Water Supply Forecast Rodeo - Hindcast Model Report

Author: Rasyid Ridha (rasyidstat)
Placement: 1st Hindcast LB
Model ID: `lgb_sweK9L2S1_diffp_S4_m3_ff_ens9_pfs`
Model date: 2023-12-19

Summary

We use ensembles 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 SWE and cumulative precipitation averaged within and near the basins from top $K = 9$ sites as the input of the model. Using additional features can cause overfitting easily due to limited training sample size for each site and lack of signal/information loss due to data aggregation. We use synthetic data generation to mitigate small sample size problems and increase 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. In addition, LOOCV for even years 2004-2023 is used for evaluating the model internally and it shows good consistency between validation and test years set.

We started with a climatological baseline which is the median, percentile 10% and 90% of historical seasonal water supply volume before 2005. Using this benchmark, we noticed that there's a big gap between validation and test set which means that water years in the test set are wetter compared to the validation set. Our best models shorten the gap with average MPL score ~84 KAF in validation set and ~88 KAF in test set, 36%-54% improvement compared to climatological baseline. Details about model experimentation log can be accessed in Appendix: Table 1.

We also tried other data sources such as teleconnection indices, PDSI and RCC-ACIS PRISM (precipitation and temperature) but the forecast skill is not better even though there's a correlation and signal from training data. Recently, we found out that incorporating PDSI as an ensemble member improves forecast skills, and this improvement occurs during the dry period.

Keywords: LightGBM, Tweedie, ensemble, SNOTEL, 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.

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 based on validation and test scores. 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 each year with different training iterations but it does not reflect well on test years. Instead, we select a fixed number of training iterations which is generalized to any water year even though it is suboptimal on individual water years. It also gives us a better picture of the gap between validation and test set.

Synthetic data generation

On this challenge, we face the issue of small sample sizes for each site. 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. Using a larger number of S synthetic years can lead to no improvement or even worse results based on the validation score.

This approach significantly improves forecast skill, prediction interval reliability, and generalizability. It's estimated that there is 1-3 KAF improvement using this approach.

Ensemble

For each quantile, we combine forecast output from 9 base LightGBM models with different random seeds by aggregating them with averaging for point forecast and percentile 10% and 90% for quantile forecast. Before that, we do simple averaging from forecasts generated from 10-fold validation years. It's estimated that there is 1-4 KAF improvement using this approach. In addition, we already get a calibrated prediction interval with coverage of 0.78-0.82. Using the base model, we still need to do post-processing calibration since the prediction interval is overconfident with coverage of 0.67-0.72.

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

Rather than ensembling models from different 10-fold years x 9 random seeds = 90 models, we can use full single model training or reduce ensemble members with a tradeoff on decreased forecast skill, 1-4 KAF based on validation and test score. The benefit is a significant decrease in training time.

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. 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. 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 to experiment faster.

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. In addition, since the sample size is small, we want to avoid overfitting by having fewer features (ideally $p < n$) and only include the most significant one in the model.

Data sources

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 forecast skill
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 SWE and precipitation in-situ measurement include long history data which satisfy criteria (1) 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
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 results.
Drought condition	gridMET PDSI	E	gridMET PDSI includes long history data which satisfy criteria (1). For the Hindcast stage, we do not use it since previously, we did not see significant improvement in forecast skill. In 2023-12-22, we found out that incorporating PDSI as an ensemble member improves forecast skill, and this improvement occurs during the dry period.
Weather	RCC-ACIS PRISM monthly	E	PRISM gridded data includes long history data which satisfies criteria (1). However, there is no

