Conditional time series generation using deep generative model like flow model or diffusive model

A Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Sasidhar Reddy Navuluri
122101025

Mantripragda Venakta Karthikeya
112101027

ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

CERTIFICATE

This is to certify that the work contained in the project entitled “ ***Conditional time series generation using deep generative model like flow model or diffusive model*** *” is a bonafide work of* ***Sasidhar Reddy Navuluri (Roll No. 122101025)*** *&* ***Mantripragda Venakta Karthikeya (Roll No. 112101027)****, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Palakkad under my guidance and that it has not been submitted elsewhere for a degree.*

Dr. Sahely Bhadra

Associate Professor

Department of Data Science

Department of Computer Science & Engineering

Indian Institute of Technology Palakkad

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Streamflow (SF) refers to the movement of water in rivers, streams, and other water bodies. The management of natural resources and the mitigation of the effects of extreme hydrological events, such as droughts and flooding, are contingent upon the precise prediction of streamflow.

Flooding can result from high streamflow, which can damage infrastructure and ecosystems. Conversely, water scarcity can result from low streamflow, which can impact agriculture, energy production, and potable water supplies.

Streamflow forecasting is essential for comprehending evolving hydrological circumstances due to climate change. It facilitates improved planning and response tactics for increasingly frequent and severe weather occurrences.

Precise SF forecasting is crucial for water resource management, enabling authorities to enhance reservoir operations, regulate irrigation systems, and safeguard ecosystems. It facilitates the computation of essential drought indicators such as the Standardized Streamflow Index (SSI), enhancing disaster preparedness and response.

Recent breakthroughs in machine learning and deep generative models, including flow-based and diffusion-based models, offer robust tools for streamflow prediction. These models can elucidate intricate temporal patterns and correlations in streamflow data, enhancing predictive accuracy.

Conditional generative models facilitate the prediction of streamflow by incorporating external variables such as precipitation, temperature, and soil moisture, hence enhancing the forecasting of extreme hydrological events and aiding adaptive water management methods in evolving climatic conditions.

1.1 Motivation

The rising frequency and severity of hydrological extremes, exacerbated by climate change, present substantial problems for water resource management. Comprehending streamflow dynamics is essential for tackling these issues, as streamflow directly affects flood risk, drought intensity, and total water accessibility. Furthermore, streamflow influences essential biological processes, such as the health of aquatic habitats and the pollutant transport capacity of rivers.

Consequently, precise streamflow forecasting can facilitate proactive decision-making for sustainable water management, disaster readiness, and environmental preservation. Utilizing sophisticated deep generative models for streamflow forecasting can augment our capacity to adjust to evolving hydrological conditions and bolster resilience against natural disasters.

1.2 Problem Statement

The capacity to replicate streamflow at ungauged locations continues to be a substantial unresolved issue in hydrology. Precise assessment of water resources is essential for efficient water resource management, especially concerning flood control, drought mitigation, and sustainable utilization. Although gauged areas possess extensive hydrological data, forecasting streamflow patterns in ungauged basins presents significant challenges due to the absence of direct measurements.

In these ungauged basins, the available data is frequently restricted to non-temporal attributes, such as catchment characteristics, in addition to temporal data like precipitation time series. Therefore, there is an urgent necessity to formulate approaches capable of producing dependable streamflow data from inception for these ungauged basins. This involves utilizing streamflow data from hydrologically comparable gauged locations in conjunction with the catchment attributes of the ungauged sites. Quantifying the similarity of several basins based on their characteristics is crucial for this methodology.

Additionally, it is very important to make sure that the suggested method can also accurately predict streamflow in gauged areas, using data that is very close to what has been observed. We can improve our ability to predict streamflow in both gauged and ungauged basins by using advanced deep generative models, such as flow-based or diffusion-based models, to create conditional time series. This will lead to better water resource management strategies in the long run.

1.3 Objective

- Predict streamflow in ungauged basins using non-temporal features (e.g., catchment characteristics) and temporal data from gauged basins.

- Integrate temporal and non-temporal data to enhance model performance and accuracy.

- Utilize advanced deep generative models (diffusion-based) for capturing complex streamflow dynamics.

- Transfer knowledge from gauged to ungauged basins by quantifying basin similarity based on attributes.

