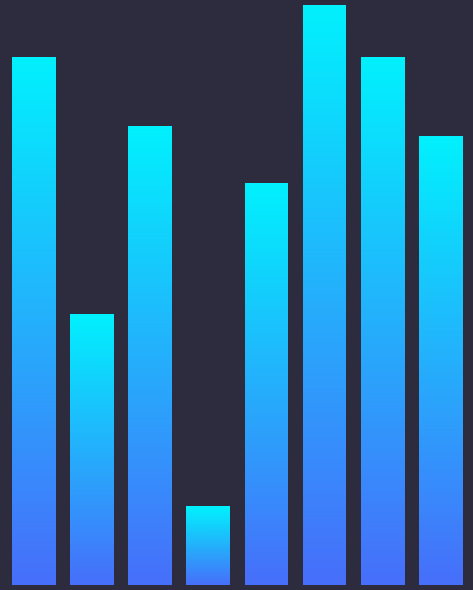
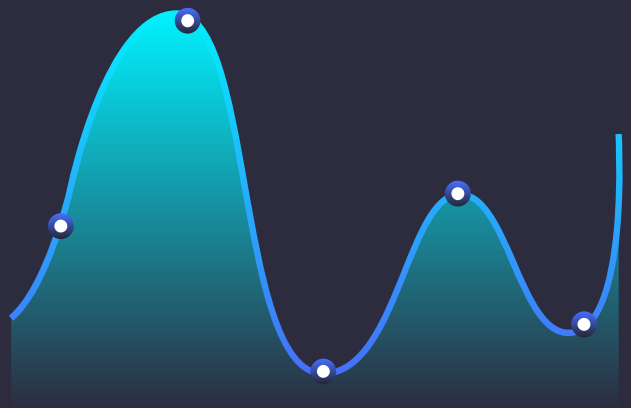
Prediction of Reservoir inflows using ML algorithms

Joel Alex (2020101118)
Gokulraj R (2020102042)



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 establishes a simple and effective framework to combine various data-driven machine learning (ML) algorithms for short-range reservoir inflow forecasting

Datasets Used

CLE Reservoir

The CLE reservoir plays a vital role in diverting water from North California to Central and Southern parts of California and the data had the following features:

- Daily inflow, evaporation, and daily precipitation data was collected from the California Department of Water Resources
- Data was collected from November 1962 to August 2020

Bhadra Reservoir

The Bhadra Reservoir, located on the Bhadra River, is a multipurpose reservoir providing irrigation, hydropower, and low flow augmentation requirement. The data had the following features:

- Daily inflow, evaporation, and daily precipitation data was collected from the Advanced Centre for Integrated Water Resources Management (ACIWRM), India
- Data was collected from June 2004 to May 2018

ML Models used

01 Random Forest

02 Gradient Boosting Regressor

03 K-Nearest Neighbors Regressor

04 Long Short-Term Memory

Random Forest

01

- Random Forest is an ensemble technique capable of performing both regression and classification, using multiple decision trees and a technique called Bootstrap and Aggregation.
- The main idea in Random Forest is to combine multiple decision trees in determining the final output, rather than relying on individual decision trees.

Gradient Boosting Regressor

02

- In Gradient Boosting, each predictor corrects its predecessor's error.
- Each predictor is trained using the residual errors of the previous predictor as labels.
- Gradient Boosting Regressor takes small steps towards the direction that yields better predictions, i.e., a lower variance.

K-Nearest Neighbors Regressor

03

- KNN considers all samples and finds k-nearest neighbors.
- KNN Regressor takes the average of the k-nearest neighbors.
- The distance here is calculated using the Euclidean distance between the sample points.