	precipitation and temperature		significant improvement in forecast skills. This might be because we use a monthly scale dataset instead of a daily scale. In addition, we use bounding boxes of drainage basins rather than exact polygon clipping which reduce the accuracy of estimation.
Snowpack	CDEC SWE	C	Currently, we use the nearest SNOTEL station as a snow proxy for 3 sites in the California basin which ranks lower scores as seen in Table 2. Using CDEC data might give better snowpack estimates since it's located within the basin.
Snowpack	SNODAS	C	We do not use these datasets since it does not fulfill the criteria (1). However, we can consider experimenting with these datasets for the Forecast/Overall stage as there are more training years. SNODAS data might be useful when SNOTEL station data is limited. For example, we notice that Skagit Ross Reservoir and Libby Reservoir intersect with Canada and there are no SNOTEL stations in that location
Snowpack	MODIS Snow Cover	C	
Vegetation	MODIS Vegetation Indices	C	
Weather forecast	CPC Seasonal Outlooks	C	Incorporating forecasts of precipitation and temperature into the model might have potential benefits such as improved long lead time forecast skill, more robust quantile forecast and better view of future snow accumulation/melting process. However, it depends on the skill of the forecast product itself and whether it exhibits consistent multi-year skill. Further assessment needs to be done to understand the improvement in forecast skills.
Weather forecast	Seasonal meteorological forecasts from Copernicus	C	

Feature engineering

SNOTEL SWE and precipitation lag features

For each issue date, we use lagged data of SNOTEL SWE and cumulative precipitation averaged within and near the basins from top $K = 9$ sites. By using lagged data, we ensure that only previous known values are used before the issue date. The chosen lags are $t - 1$, $t - 8$ and $t - 15$. Using more lags can lead to no improvement or even worse results because of overfitting based on the validation score.

Pair between basins and SNOTEL 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. Top $K = 9$ SNOTEL sites for each basin are chosen based on the highest coefficient of determination between SWE and the target variable from the training dataset.

Static 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. The forecast skill improvement with static features is minimal since we only have 26 basins in the training dataset. Training with more basins can bring more potential benefits of static features, enabling cross-learning between similar basin characteristics and improving model generalizability (Kratzert et al, 2019).[3]

Target preprocessing
We apply target preprocessing by differencing the ground truth with the latest known naturalized flow. This approach significantly reduces error for issue dates within the seasonal month target (Apr-Jul). For example, in May, we subtract the original target with April naturalized flow. In June, we subtract the original target from the sum of April-May naturalized flow, and so on. The preprocessed target is used as the final target when training the model and we do the inverse transformation for inferencing by summing back the forecast with known naturalized flow.

To further improve the performance, we also incorporate the latest known daily USGS and USBR observed flow from sites with minimal impairment. There are selected 14 sites that have a ratio of historical observed flow and naturalized flow near 1. Before that, we convert the unit of the metric from cfs (cubic feet per second) to KAF (kilo acre-feet).

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.

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. 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 snow, prediction might be more difficult as we rely on historical precipitation only and have many unknowns on future weather. It would be interesting to see more detailed explanations in the Model Explainability part and link it with the known hydrological process.

There might be some missing equations in our model, e.g. how the basin responds to the precipitation, land, vegetation, soil condition, temperature, evapotranspiration, and other complex hydrological processes. However, some hydrological processes might be too detailed

and can only be revealed or explained at a more granular level (hourly, daily). Data-driven models built in seasonal granularity might be not able to capture this importance or link it with the known hydrological process.

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

Initially, we defined the model evaluation scheme by splitting the dataset into training (1980-2005), validation (even years 2005-2023) and test set (odd years 2005-2023). This scheme is relatively fast because we only need to train the model once. During the training, we implemented early stopping based on the validation set score. However, we noticed that validation and test scores are not consistent. Training with early stopping will give the best score in the validation set but it does not reflect well in the test set. In addition, using this scheme will reduce sample size a lot resulting in poorer forecast skills.

To have a more robust evaluation scheme, we changed our strategy by using LOOCV (leave-one-out cross-validation). We use a single year as validation for each even year 2004-2023. This scheme increases training time by 10x but with benefits on more robust validation and test scores. In addition, this scheme mimics evaluation for the Hindcast stage (10 years test set of odd years 2004-2023) and also mimics evaluation for the Forecast stage where there is only one year used for validation. We also disable early stopping and define fixed parameters to have more robust results and avoid overfitting in a single validation year. Later on, we realize that LOOCV is also included as the scoring metric for the Overall Stage which is aligned with our evaluation scheme here.

