall, our model performs generalized and balanced across different water years and weather conditions (Table 5, Figure 2). This is attributed to the diverse models used for the ensemble which have their own unique strengths. Models with SNOTEL/CDEC snowpack data tend to be more accurate in predicting water supply volume in wet years but may overpredict in dry years. Conversely, models without SNOTEL/CDEC snowpack data may underpredict in wet years due to potential biases in snowpack estimation, which is based on data assimilation and basin averaging. However, its spatial attributes are better at capturing dry years, especially when combined with drought indicators.

Model performance varies across different locations with normalized MPL ranging from 0.05-0.20 (Table 7). This varying performance is expected since the location used for this challenge has diverse characteristics. For instance, forecast skill is highly correlated with the percentage of snow cover area within basin (Figure 5). On the other hand, it is less correlated with amount of volume (Figure 6), drainage area size (Figure 3), and site elevation (Figure 4) which might seem to be less obvious due to the small amount of basin samples used for this challenge.

Relative to the baseline, our model has the highest MPLSS for basins located in Sierra Nevada (Table 6). Nevertheless, these basins are challenging to forecast since the location has a mix of snow areas and lower normalized MPL compared to other sites (Table 7). The fact that this region has the highest MPLSS is because there is high variability and rarely water supply volume within the normal range.

Performance limitation in the operational setting

Despite having good cross-validation, the hindcast score provides an overoptimistic view of forecast skill since all the data is present, available, and already approved for use. In the operational setting, some limitations such as data availability and data quality can worsen the forecast skill.

To mitigate the issue of data availability, we can have a model trained with larger lag gaps as we did in the Forecast Stage by including lag t-3 for SNOTEL data. Also, we rely on the latest known volume from NRCS where having lag t-1 availability might be too overoptimistic, we recommend developing a model using lag t-2 where monthly volume can be used on 8th of the

month. In addition, for the model that we develop for the Overall Stage, we recommend using ERA5-Land data with lag t-5 or even lag t-7. Previously, in the Hindcast Stage, we developed the model with larger lag gaps with estimated 1-2 KAF reduction in forecast skill.

For data quality, features and target variables are still provisional and can be updated later because of some adjustments. Since we rely on the latest known volume, slight deviation can make the prediction interval unreliable, especially in later issue dates. Hence, having a wider prediction interval might be preferred for issue dates within season months, without the need for recalibration (Table 4).

Recommendations

We presented an ensemble of global LightGBM models to improve seasonal water supply forecasting. Through the analysis, we showed that having accurate snowpack estimation is pivotal to enhancing forecast skills. Our model uses various snowpack data, leveraging in-situ measurement using SNOTEL/CDEC data, combined with various gridded data products which are better at capturing the spatial variability. By incorporating these diverse data sources and models within the ensemble, we further improve generalization and applicability across varying conditions.

In an operational setting, our model can handle problems when a data source is not available by having multiple models trained with varying dependencies. These models can be used for ensemble and still yield better performance compared to a single model. It also addresses the dependency issue when there is no in-situ measurement available by having models trained only with gridded data.

Our global model approach is also scalable and can be further extended for training with all supplemental sites. Our first empirical finding suggests that training with all sites can enhance forecast skills. We recommend increasing samples across basins with similar characteristics rather than increasing samples within the basin by using synthetic data generation. Global models need fewer parameters which gives us room to build a more complex model with more diverse data sources and features engineering. Incorporating more static features, such as those derived from DEM, will be more relevant to further benefit from the cross-learning capability of a global model.

Nevertheless, some adjustments still can be made to enhance the forecasting skills of our ensemble model. For instance, we can apply post-processing or re-calibrating prediction intervals as we have more diverse ensemble members which can make the prediction interval to be wider. In addition, we also can do weighted ensembles or develop meta-learning models to further optimize the ensemble selection rather than simple averaging. However, we decided to not do this for the final submission since it can increase the risk of overfitting, especially when we only have small datasets to be used for the optimization and evaluation. We recommend validating with all training years, using nested LOOCV, or extending the training using all supplemental sites.

In addition, there are many other ways to approach seasonal forecasting tasks such as by using sub-model, semi-distributed, and bottom-up forecasting settings. For example, we can use a monthly forecasting approach with weekly initialization and condition, then sum up the forecast prediction flexibly based on the season month of interest. Also, we can build long-range multi-year forecasting, using different features such as teleconnection index and other long-term weather features, combined with medium to short-range forecasting which uses initial conditions and time series of key parameters. For larger basins, it's recommended to train the model by sub-basins or use features engineering calculated based on different zones (snow vs arid, high vs low elevation, high vs low vegetation index, etc.). While the approach increases complexity, it has the potential to enhance forecast skills, improve explainability, and flexibly address diverse sub-tasks within hydrological forecasting.

