Improving Runoff Forecasting Performance Through Deep Learning Error Prediction

John Waugh
Appalachian State University
waughjm@appstate.edu

Abstract – Numerical weather models like the National Water Model (NWM) provide critical runoff forecasts but often exhibit systematic errors, particularly during peak flow events. This project aimed to improve the accuracy of NWM forecasts by training deep learning models to predict and correct these inherent errors. Three distinct neural network architectures – a Simple Recurrent Neural Network (RNN), a Long Short-Term Memory (LSTM) network, and a Transformer network – were developed and evaluated for two separate monitoring stations. The models were trained to predict the NWM's forecast error based on a 24-hour lookback period of NWM output and time-based features. Results show that all three architectures successfully learned to correct the NWM's systematic underestimation of streamflow. The LSTM and Transformer models demonstrated superior performance over the baseline RNN, significantly improving key hydrologic metrics such as the Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) across all 18 forecast lead hours. This study confirms that a deep learning-based error correction is a highly effective strategy for enhancing the reliability of operational runoff forecasts.

I. INTRODUCTION

Accurate streamflow and runoff forecasting are essential for effective water resource management, flood prediction, and ecological monitoring. While the National Water Model (NWM) provides invaluable, high-resolution forecasts across the continental United States, it is known to contain systematic biases. These biases, often manifesting as an underestimation of peak flows and a time-dependent increase in error, can limit the model's utility for critical applications.

The primary objective of this project is to develop and evaluate a post-processing methodology to improve NWM forecast accuracy. Instead of building a new forecasting model from scratch, this work focuses on a more targeted approach: predicting the NWM's inherent error. By training deep learning models to learn the complex, non-linear patterns in the NWM's forecast errors, we can apply these predictions as a correction to the original forecast, yielding a more accurate final product.

To achieve this, we compare three distinct and progressively complex deep learning architectures:

1. A **Simple Recurrent Neural Network (RNN)**, a foundational sequence model.
2. A **Long Short-Term Memory (LSTM)** network, the industry standard for time-series forecasting, designed to overcome the limitations of simple RNNs.
3. A **Transformer** network, a state-of-the-art architecture that uses attention mechanisms to capture complex temporal dependencies.

This comparative analysis will be performed independently for two streamflow stations, allowing for a robust evaluation of model performance and the development of location-specific correction models.

II. DATA

The project utilizes two primary sources of time-series data for two monitoring stations, designated **Station 1** and **Station 2**, spanning from April 2021 to April 2023.

- **USGS Observations:** Hourly observed streamflow data was obtained from the U.S. Geological Survey (USGS). This dataset serves as the "ground truth" for model training and evaluation.
- **NWM Forecasts:** Corresponding NWM forecasts were collected for the same period. These forecasts are initialized at regular intervals and provide predictions for 18 lead hours.

The data was chronologically partitioned into a **training set** (April 2021 - September 2022) and a **test set** (October 2022 - April 2023) to ensure a rigorous and fair evaluation of model performance on unseen data.

An initial exploratory data analysis (EDA) revealed key characteristics of the data at both stations. The streamflow is "flashy," characterized by long periods of low flow punctuated by sharp, high-flow events. Furthermore, a direct comparison showed that the NWM baseline forecast systematically underestimates these peak flow events.

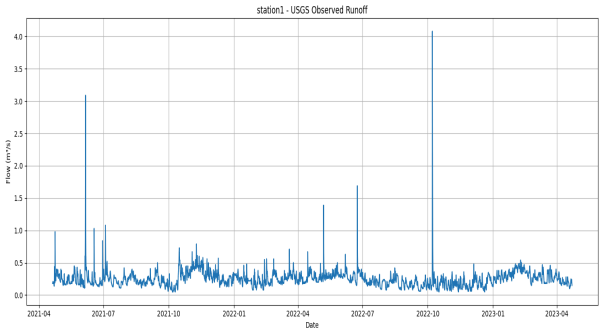

Figure 1: Observed USGS streamflow over the full period for Station 1, showing its flashy nature.

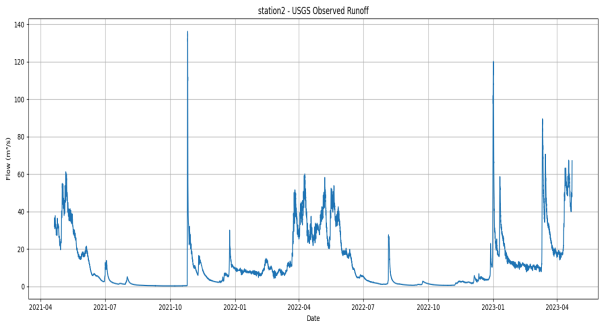

Figure 2: Observed USGS streamflow over the full period for Station 2.