Training process

For each validation year, we train a single LightGBM model for all basins and issue dates with 4 different loss functions (Tweedie, quantile with alpha 0.1, 0.5 and 0.9). The training dataset is formed by excluding a single validation year of interest and all test years. This computation below only uses the training dataset to avoid data leakage and improve variety:

- Selection of pair between SNOTEL 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 improves generalizability. We also cached pairs between SNOTEL sites and basins to be used later for inference for the test years. In the end, for each validation year, we will form a dataset as follows:

- ~100K rows of training dataset (~150 samples for each basin and issue dates)
    - ~20K rows original dataset (~30 samples for each basin and issue dates, training years mostly start from 1980 based on SNOTEL sites availability)
    - ~80K rows synthetic dataset (~120 samples for each basin and issue dates)
- 724 rows of validation dataset

For a single experiment of the training process above, there will be 10 single LightGBM models with Tweedie loss from 10 validation years. Hence, there will be 10 forecasts generated for a single issue date and basin that applied to the test set. The point forecast for a single experiment is the average of all 10 forecasts. This also applied the same for 0.10 and 0.90 quantile forecasts with an additional 20 single LightGBM models with quantile loss.

To further improve forecast skills, we also train an additional 8 base models with different random seeds. Finally, we do the ensemble mentioned in Algorithm and Architecture Selection to generate the final forecast.

Endnote

Our final models for the Hindcast Stage get the 1st placement in the leaderboard. It will be interesting to know how the forecast skill performance compares with the existing operational benchmark. Given more time to work on this challenge, there is an opportunity to improve the forecast skill. We recommend to:

- Incorporate other snowpack data (CDEC, SNODAS, MODIS) to have more accurate snowpack estimates
- Incorporate weather forecast products to improve long lead time forecast skills
- Experiment with two-stage models (meta-learning) trained on different data sources, features and configurations
- Train on additional basins to account for small sample size problems and improve generalizability
- Experiment with sub-model, semi-distributed and bottom-up forecasting settings (daily and monthly forecasting, large basin divided into sub-basins)

Our models indicate generalized results for different conditions of wet and dry years based on a small gap between validation and test forecast skills. Further post-evaluation of the models will bring additional insights for the water managers. Several next steps are to:

- Assess model generalization in wet years and dry years. Identify where the model gives poorer results and potential gaps or biases
- Understand feature importance from the model and how the forecast evolves from different issue dates
- Understand which features are dominant in different conditions (wet years vs dry years, snow-dominated vs semi-arid regions) and how they are linked with the hydrological process

Machine Specifications

- CPU: Core i5
- RAM: 8GB
- Training duration: ~2 hours for all 360 models (10-fold years x 9 random seeds x 4 losses)
- Inference duration: less than 5 minutes for 10 years test set (not including data download time)

Appendix

Table 1: Model experimentation log

Date	Model ID with a short description	Val	Test
2023-10-28	**`climatological_baseline`**	131.72	192.46
2023-11-01	**`lgb_swe_tp3`** Initial model with SWE	99.35	118.57
2023-11-07	**`lgb_sweonly_tp3_diff_v3_wd5`** Adjust model params, use target diff preprocessing, calibrate PI	96.31	102.06
2023-11-26	**`lgb_sweK9L2_diff_S0_m02`** Tweedie loss, adjust lag, adjust SNOTEL sites selection approach, 10Y LOOCV instead of train/val/test split, bigger train size and more robust validation	91.82	96.62
2023-11-27	**`lgb_sweK9L2_diff_S4_m12`** Use synthetic data, increase training size by 5x	90.57	93.92
2023-12-11	**`lgb_sweK9L2S1_diff_S4_m3`** Update monthly naturalized flow to use recent lag t-1 from t-2 for 1st week of the issue date	88.67	92.37
2023-12-17	**`lgb_sweK9L2S1_diff_S4_m3_ff`** Update SWE to use recent lag t-1 from t-2, only use available SNOTEL sites in competition runtime	88.12	91.85
2023-12-17	**`lgb_sweK9L2S1_diff_S4_m3_ff_ens9`**	86.69	91.03
2023-12-18	**`lgb_sweK9L2S1_diffp_S4_m3_ff`** Incorporate daily USGS and USBR observed flow from sites with minimal impairment	84.92	89.24
2023-12-18	**`lgb_sweK9L2S1_diffp_S4_m3_ff_ens9`**	83.63	88.06
2023-12-19	**`lgb_sweK9L2S1_diffp_S4_m3_ff_ens9_pfs`** ✅ For sites without USGS and USBR observed flow, the forecast in	83.66	87.76