Changes Between Stages

Stage	Changelog
Hindcast	Submitted model is an ensemble of 90 LightGBM models originating from single model variant (SWE) with 9 random seeds and 10-fold validation years
Forecast	Submitted model is an ensemble of 36 LightGBM models originating from 4 model variants = 2 base features (SWE; SWE+PDSI) x 2 varying top-K SNOTEL sites(K=5; K=9), with 3 random seeds and 3-fold validation years Changes in general: ● Retrain with additional data from odd years 2004-2023 ● Exclude anomalous years for training (2011 and 2015) ● Add PDSI features ● Scale SNOTEL SWE and cumulative precipitation with 0-max so averaging still can be done in case of missing SNOTEL sites Changes only applied in the Forecast Stage for simplification and deployment: ● Only include recent lag t-1 or t-3 for SNOTEL data features to reduce dependency ● Reduce random seeds iteration for ensemble (9 → 3) to reduce model size ● Inner CV ensemble is based on recent 3 validation years (2020-2022) rather than all 20 validation years to reduce model size ● Corner case handling if data not available ○ Minimum 65% of SNOTEL sites is available for a given site, otherwise forecast skill decline will be higher than 3 KAF ○ Backup models with K=5 SNOTEL sites and lag t-3 to reduce dependency in case of missing stations ○ In case less than 65% of SNOTEL sites are available for a single site or any dependent features are missing, the inference code will always pick the ensemble of models with the latest data available ● Update inference code to filter date based on issue date for clarity (in fact, it does not change the result since we use lagged variables as the input and this already guarantees no future data leak) ● Update inference code to preprocess data for each issue date rather than in bulk like in the Hindcast Stage

Final (Overall)	Submitted model is an ensemble of 9 LightGBM model variants (Table 1) Changes in general: ● Add CDEC SWE features ● Add UA-SWANN SWE features ● Add seasonal forecast ECMWF SEAS51 features ● Add ERA-5 land features ● Scale target variable with 0-max ● Ensemble ○ Disable inner CV ensemble (averaging across inner fold iteration, this can still be done if we use nested LOOCV with drawback of longer training time) ○ Disable different random seeds variation (3 → 1) ● Synthetic generation function is adjusted for features with negative values or negative correlation as we add new features (e.g. temperature) ● Use climatological baseline with target differencing for more equivalent relative skill score calculation

Machine Specifications

- CPU: Core i5
- RAM: 8GB
- Training data preprocessing for all years 1981-2023: 3-4 hours
- Training duration: ~3 hours for all 720 models (20-fold years x 9 model variants x 4 losses)
- Inference duration: less than 3 minute for a single issue date and 26 sites (not including data download time)

In general, download time is fast and not a big issue, except when retrieving data from CDS API, which on our side, queue time can take ~1 hour for a single date or month. In addition, there are some optimizations to reduce training and processing time: (1) exclude MAE loss since it's not needed and we use Tweedie loss for the final solution, (2) caching some data processing steps, e.g. selection of SNOTEL sites and basins within the inner loop of cross-validation

Appendix

Glossary

- **MPL**: averaged mean quantile loss (primary metric of the challenge)
- **MPLSS**: MPL skill score with climatological as reference
- **IC**: interval coverage of prediction within 10% and 90% range (secondary metric of the challenge)
- **ens_all:** ensemble of all 9 model variants
- **ngm:** model without SNOTEL/CDEC data
- **ens_ngm**: ensemble of 4 model variants without SNOTEL/CDEC data

Table 1: Overall forecast skill (ensemble and single models)

Model variant	MPL	MPLSS	MAE	IC
ens_all (submitted for Overall Stage)	**79.49**	**0.420**	**126.20**	83.74%
ens_ngm	80.98	0.409	127.87	78.53%
base_swe	86.25	0.371	137.78	70.76%
swe_ua	85.93	0.373	135.17	67.54%
pdsi_swe_s51	86.60	0.368	130.65	59.66%
pdsi_swe_era5	**85.54**	**0.376**	129.90	61.69%
pdsi_swe_era5_s51	86.34	0.370	**129.09**	59.48%
ngm_ua	88.71	0.353	138.89	68.14%
ngm_pdsi_ua_s51	88.56	0.354	131.36	57.74%
ngm_pdsi_era5_s51	90.68	0.338	133.33	60.73%
ngm_pdsi_ua_era5_s51	88.55	0.354	129.51	58.89%

Table 2: Overall forecast skill - baseline and previous models (SWE and PDSI+SWE)

Model variant	MPL	MPLSS	MAE	IC
lgb_sweK9L2S1_pdsi_diffp_S4_m3_ff_ens18_pfs (equivalent to ensemble of swe and pdsi_swe, retrain with all data, no inner CV ensemble, same configs in Overall Stage)	83.41	0.391	133.58	75.32%
lgb_sweK9L2S1_pdsi_diffp_S4_m3_ff_ens18_pfs (model variant for Forecast Stage)	84.57	0.383	135.02	82.38%
lgb_sweK9L2S1_diffp_S4_m3_ff_ens9_pfs (submitted model for Hindcast Stage)	85.72	0.374	138.23	81.04%
climatological_baseline (with target differencing, reference for skill score)	137.02	0.000	220.70	75.90%
climatological_baseline	163.79	-0.198	266.01	72.96%