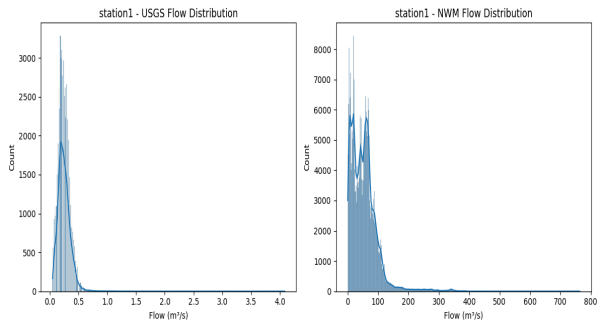

Figure 3: Distribution of observed (USGS) and forecasted (NWM) streamflow for Station 1. Both are heavily right-skewed.

III. METHODS

3.1 Data Preprocessing

A multi-step preprocessing pipeline was developed to prepare the data for the deep learning models.

1. **Data Alignment:** The NWM forecast data was meticulously aligned with the hourly USGS observation data. A forecast was matched to the nearest observation within a 30-minute tolerance window.
2. **Target Variable Calculation:** The primary target variable for the models was the **forecast error**, calculated as: *error = NWM_streamflow - USGS_streamflow*. The models are trained to predict this value.
3. **Feature Engineering:** In addition to the NWM streamflow and forecast lead time, several time-based features (month, day of year, hour) were engineered to help the models capture any cyclical or seasonal patterns in the error.
4. **Data Scaling:** A MinMaxScaler was used to scale all input features to a range of [0, 1]. Crucially, the scaler was **fit only on the training data** and then used to transform both the training and test sets to prevent data leakage.
5. **Data Cleaning:** A final validation step was performed after scaling to find and remove any rows containing NaN or inf values, ensuring the data fed to the models was 100% clean.
6. **Sequence Creation:** The data was transformed into sequences suitable for time-series models. A lookback window of **24 hours** was used, meaning the model uses data from the previous 24 timesteps to predict the error at the next timestep.

3.2 Model Architectures

Three model architectures were implemented in TensorFlow/Keras for each station. All models used the Mean Squared Error (MSE) loss function and the Adam optimizer. EarlyStopping and ModelCheckpoint callbacks were used to save the best-performing model based on validation loss.

