Lab for Spatial Informatics

BTP - 1

Instructor: Prof. Rehana Shaik

Students: Gokulraj R, Joel Alex

May 2023

Final Report

Overview

1 Introduction

Predicting reservoir inflows is important to ensure the efficient and safe management of water resources, as well as to mitigate the impacts of floods and droughts.
The paper[**1**] establishes a simple and effective framework to combine various data-driven machine learning (ML) algorithms for short-range reservoir inflow forecasting

1.1 Datasets

1. **CLE Reservoir**: The CLE reservoir plays a vital role in diverting water from Northern California to the Central and Southern parts of California. Daily inflow, evaporation, and daily precipitation data were collected from the California Department of Water Resources from November 1962 to August 2020

2. **Bhadra Reservoir**: The Bhadra Reservoir, located on the Bhadra River, is a multipurpose reservoir providing irrigation, hydropower, and low-flow augmentation requirement. Daily inflow, evaporation, and daily precipitation data were collected from the Advanced Centre for Integrated Water Resources Management (ACIWRM), India from June 2004 to May 2018.

3. **Climate Indices**: The 10 climate indices considered in the study were Arctic Oscillation (AO)[**2**], East Pacific/North Pacific Oscillation (EPO)[**2**], North Atlantic Oscillation (NAO)[**2**], NINO1 + 2 (Extreme Eastern Tropical Pacific SST)[**3**], NINO3 (Eastern Tropical Pacific SST)[**3**], NINO4 (Central Tropical Pacific SST)[**3**], NINO34(East Central Tropical Pacific SST)[**3**], PNA (Pacific North American Index)[**2**], SOI (Southern Oscillation Index)[**4**], WP (Western Pacific index)[**2**].

1.2 ML Models

Four ML models Random Forest, Gradient Boosting Regressor, K Nearest Neighbours, and Long Short Term Memory (LSTM), are combined with sufficient tests of performance measures of the models. An ensemble model using an arithmetic weighted voting ensemble method to quantify forecasting uncertainty by combining the inflow results of each ML model and the

highest vote is chosen based on the weights assigned to them.

2 Objectives

1. Increase the accuracy of the developed framework in comparison to forecasted inflows via data pre-processing, parameter optimization, and post-processing.
2. Improve the predictability of extreme values

3 Methodology

3.1 Dataset Collection

The Datasets used in the paper were outdated so our first task was to collect updated datasets. Since some of the links in the paper were invalid, we were required to obtain the following datasets ourselves:

1. **CLE Reservoir**: Used California Data Exchange Centre's Database [**5**] to query Inflow, Evaporation, and Precipitation data from 1962 to 2023. The site only had recent data available so we had to combine the data with the previous dataset after making the format uniform.
2. **Bhadra Reservoir**: Inflow, Evaporation, and Precipitation data were collected from 6 rain gauge stations, on a request basis, from June 2004 to May 2018 from ACIWRM.
3. **Climate Indices**: 10 unique climate Indices, AO, EPO, NAO, NINO1+2, NINO3, NINO3+4, NINO4, PNA, SOI and WPO, were collected from the Physical Sciences Laboratory of the NOAA[**2**], the KNMI Climate Explorer tool[**3**], and The Long Paddock[**4**]

3.2 Code Quality Improvements

Improved the general quality of the code to make it more extendable and flexible. Performed the following improvements:

1. *Originally, the datasets are in different formats but need to be used for the same tasks.* We Defined functions for converting the datasets into the same format.
2. *Originally, the ranges for data extraction from a common timeline were hard coded and a change in the datasets would require us to manually go through every dataset and find the*

indices of the overlapping time period for each dataset and update them in the code. This Data extraction from a common timeline has been automated.

3. *Originally, the code for running models with different hyper-parameters was repetitive and hard to compare and test.* We modularized the code for running models with different hyperparameters and ran ablation studies to determine the best hyperparameter for each model.