Table 3: Forecast skill for different lead times

Lead time	MPL	MPLSS	MAE	IC
Long lead time (issue dates from 1 Jan to 15 Mar)	120.23	0.266	194.07	77.97%
Short lead time (issue dates from 22 Mar to 22 Jul)	52.88	0.558	81.88	87.50%

Table 4: Forecast skill for different lead times (monthly)

Month issue date	MPL	MPLSS	MAE	IC
January	134.90	0.176	219.54	78.85%
February	118.24	0.278	191.57	76.92%
March	100.52	0.386	158.74	78.99%
April	76.66	0.526	119.90	85.67%
May	62.33	0.566	96.76	87.16%
June	41.47	0.614	63.35	89.42%
July	20.00	0.614	29.80	89.35%

Figure 1: Boxplot of forecast skill for different lead times (monthly)

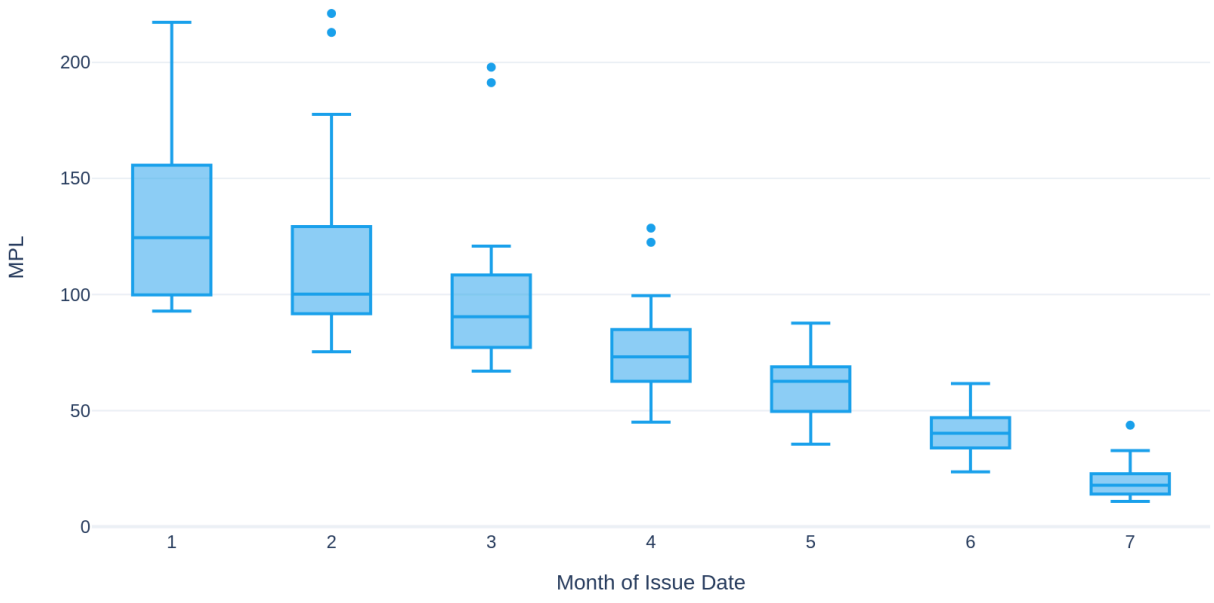


Table 5: Forecast skill under different climate conditions

Condition	MPL	MPLSS	MAE	IC
Dry (<775 KAF)	74.47	0.425	122.01	76.96%
Normal (775-975 KAF)	75.11	0.328	116.12	87.11%
Wet (>975 KAF)	90.23	0.502	144.80	84.32%

Figure 2: Yearly forecast skill of ensembles, the best single model and base model