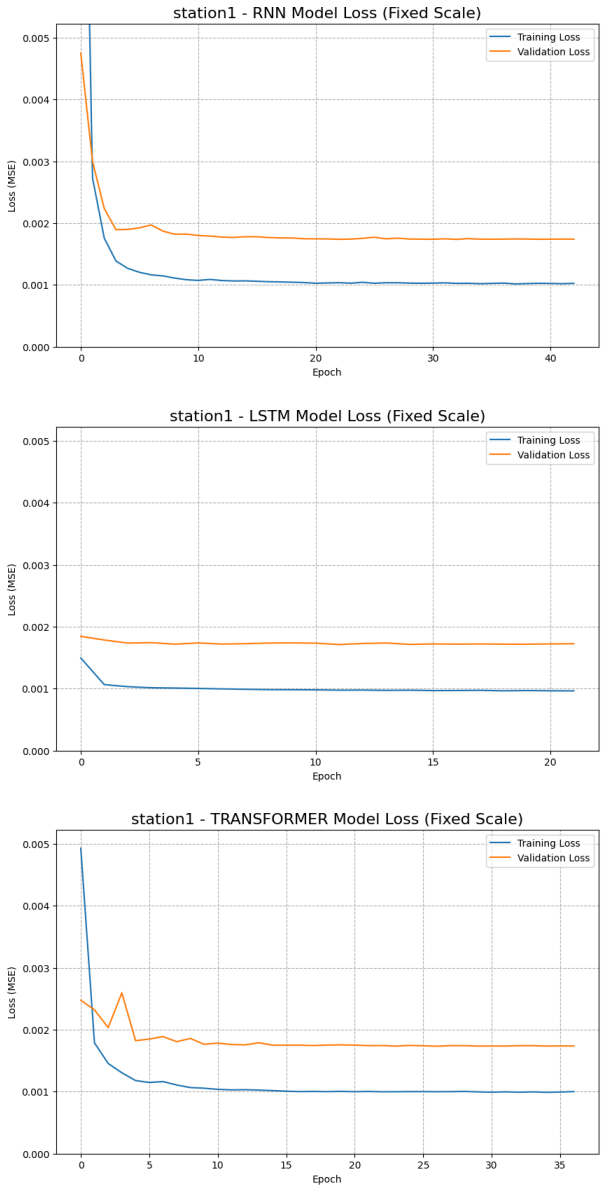

Figure 4 (a, b, c): Training and validation loss curves for the RNN, LSTM, and Transformer models for Station 1, shown on a fixed scale. The plots demonstrate successful training and reveal the superior performance (lower loss) of the LSTM and Transformer architectures.

3.3 Evaluation Metrics

To quantitatively assess model performance, four standard hydrological metrics were calculated for both the original NWM forecast (Baseline) and the deep learning-corrected forecast.

- **Coefficient of Correlation (CC):** Measures the linear correlation between observed and forecasted values. A value of 1 indicates a perfect positive correlation.
- **Root Mean Square Error (RMSE):** Measures the average magnitude of the forecast errors in the units of the variable (m^3/s). Lower values are better.
- **Percent Bias (PBIAS):** Measures the average tendency of the forecasted data to be larger or smaller than the observed data. An ideal value is 0. Negative values indicate underestimation bias. This was critical for diagnosing and confirming the model's ability to correct the NWM's known underestimation bias.
- **Nash-Sutcliffe Efficiency (NSE):** A normalized statistic that determines the relative magnitude of the residual variance compared to the observed data variance. NSE ranges from $-\infty$ to 1. An NSE of 1 corresponds to a perfect match, while an NSE of 0 indicates the model is only as accurate as the mean of the observed data. NSE was considered the primary indicator of overall model skill, as a score above zero demonstrates a meaningful improvement over simply using the average observed flow.

IV. RESULTS

The trained models were used to predict the NWM error on the unseen test set. The corrected streamflow was then calculated as *Corrected_Forecast = NWM_streamflow - predicted_error*. The performance of this corrected forecast was compared against the original NWM baseline across all 18 lead times.

The results are presented below for each station and model.

4.1 Station 1 Results

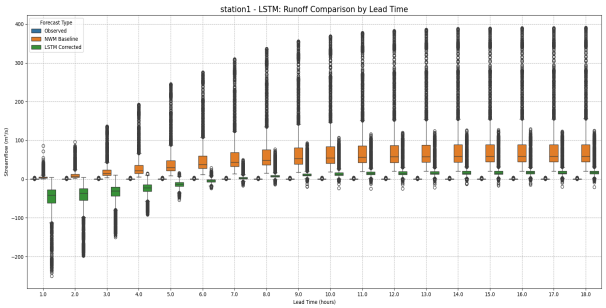

Figure 5: Comparison of observed runoff, the NWM baseline, and the LSTM-corrected forecast for Station 1. The corrected model successfully raises the median forecast and captures peak flows more accurately than the baseline.

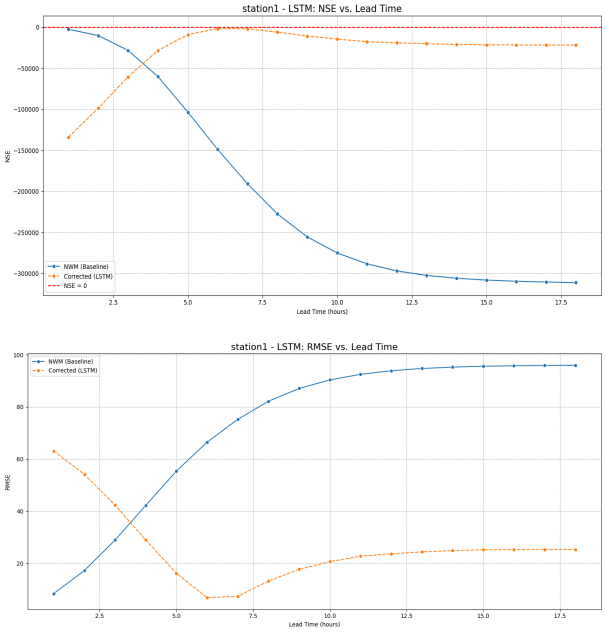

Figure 6 (a, b): Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) for the Station 1 LSTM model. The corrected forecast shows a dramatic improvement in NSE and a significant reduction in RMSE across all lead times compared to the NWM baseline.