3.3 Data Processing

- Common time frame automatically selected for different variables
- Data gap handling using linear interpolation has also been automated by checking for data gaps in all the datasets and immediately inserting missing values
- 3 Day lagged data is generated for all the datasets after considering their autocorrelation and partial autocorrelation coefficients as well as considering inputs used in a similar paper [**7**].
- The Pearson Correlation of Features has been computed and Similarity Matrix is plotted to manually perform feature reduction.
- Decision variables were ranked and pruned using the Gini Diversity index

3.4 Parameter Selection

- Ideally only 2 days of lagged data should be considered considering autocorrelation and partial autocorrelation. This is visualized using a similarity matrix using Pearson correlation score between the main features.
- LSTM is tested on lagged data as well as non-lagged data since LSTM, unlike the other classical models used in this paper, can store information about the previously passed features and hence doesn't need to be passed lagged data separately.
- Precipitation data contributes to inflow prediction significantly more than the other parameters and this is shown by the Similarity matrix and supported by Mao et al.[**6**] and Lee et al.[**7**].
- Since Code has been modularised, an ablation study has been conducted on the less

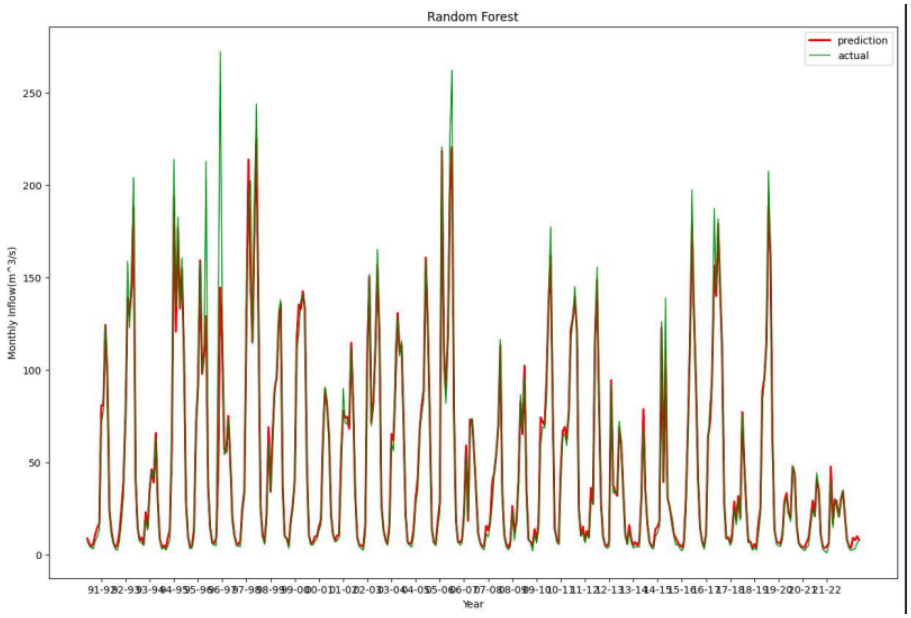

Figure 1: Random Forest model on CLE dataset

computationally intensive models using partial datasets to choose more optimal hyper-parameters.

- Weights were originally assigned in a 4:3:2:1 ratio based on performance but no weights are assigned proportional to their accuracy scores. This way the voting is more representative of the model's actual ability to predict

4 Results

4.1 CLE Reservoir

Since the dataset for the CLE reservoir is huge, the selection of the best hyper-parameters for the ML models is done by running the models for a hand-picked set of values of hyper-parameters and determining the hyper-parameter that provides minimum RMSE. The predicted monthly inflow values by the models using the best set of hyper-parameters are then plotted against the actual values in the figures 1, 2, and 3. Figure 4 shows the comparison of original metrics presented in the paper [**1**] given to us, and the improved metrics obtained using the data processing and the parameter selection mentioned in sections 3.3 and 3.4 on the updated CLE dataset.