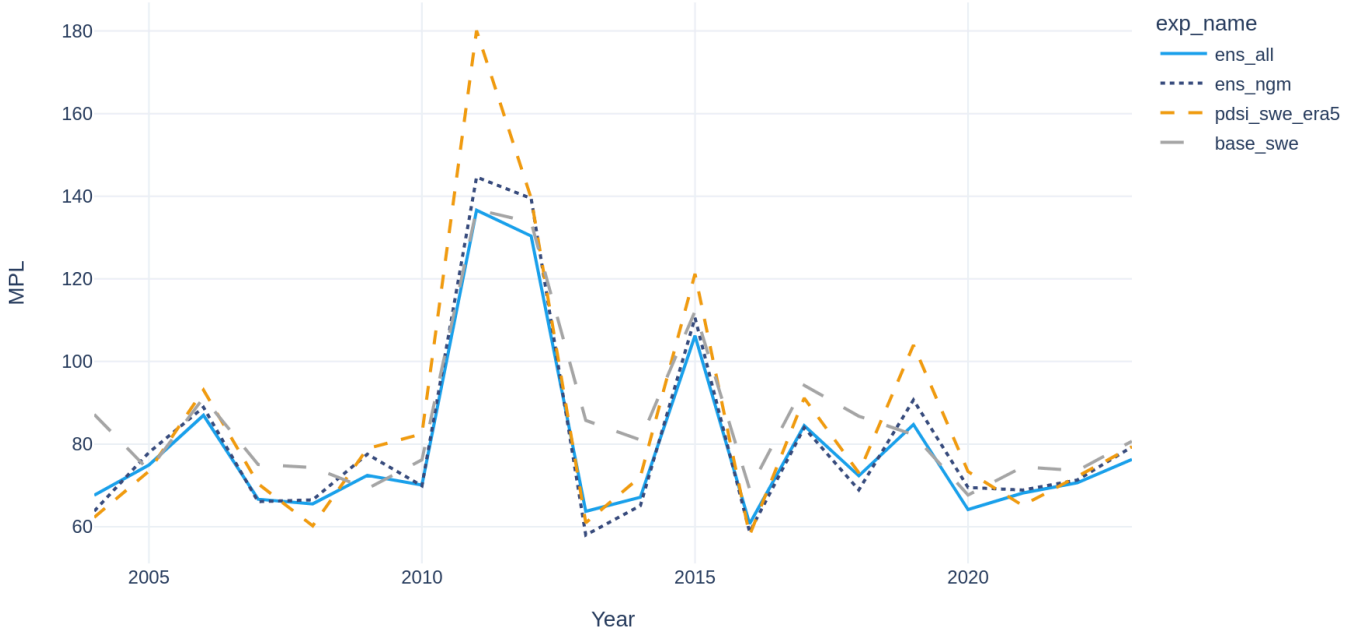


Table 6: Regional bonus prize forecast skill

Region	# sites	MPL	MPLSS	MAE	NMPL	IC
Cascades	4	50.53	0.346	81.36	0.083	81.76%
Sierra Nevada	3	118.68	0.542	191.51	0.119	80.95%
Colorado Headwaters	4	16.53	0.449	26.09	0.088	85.89%
Challenging Basins	3	23.55	0.400	37.34	0.139	84.46%

Table 7: Site-level forecast skill, sorted from low to high based on normalized MPL

Site ID	MPL	MPLSS	MAE	IC	NMPL
stehekin_r_at_stehekin	34.36	0.457	52.29	87.86%	0.049
snake_r_nr_heise	203.28	0.514	333.39	87.50%	0.063
libby_reservoir_inflow	375.91	0.335	571.05	83.93%	0.069
skagit_ross_reservoir	90.40	0.386	141.29	81.79%	0.069
hungry_horse_reservoir_inflow	142.46	0.389	224.07	78.57%	0.071
weber_r_nr_oakley	8.43	(T3) 0.549	13.63	86.43%	0.078
ruedi_reservoir_inflow	10.40	0.445	16.80	85.54%	0.081
boise_r_nr_boise	102.83	0.444	168.73	86.96%	0.085
merced_river_yosemite_at_pohono_bridge	33.64	(T1) 0.640	52.67	84.29%	0.088
animas_r_at_durango	33.15	0.454	51.07	87.32%	0.090
dillon_reservoir_inflow	14.33	0.439	24.00	84.11%	0.091
taylor_park_reservoir_inflow	8.23	0.447	12.51	86.61%	0.092
green_r_bl_howard_a_hanson_dam	23.89	(B3) 0.259	38.31	81.07%	0.096
yampa_r_nr_maybell	90.33	0.414	141.52	84.64%	0.100
fontenelle_reservoir_inflow	74.60	0.450	124.67	87.86%	0.102
missouri_r_at_toston	196.43	0.271	304.55	84.11%	0.110
pecos_r_nr_pecos	5.02	0.492	8.03	86.07%	0.112
detroit_lake_inflow	57.94	(B2) 0.178	95.52	75.42%	0.117
virgin_r_at_virtin	7.06	0.538	11.43	88.04%	0.120
american_river_folsom_lake	144.29	0.438	235.26	81.61%	0.120
pueblo_reservoir_inflow	39.08	0.323	61.95	86.61%	0.122
boysen_reservoir_inflow	102.11	0.324	170.90	82.68%	0.145
colville_r_at_kettle_falls	17.87	(B1) 0.166	26.34	75.89%	0.145
san_joaquin_river_millerton_reservoir	178.21	(T2) 0.582	286.80	77.14%	0.148
owyhee_r_bl_owyhee_dam	58.58	0.368	92.57	78.75%	0.186
sweetwater_r_nr_alcova	10.78	0.282	17.39	84.29%	0.203