	weeks 2-4 is not better compared to week 1 in months 5-7. Using the previous forecast on week 1 of the month slightly improves the score	+36.5% vs baseline	+54.4% vs baseline
2023-12-21	Hindcast Stage Submission Deadline		
2023-12-22	**`lgb_sweK9L2S1_pdsi_diffp_S4_m3_ff_ens18_pfs`**	82.16	86.95

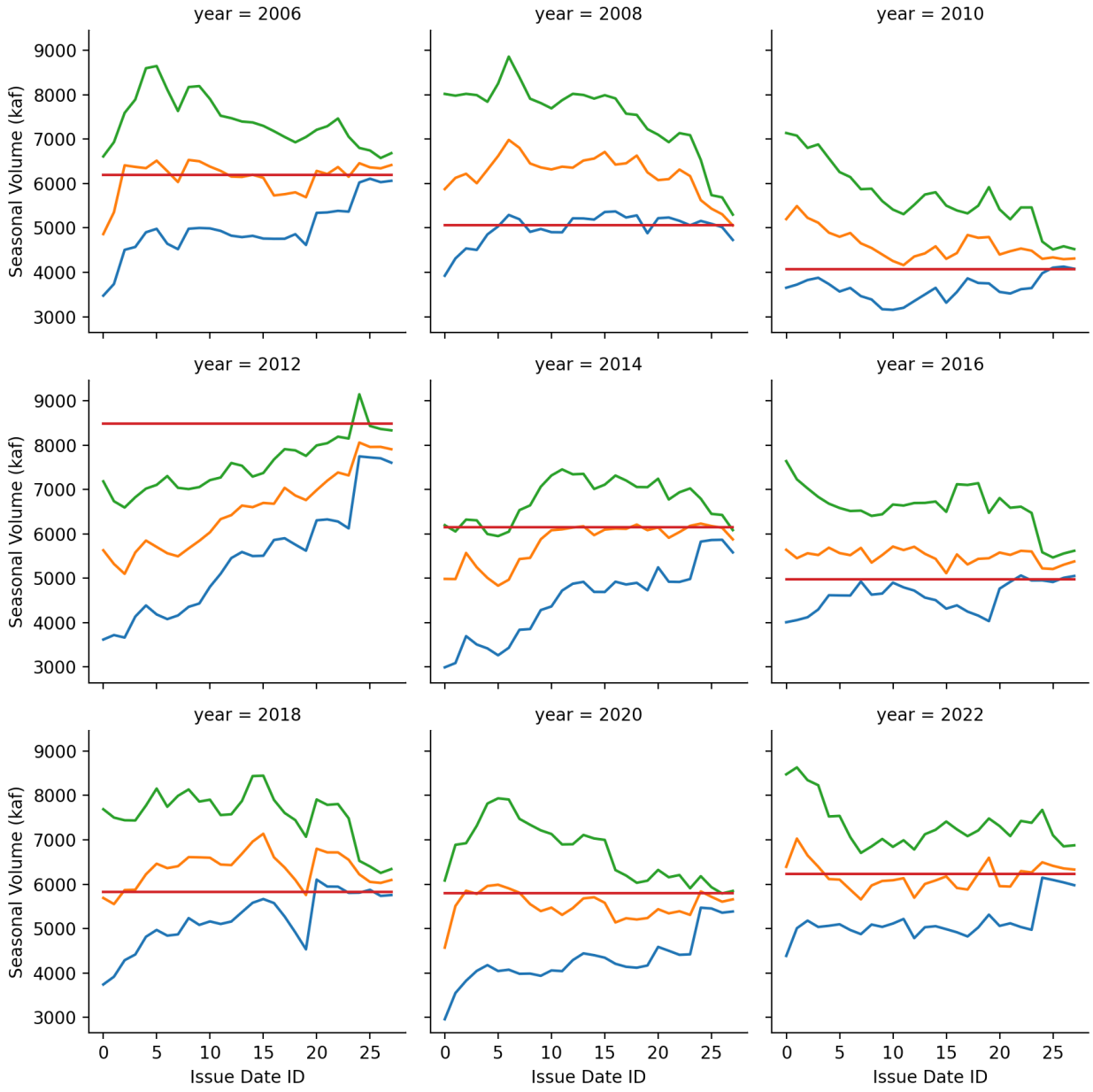
Table 2: Validation score per site

Validation score per site for all issue dates sorted by smallest normalized MPL (normalized by the average ground truth value of water supply volume)

site_id	mpl	mpl10	mpl50	mpl90	int_cvr	nmpl
stehekin_r_at_stehekin	42.152696	29.696864	60.031611	36.729612	0.889286	0.056201
hungry_horse_reservoir_inflow	137.191865	81.777182	215.804995	113.993419	0.792857	0.062117
snake_r_nr_heise	204.864358	133.232617	336.264960	145.095496	0.771429	0.064683
boise_r_nr_boise	92.052085	56.925284	144.978208	74.252764	0.853571	0.076253
libby_reservoir_inflow	464.242099	241.445710	706.341329	444.939259	0.775000	0.081581
weber_r_nr_oakley	7.940027	4.196735	12.675226	6.948120	0.839286	0.085670
yampa_r_nr_maybell	76.377436	44.397182	127.747617	56.987511	0.935714	0.091618
skagit_ross_reservoir	131.347193	68.654821	204.988910	120.397847	0.664286	0.092885
ruedi_reservoir_inflow	11.370981	6.103665	17.167287	10.841991	0.867857	0.096474
green_r_bl_howard_a_hanson_dam	25.397095	15.231630	41.611172	19.348484	0.750000	0.096925
animas_r_at_durango	30.124358	16.268144	47.215043	26.889886	0.839286	0.098069