Long Short-Term Memory

04

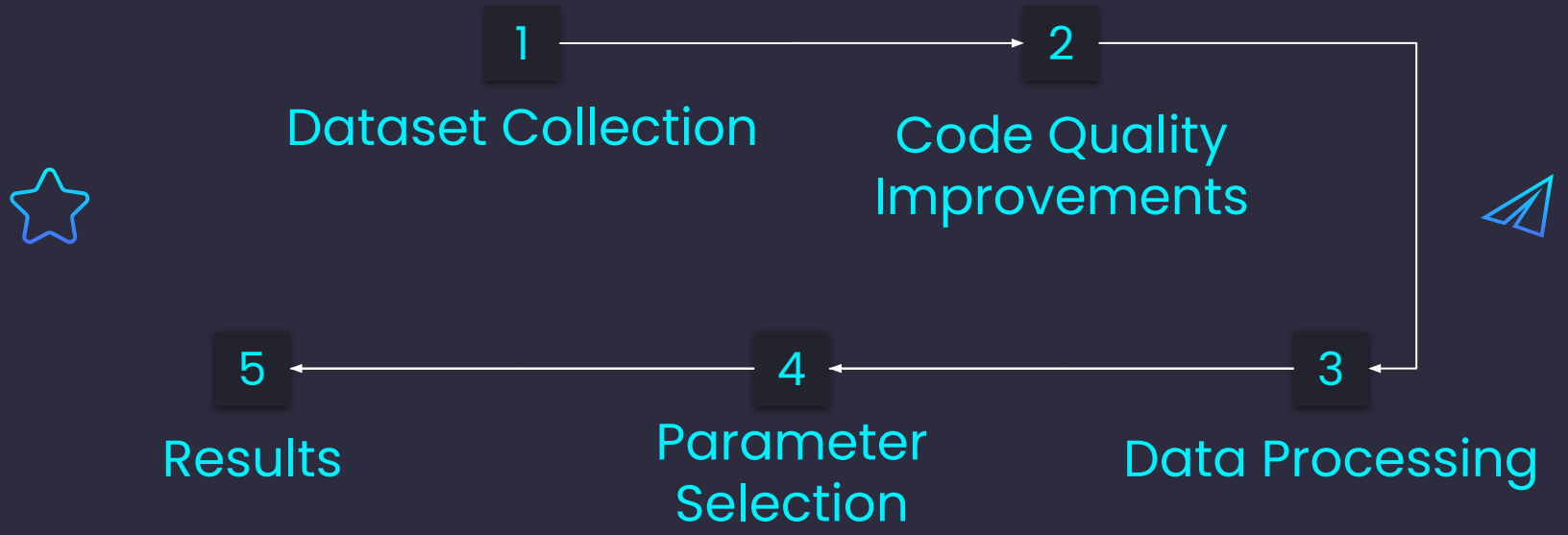
- LSTM is a Recurrent Neural Network (RNN) that is designed to remove the struggle of remembering long-term dependency
- LSTM is designed to remove the struggle of remembering long-term dependency
- LSTM has a chain-like structure, and every module has four neural networks connecting in a special way. The sigmoid layer in LSTM, called “forget gate layer”, decides on keeping or discarding the information.

An ensemble model using a robust weighted voting ensemble method to quantify forecasting uncertainty and to improve the model performance by combining the inflow results of the single ML model and the highest vote is chosen based on the weights assigned to the single ML model.

Objectives

1	Increase accuracy of the developed framework in comparison to forecasted inflows
2	Improve predictability of extreme values

Methodology



Dataset Collection

Datasets are outdated and links in the paper are invalid

Collected 3 Major Datasets

CLE

Used California Data Exchange Centre's Database to query Inflow, Evaporation and Precipitation data from 1962 to 2023

Bhadra

Inflow, Evaporation and Precipitation data were collected from 6 rain gauge stations, on a request basis, from June 2004 to May 2018 from ACIWRM

Climate Indices

10 unique climate Indices AO, EPO, NAO, NINO1+2, NINO3, NINO3+4, NINO4, PNA, SOI, WPO from psl.noaa.gov, climexp.knmi.nl, and longpaddock.qld.gov.au

Code Quality Improvements

Problems

- Datasets are in different formats.
- Hard coded ranges for data extraction from common timeline.
- Repetitive code for running models with different hyper-parameters.

What we have done

- Defined functions for converting the datasets into same format.
- Data extraction from common timeline has been automated.
- Modularized code for running models with different hyperparameters to determine the best hyper-parameter for each model.