Figure 3: Scatter plot of site-level normalized MPL and drainage area

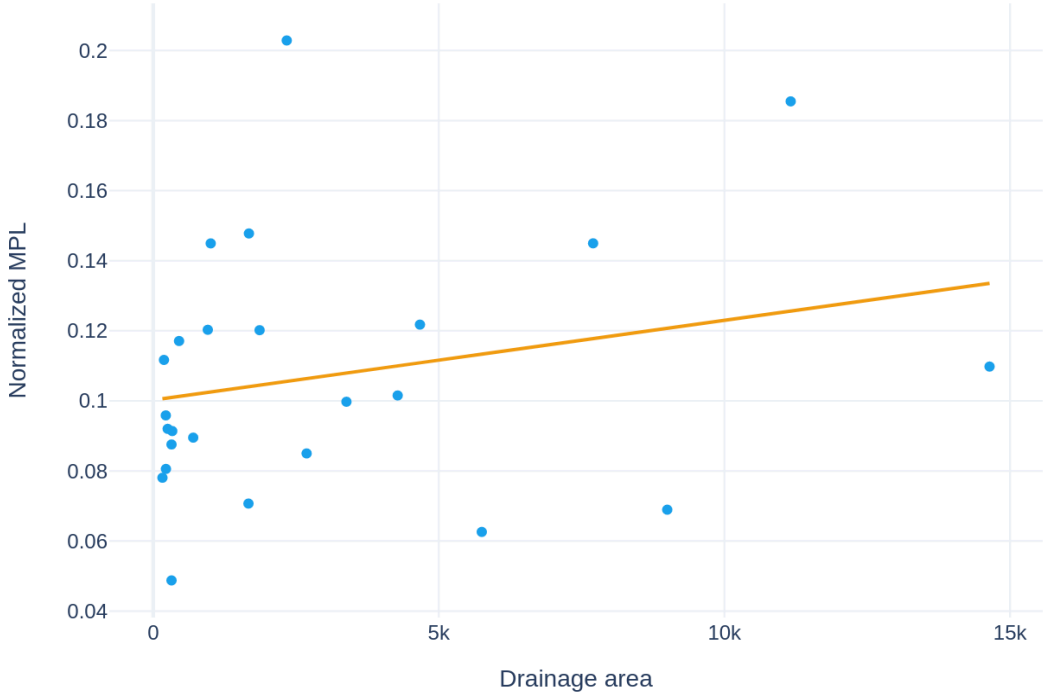


Figure 4: Scatter plot of site-level normalized MPL and site elevation

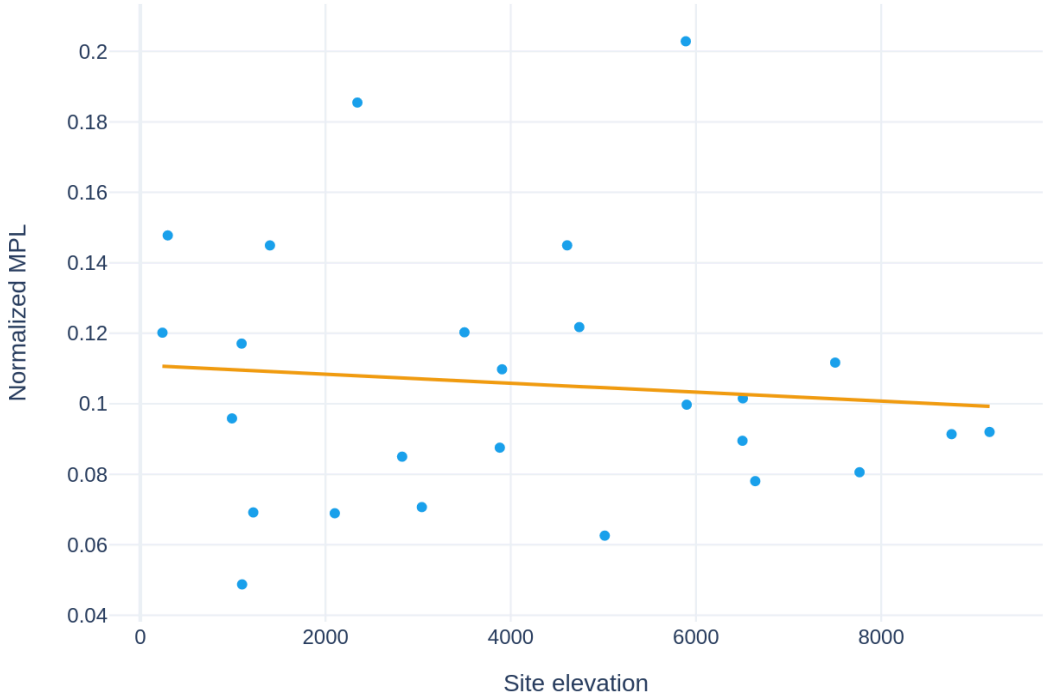


Figure 5: Scatter plot of site-level normalized MPL and average % snow cover area