fontenelle_reservoir_inflow	67.864425	38.096501	110.744181	54.752594	0.925000	0.104735
dillon_reservoir_inflow	14.743281	7.897170	23.779402	12.553271	0.846429	0.105137
missouri_r_at_toston	198.662286	127.719926	311.486427	156.780504	0.878571	0.107627
merced_river_yosemite_at_pohono_bridge	35.381370	25.107851	55.169527	25.866734	0.814286	0.116130
pecos_r_nr_pecos	4.474079	2.318540	6.734965	4.368732	0.875000	0.121143
taylor_park_reservoir_inflow	10.273656	6.329075	15.551398	8.940496	0.775000	0.134820
colville_r_at_kettle_falls	17.674034	10.835011	26.910340	15.276751	0.789286	0.138380
detroit_lake_inflow	74.494474	41.311094	115.712185	66.460142	0.629167	0.142394
american_river_folsom_lake	150.859427	76.613561	257.791724	118.172997	0.785714	0.145022
san_joaquin_river_millerton_reservoir	159.672524	98.146984	276.198797	104.671791	0.746429	0.167203
boysen_reservoir_inflow	102.031050	56.770034	163.984460	85.338655	0.753571	0.167697
pueblo_reservoir_inflow	47.189352	16.741391	73.295313	51.531351	0.789286	0.174539
virgin_r_at_virtin	6.455161	2.961378	10.082884	6.321222	0.821429	0.178131
owyhee_r_bl_owyhee_dam	50.622820	22.281454	81.254351	48.332657	0.796429	0.195470
sweetwater_r_nr_alcova	10.981045	4.151523	18.536103	10.255508	0.867857	0.258893

Figure 1: Sample forecast result

Sample forecast result for Libby Reservoir Inflow for validation years (even years 2004-2023, excluding 2004)



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