Data Processing

Three Main Means of Processing Done

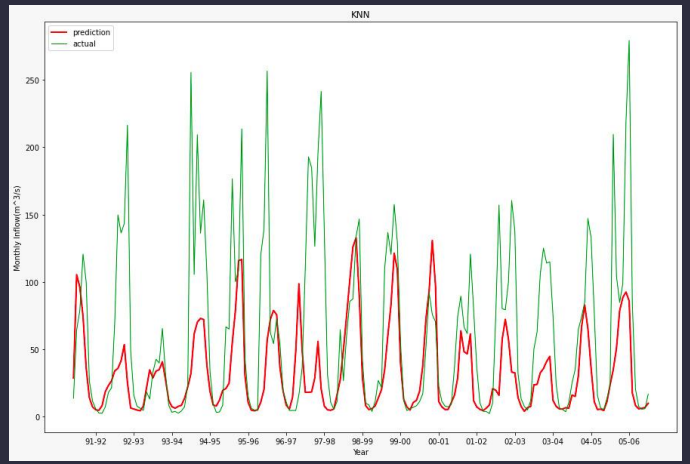
Initial Processing	Time Lag	Feature Reduction
Common time frame automatically selected for different variables	Hydroclimatic series tend to depend on earlier data	High dimensionality of data causes overfitting
10 Climate indices: AO, EPO, NAO, WP, PNA, Nino1 + 2, Nino3, Nino4, Nino34 and SOI	Autocorrelation and partial Autocorrelation coefficients for climate and hydrological variables estimated	Pearson Correlation of Features computed and Similarity Matrix plotted to manually remove irrelevant features
Data gap handling using linear interpolation	3 Day lagged data is generated for CLE, Bhadra and climate indices datasets.	Decision variables were ranked and pruned using the Gini Diversity index

Parameter Selection

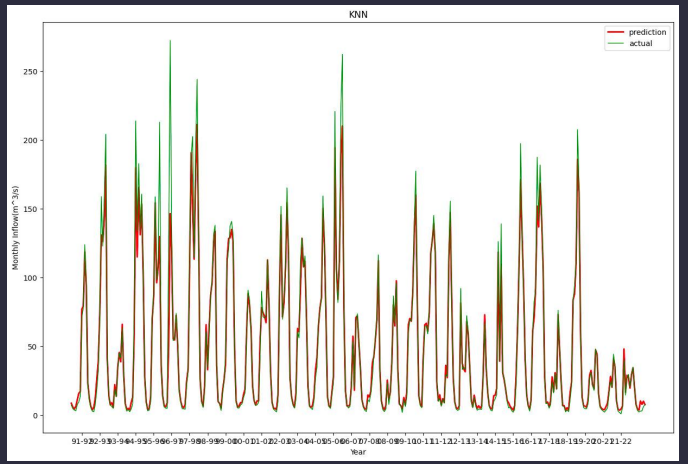
- Ideally only 2 days of lagged data should be considered considering autocorrelation and partial autocorrelation. This is visualised using a similarity matrix
- LSTM is tested on lagged data as well as non-lagged data.
- Precipitation data contributes to inflow prediction significantly more than the other parameters