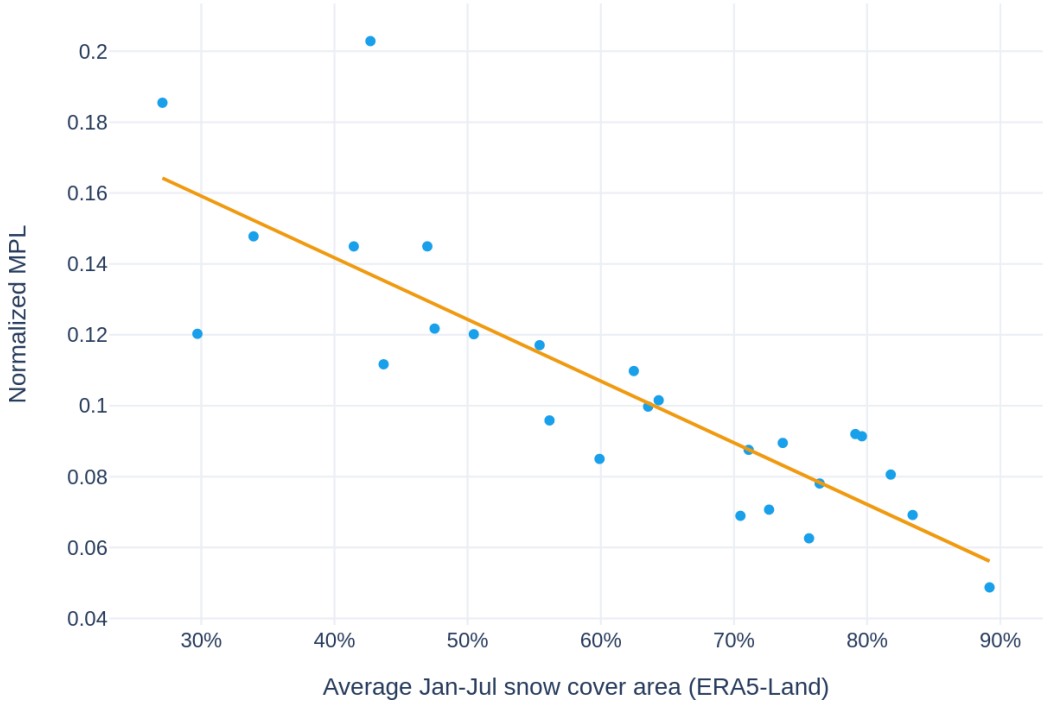


Figure 6: Scatter plot of site-level MPLSS and average seasonal water supply volume

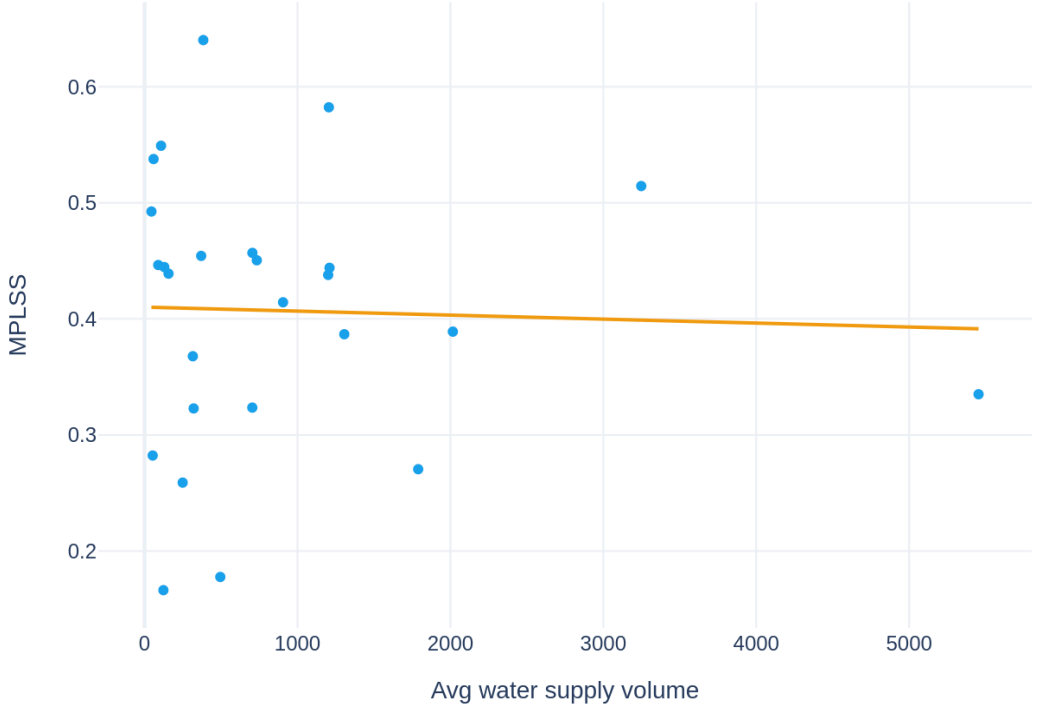


Table 8: Forecast skill of model trained using supplemental NRCS sites and ERA5-Land data only. The forecast skill is worse for sites in Sierra Nevada. This result may arise from the fact that supplemental NRCS sites do not cover all sites in Sierra Nevada, as it is managed by CDEC

Model variant	MPL					
	Overall	**Long Lead Time**	**Cascades**	**Sierra Nevada**	**Colorado Headwaters**	**Challenging Basins**
ERA5-Land (with supplemental NRCS sites)	**87.55**	**126.98**	**53.14**	156.26	**18.48**	**22.91**
ERA5-Land (only 26 sites)	89.44	132.69	54.96	**144.51**	19.39	25.37