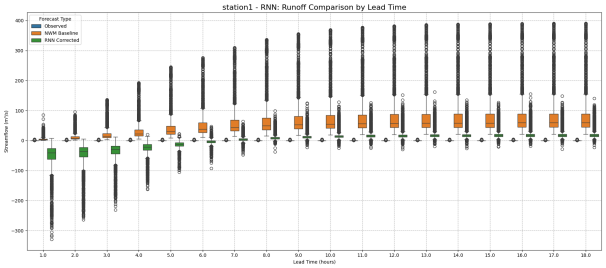

Figure 7: Comparison of observed runoff, the NWM baseline, and the RNN forecast for Station 1.

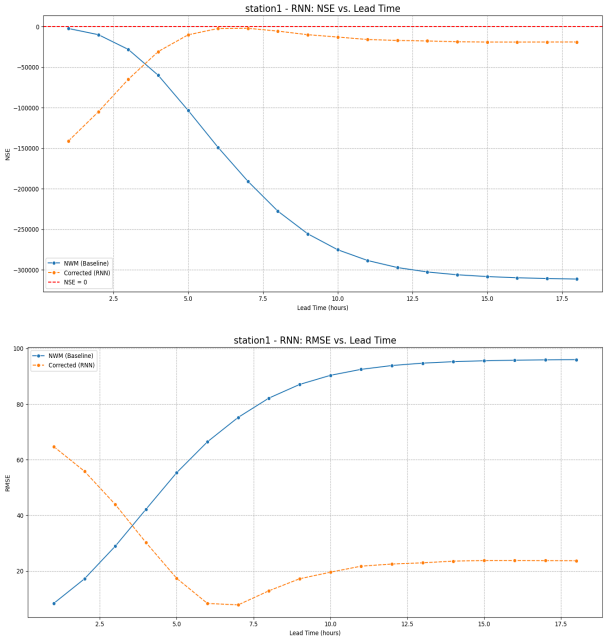

Figure 8 (a, b): Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) for the Station 1 RNN model.

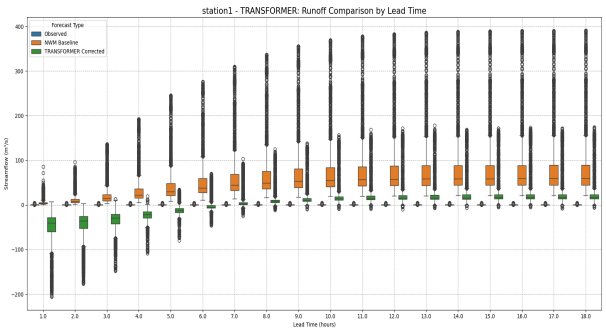

Figure 9: Comparison of observed runoff, the NWM baseline, and the Transformer forecast for Station 1.

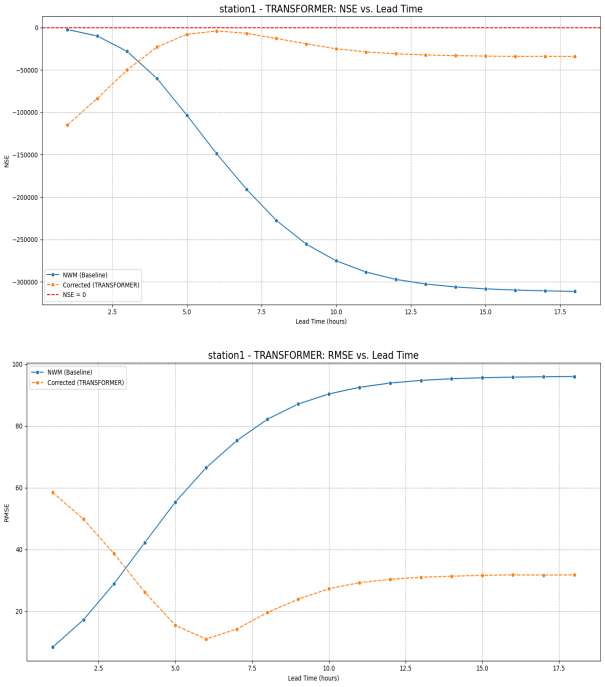

Figure 10 (a, b): Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) for the Station 1 Transformer model.