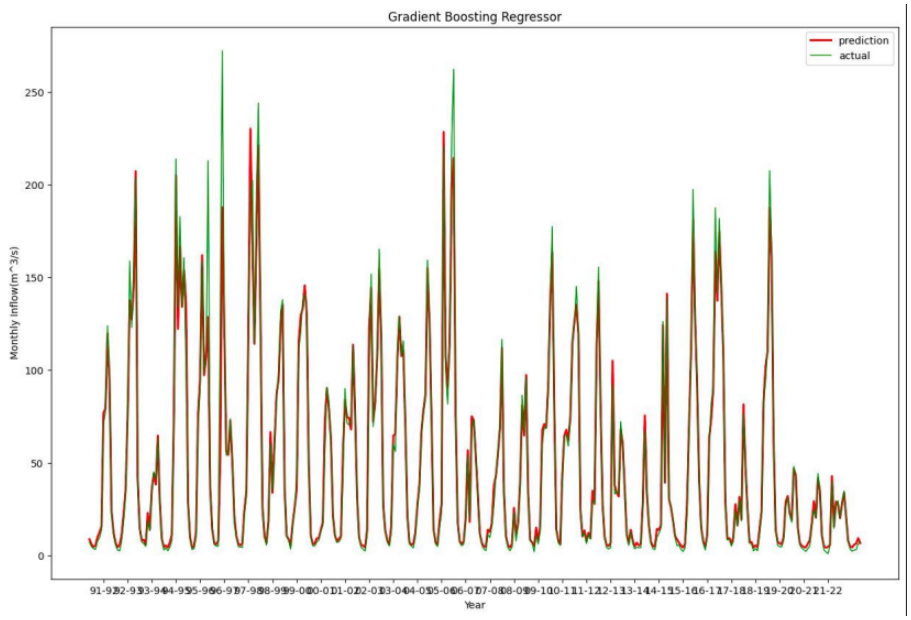

Figure 2: GBR model on CLE dataset

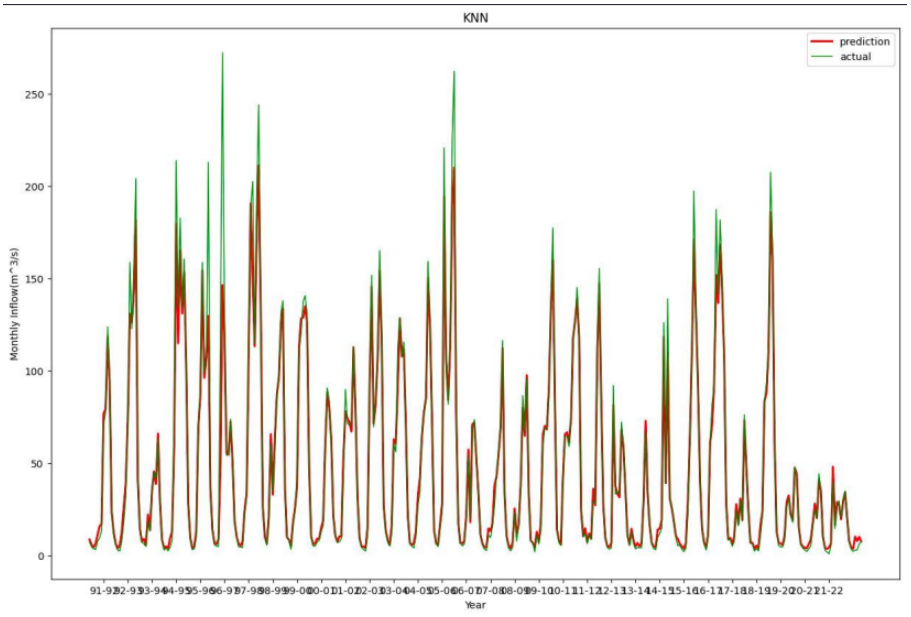

Figure 3: kNN Regressor model on CLE dataset

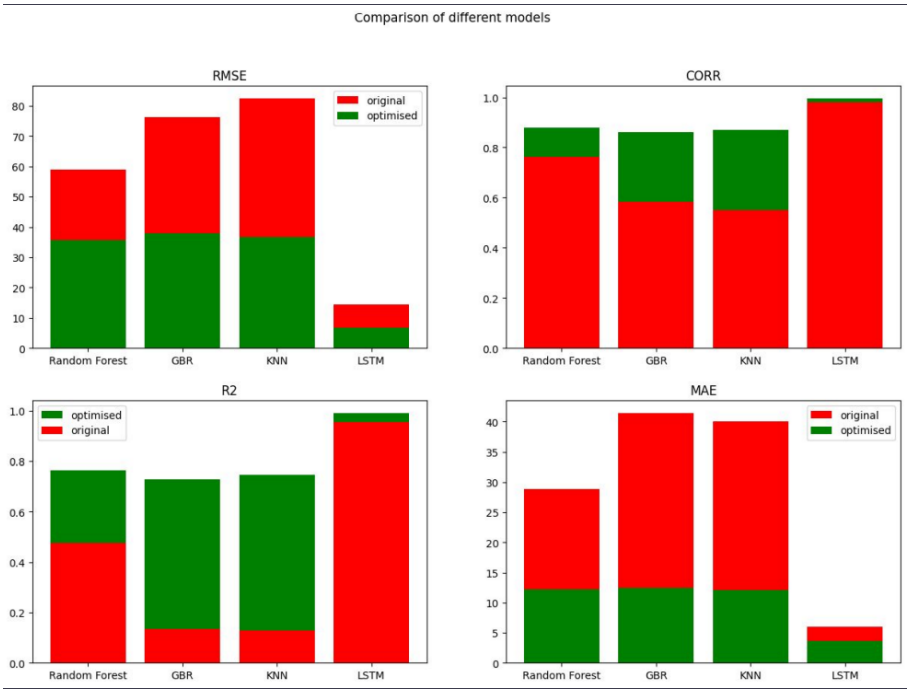

Figure 4: Comparison between original and improved metrics for CLE dataset

4.2 Bhadra Reservoir

Since the dataset for the Bhadra reservoir is relatively small, the best hyper-parameters are selected by running the models in a loop over a range of hyper-parameter values and determining the hyper-parameter that provides minimum RMSE. Results observed for predicted vs. actual monthly inflow for the Bhadra dataset are similar to the results for the CLE dataset. The predictions of all the models improved significantly. Figure 5 shows the comparison of original metrics presented in the paper [**1**] given to us, and the improved metrics obtained using the data processing and the parameter selection mentioned in sections 3.3 and 3.4 on the updated Bhadra dataset.

We notice that LSTM presented in the paper [**1**] clearly outperforms the other three models on both datasets. Using non-lagged data for the LSTM improves its performance further.

5 Challenges Faced

- ***Acquiring the datasets.*** Most links in the paper [**1**] for the dataset were outdated and invalid. We had to manually search and find each of the datasets in the same formats.
- ***Limited Compute.*** The ML models were computationally quite heavy and could not be trained locally. Especially, LSTM required GPU for faster training and we only had