Table 9: List of 14 sites with minimal impairment, utilizing USGS/USBR observed flow for target differencing. Observed and naturalized volume ratio is calculated for season target month using years 1980-2003

Site ID	Observed and Naturalized Vilume Ratio
american_river_folsom_lake	1.096
fontenelle_reservoir_inflow	1.000
boysen_reservoir_inflow	1.109
taylor_park_reservoir_inflow	1.000
colville_r_at_kettle_falls	1.000
animas_r_at_durango	1.000
stehekin_r_at_stehekin	1.000
pecos_r_nr_pecos	1.000
virgin_r_at_virtin	1.000
yampa_r_nr_maybell	1.002
merced_river_yosemite_at_pohono_bridge	1.000
missouri_r_at_toston	1.044
weber_r_nr_oakley	1.014
green_r_bl_howard_a_hanson_dam	1.096

Figure 7: Example forecast result for site with latest known volume from NRCS

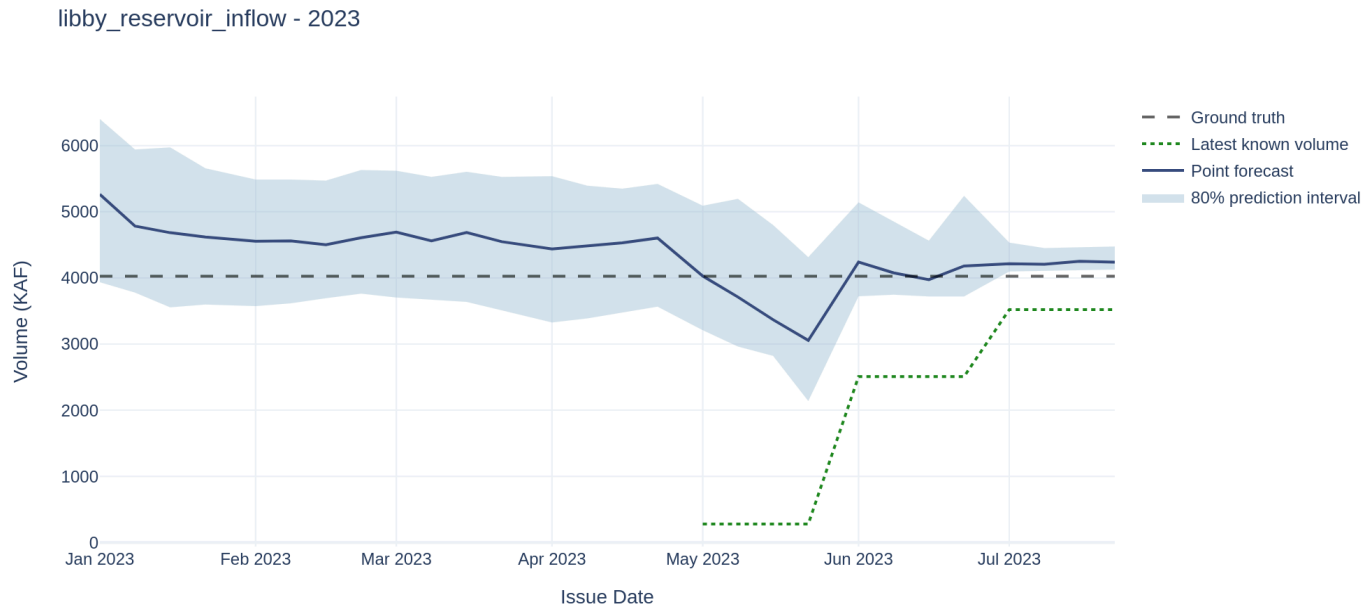


Figure 8: Example forecast result for site with latest known volume from NRCS/USGS/USBR