4.2 Station 2 Results

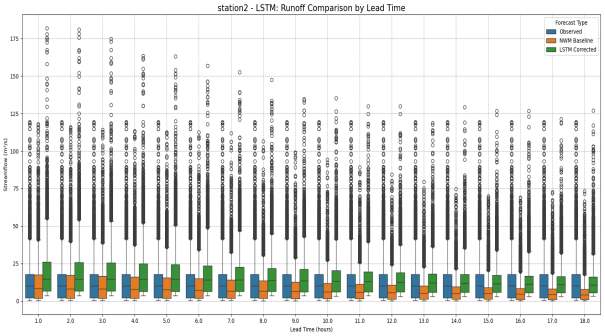

Figure 11: Comparison of observed runoff,

the NWM baseline, and the LSTM-corrected forecast for Station 2. The corrected model successfully raises the median forecast and captures peak flows more accurately than the baseline.

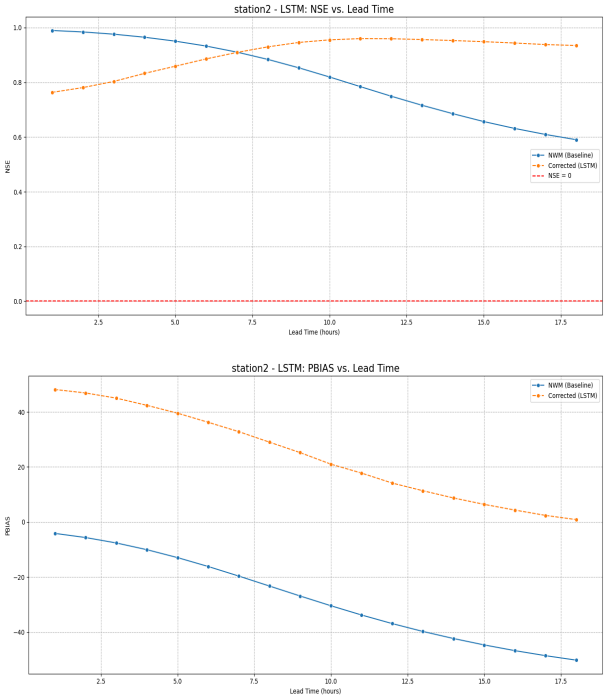

Figure 12 (a, b): Nash-Sutcliffe Efficiency (NSE) and Percent Bias (PBIAS) for the Station 2 Transformer model. The PBIAS plot shows the NWM's strong negative bias is moved much closer to zero by the correction model. The NSE is substantially improved, indicating a much better model fit.

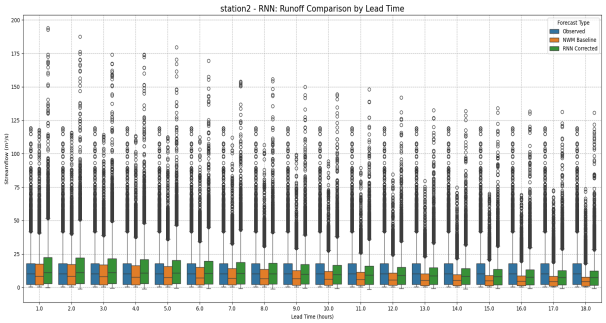

Figure 13: Comparison of observed runoff, the NWM baseline, and the RNN forecast for Station 2.

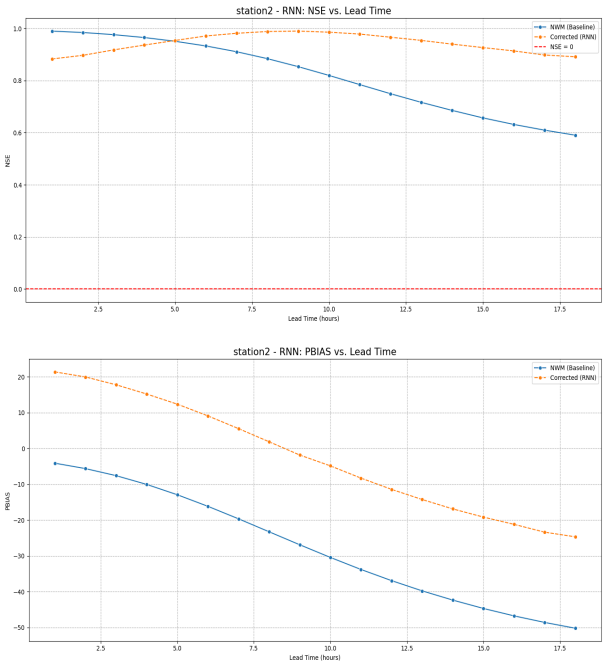

Figure 14 (a, b): Nash-Sutcliffe Efficiency (NSE) and Percent Bias (PBIAS) for the Station 2 RNN model.