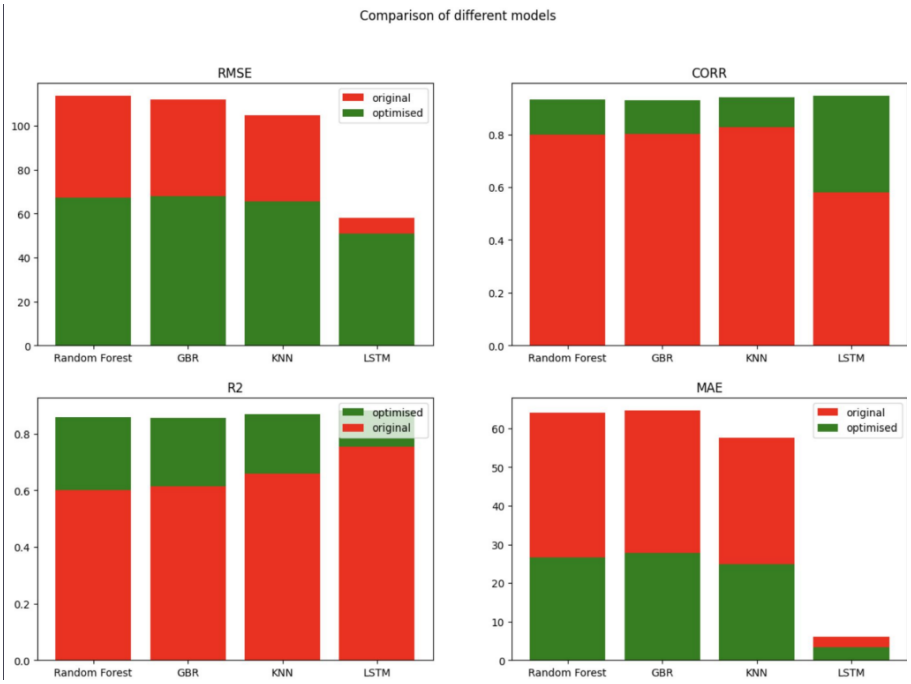

Figure 5: Comparison between original and improved metrics for Bhadra dataset

limited GPU. So, analyzing LSTM was harder than analyzing the other three models.

- ***Processing Tactics.*** Due to the limitations of the dataset and models, it was difficult to find and apply appropriate data processing methods to the given codebase.

6 Conclusion

In conclusion, our project successfully accomplished the goals of increasing the accuracy of the developed framework presented in the paper [**1**], and improving the predictability of extreme values in comparison to forecasted inflows via data pre-processing, parameter optimization, and post-processing.

Pre-processing is done in terms of better processing of the datasets to get them in the same format, varying the size of lagged data, and manual feature reduction by using Pearson Correlation of features. Parameter optimization has been done for the newer updated dataset for both CLE and Bhadra. This is accomplished by minimizing RMSE over a set of values of parameters for each model. Post-processing is improved by using assigning weights to the models for the ensemble proportional to their individual accuracy scores, rather than the ranks of the models (4:3:2:1).

Despite being new to the field of reservoir inflows and climatic conditions, we were able to bring

good perspectives to the analysis and improve the performance of the existing model.

The datasets and climatic indices used, the Jupyter notebooks containing the implementation, and the results can be found at:

`https://github.com/Gokulraj-R-002/BTP-1-Reservoir-Inflow-prediction`.

References

[1] Maddu, Rajesh, Indranil Pradhan, Ebrahim Ahmadisharaf, Shailesh Kumar Singh, and Rehana Shaik. *"Short-range reservoir inflow forecasting using hydrological and large-scale atmospheric circulation information."* Journal of Hydrology 612 (2022)

[2] https://psl.noaa.gov/data/timeseries/daily/

[3] https://climexp.knmi.nl/selectdailyindex.cgi?id=someone@somewhere

[4] https://www.longpaddock.qld.gov.au/soi/soi-data-files/

[5] https://cdec.water.ca.gov/dynamicapp/wsSensorData

[6] Mao, T., Wang, G. Zhang, T. Impacts of Climatic Change on Hydrological Regime in the Three-River Headwaters Region, China, 1960-2009. Water Resour Manage 30, 115–131 (2016). https://doi.org/10.1007/s11269-015-1149-x

[7] Lee, S.; Kim, J.; Bae, J.H.; Lee, G.; Yang, D.; Hong, J.; Lim, K.J. Development of Multi-Inflow Prediction Ensemble Model Based on Auto-Sklearn Using Combined Approach: Case Study of Soyang River Dam. Hydrology 2023, 10, 90. https://doi.org/10.3390/hydrology10040090