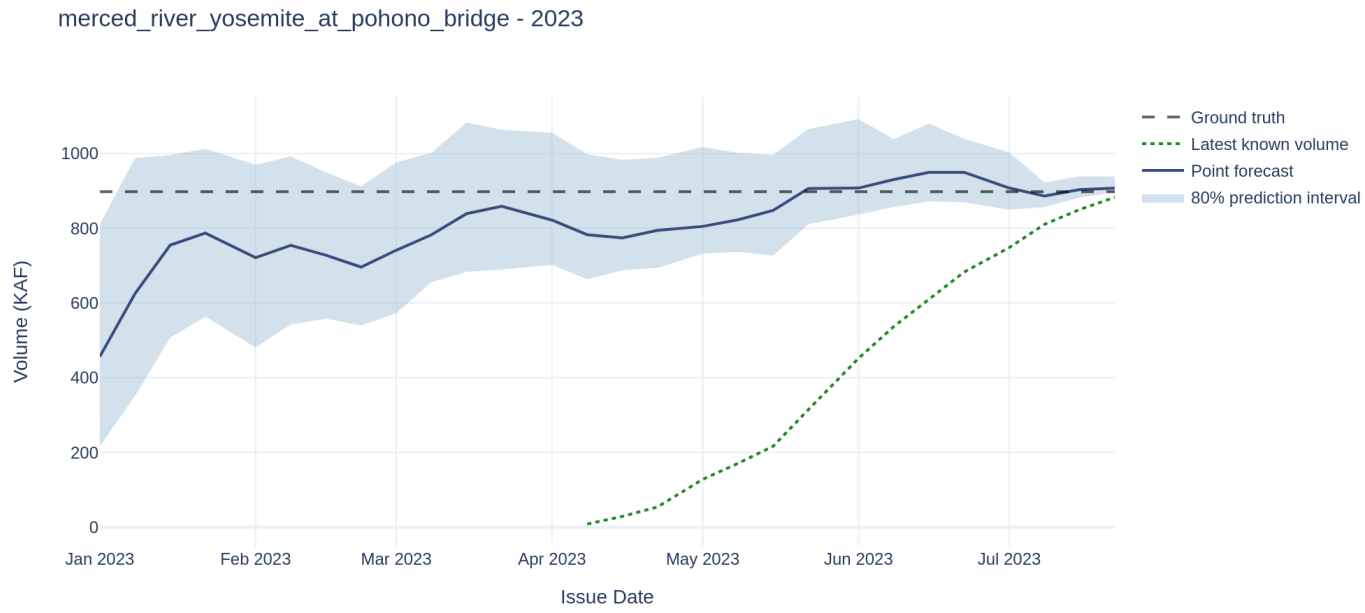
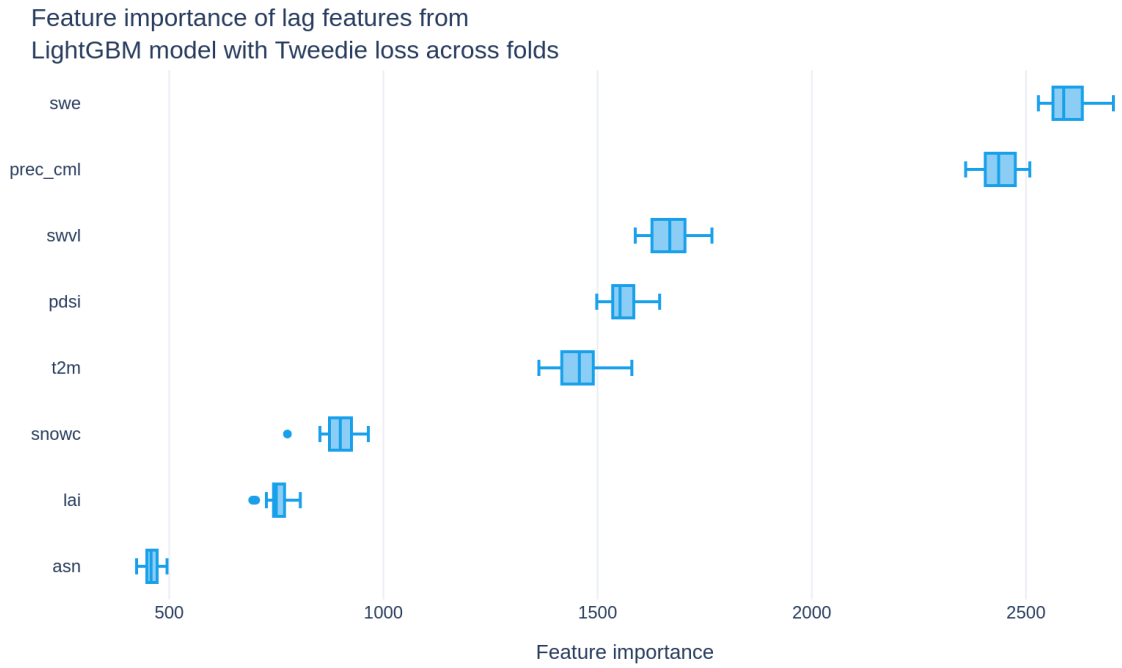


Figure 9: Feature importance of LightGBM model with Tweedie loss across folds



References

1. https://stats.stackexchange.com/questions/113994/how-to-choose-the-training-cross-validation-and-test-set-sizes-for-small-sampl
2. Herbert, Z. C., Asghar, Z., & Oroza, C. A. (2021). Long-term reservoir inflow forecasts: Enhanced water supply and inflow volume accuracy using deep learning. Journal of Hydrology, 601, 126676. https://doi.org/10.1016/j.jhydrol.2021.126676
3. Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via Machine-Learning applied to Large-Sample datasets. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.08456
4. Grillakis, M., Koutroulis, A., & Tsanis, I. (2018). Improving seasonal forecasts for basin scale hydrological applications. Water, 10(11), 1593. https://doi.org/10.3390/w10111593
5. Montero-Manso, P., & Hyndman, R. J. (2021). Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality. arXiv preprint. https://doi.org/10.48550/arXiv.2008.00444