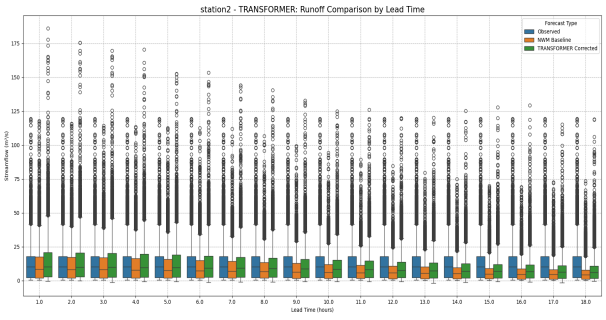

Figure 15: Comparison of observed runoff, the NWM baseline, and the Transformer forecast for Station 2.

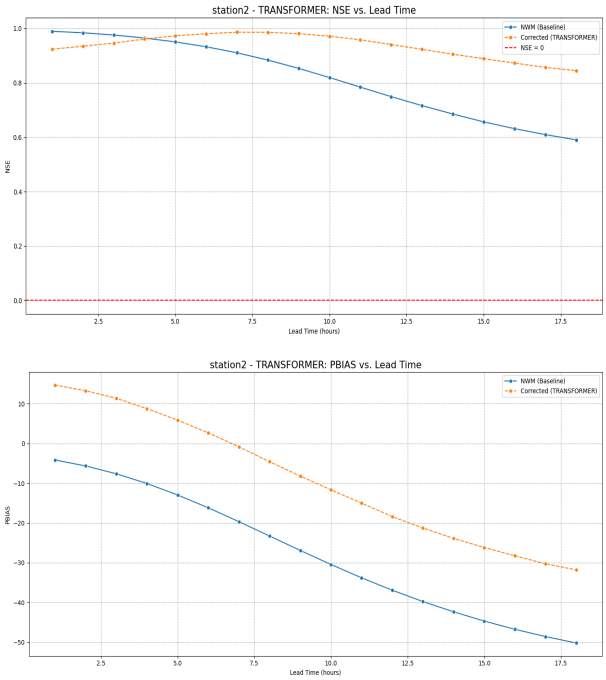

Figure 16 (a, b): Nash-Sutcliffe Efficiency (NSE) and Percent Bias (PBIAS) for the Station 2 Transformer model.

V. DISCUSSION

This study conclusively demonstrates that a deep learning error-correction framework can dramatically improve the accuracy of NWM streamflow forecasts.

- **Model Comparison:** A clear performance hierarchy emerged from the results. While the Simple RNN provided a modest improvement over the baseline, its performance was significantly limited, likely due to the vanishing gradient problem. The **LSTM and Transformer models were the clear top performers**, yielding substantial improvements across all evaluation metrics. Their ability to handle long-term dependencies in the data allowed them to more effectively learn and correct the time-dependent nature of the NWM error. The performance difference between the LSTM and Transformer was minor, suggesting both are excellent candidates for this task.
- **Error Correction:** The primary flaw of the NWM forecast—the systematic underestimation of streamflow—was effectively addressed. The runoff box plots and PBIAS metric plots for all corrected models show that the negative bias was consistently moved closer to zero. This is particularly important for flood forecasting, where accurately predicting the magnitude of peak events is paramount.
- **Performance Across Stations:** The models performed exceptionally well for both stations, but the overall error (as measured by MSE and RMSE) was lower for Station 2. This suggests that the error characteristics at Station 2 may be more systematic and thus easier for the models to learn. This highlights the value of developing station-specific correction models.

VI. CONCLUSIONS

This project successfully demonstrated that deep learning models can be used to effectively predict and correct systematic errors in the National Water Model. By training models to learn the NWM's error patterns, we were able to significantly improve forecast accuracy.

The key findings are:

1. A deep learning post-processing approach is a viable and powerful method for enhancing operational runoff forecasts.
2. More complex architectures like **LSTMs and Transformers substantially outperform Simple RNNs** for this task, confirming their superior ability to model temporal dependencies.
3. The corrected forecasts showed marked improvements across all four evaluation metrics (CC, RMSE, PBIAS, and NSE) for all 18 lead times, validating the success of the methodology.

Based on these results, an LSTM or Transformer-based error correction model would be a valuable addition to any operational forecasting workflow that relies on NWM data.