Results - CLE

Monthly inflow prediction: KNN Original vs KNN Optimised



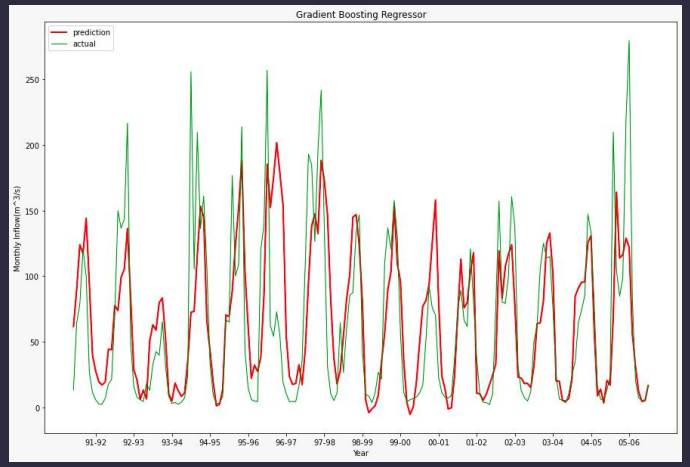
Original



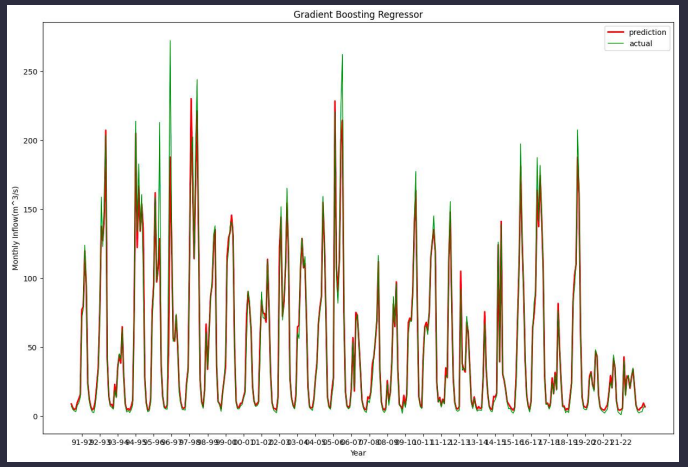
Optimised

Results - CLE

Monthly inflow prediction: GBR Original vs GBR Optimised



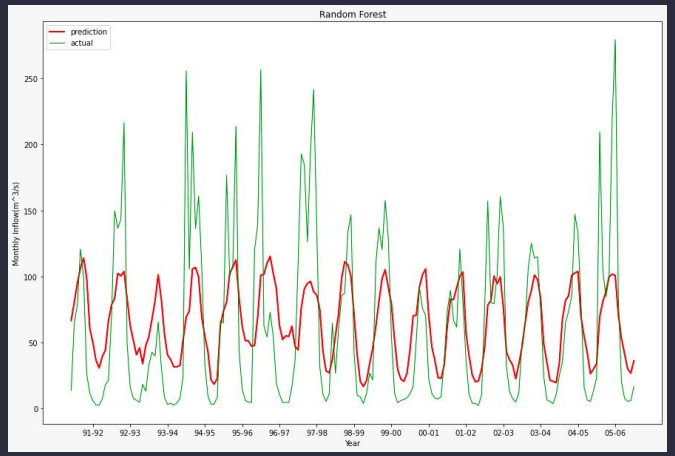
Original



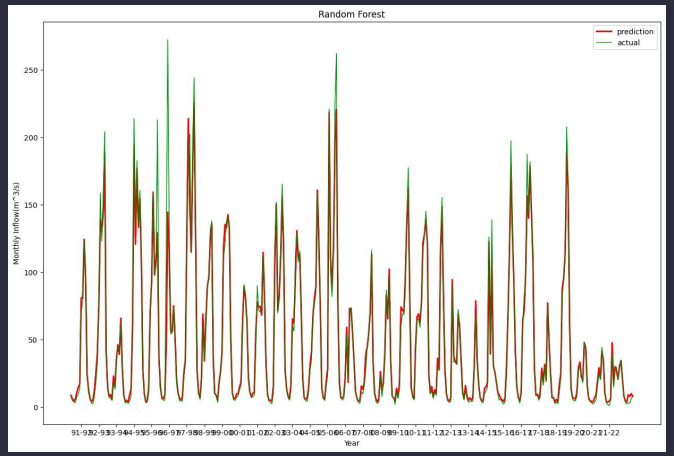
Optimised

Results - CLE

Monthly inflow prediction: Random Forest Original vs Random Forest Optimised



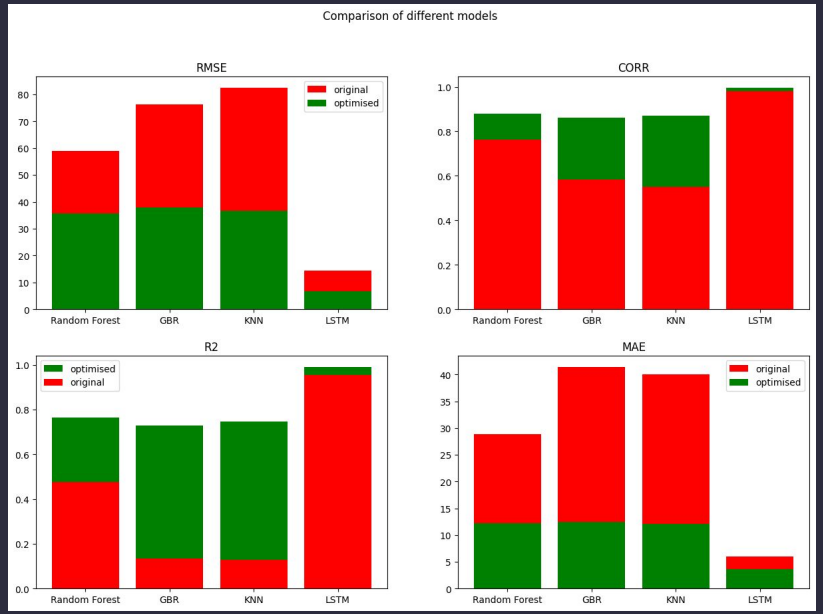
Original



Optimised

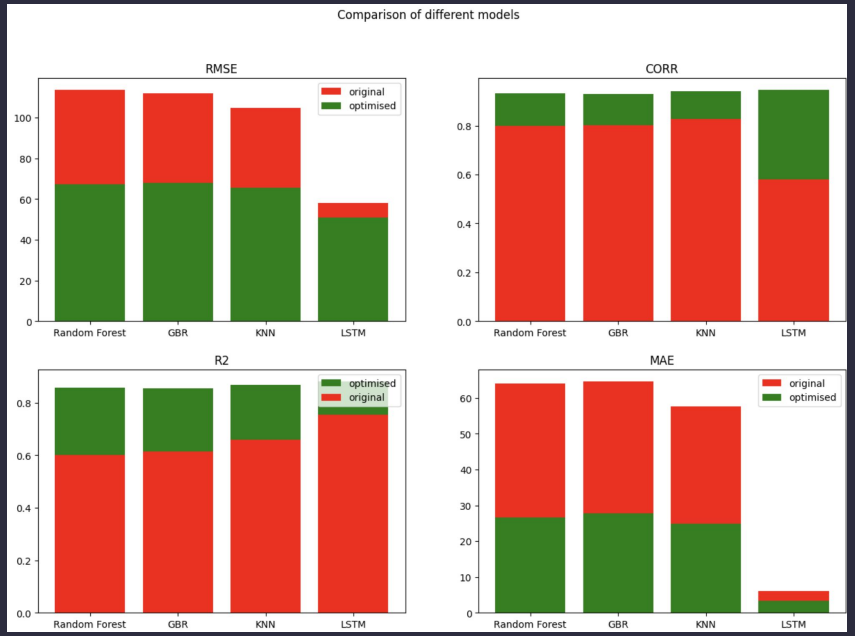
Results - CLE

Overall performance of Optimised code

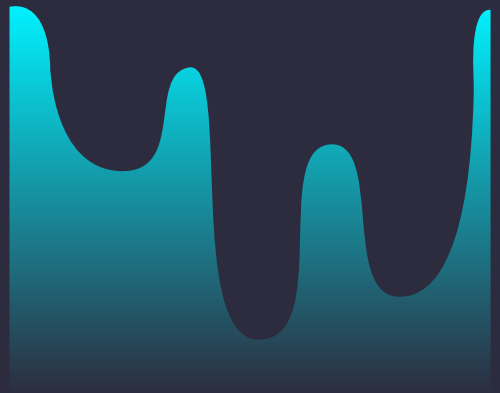


Results - Bhadra

Overall performance of Optimised code



Challenges Faced



Acquiring Datasets

Most links in the paper for the dataset were outdated and invalid. We had to manually search and find each of the datasets in the same formats.

Computation

The ML models were computationally quite heavy and couldn't be trained locally. Initially used Google Colab and then shifted to RC

Processing Tactics

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

References

- 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
- 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
- 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)
- CLE Reservoir Dataset: http://cdec.water.ca.gov/dynamicapp/wsSensorData