- Improve water resource management for better flood control, drought mitigation, and sustainable utilization.

1.4 Idea

We want to efficiently incorporate both temporal and non-temporal characteristics using Diffusion-TS, a state-of- the-art generative framework, to forecast streamflow for ungauzed basins. The capacity of the Diffusion-TS architecture to capture intricate temporal patterns in time series data and offer interpretability via seasonal-trend decomposition makes it especially appropriate for this work.

We will utilize temporal data such as precipitation and historical streamflow to model streamflow dynamics. The downstream volume of streamflow is significantly influenced by upstream conditions and precipitation patterns over time. Therefore, our model must effectively learn these temporal dependencies to generate accurate predictions.

We will add non-temporal data like catchment features (Gauge latitude, Gauge longitude, Depth to bedrock , Soil porosity etc) to Diffusion-TS to make it better at making predictions. To do this, these traits will have to be embedded and added at the decoder input stage. We can get a better picture of the connections between different basins by carefully inserting non-temporal data.

When non-temporal characteristics are added, the generation process changes into a conditional one. By using these catchment properties along with temporal data to condition Diffusion-TS, we can tell the model to make streamflow forecasts that are true to each basin's unique features. We can use learned connections from gauged basins to put together streamflow data for basins that aren't gauged.

Our method tries to figure out how similar basins are by using physical factors from the features of catchments. This way, we can be sure that the streamflow data we get is a good reflection of how basins' water levels change over time. From the time series data, the model will also learn about implicit connections, such as how tributaries and distributaries affect each other.

Considering the constraints of existing datasets, especially with respect to tributaries and various gauges, our attention will be directed towards utilizing current daily precipitation data and catchment features.This will provide a solid foundation for our predictions, allowing us to generate meaningful streamflow data for ungauged basins despite limited historical data.

1.5 Previous Work

By resolving the issues with our prior implementation, we enhanced our conditional time series generation model for streamflow prediction this semester. To improve generalization and predictive performance, we specifically enlarged the dataset and changed the model architecture. The following is the revised model structure:

Encoder

The encoder creates a latent representation of the streamflow data by processing temporal aspects. Among the inputs are:

- **Temporal Data** – Features such as precipitation, temperature, and other meteorological variables.

Decoder

The decoder reconstructs the streamflow time series using:

- **Encoder Output** – The latent representation learned from temporal inputs.
- **Streamflow Data** – Reconstruction is improved by directly utilizing historical streamflow observations as an input.

Key Improvements

- **Expanded Dataset:** 30 gauge data points were analyzed for evaluation after 533 gauge data points were used for training in order to capture a variety of hydrological trends.
- **Improved Generalization:** Instead of overfitting to one place, the model now learns variances across many basins.
- **Removal of Seasonality Block in Decoder:**By removing the seasonality block from the decoder, the learned temporal relationships—rather than predetermined periodic components—were the main focus, which reduced complexity and increased prediction accuracy.

1.6 Current Work

Key Improvements

- Integration of Non-Temporal Data: Static non-temporal basin features (such as soil type, elevation, etc) were incorporated as conditioning factors alongside dynamic temporal data, expanding on the multi-basin approach. This gives the model the necessary static background to produce streamflow patterns that are more physically accurate and basin-specific.
- Enhanced Conditional Generation: With the help of the non-temporal data, the diffusion model can now produce stream flow that is conditioned on both time-varying inputs (such as temp) and the distinct, static characteristics of each individual basin.
- Non-Temporal Data Pathway: To deal with the non-temporal input, the Transformer architecture was modified to include specialized layers and processing steps. As part of this, non-temporal features are expanded across the sequence length, embedded, and ready for merging with encoded temporal information to condition the decoder.

Adjustments made to the model architecture:

Encoder

Inputs:

- Temporal Data: Data that changes with the time.

By analyzing these inputs, the encoder creates a multi-dimensional latent representation that records changing patterns over time.

Decoder

Inputs:

- Encoded Temporal Data (Output from the encoder)
- Processed Non-Temporal Data (Static basin characteristics, prepared for conditioning)
- Noisy Streamflow Data

Key Process

- **Forward Process:** Noise is gradually applied at each time step to pure streamflow data.
- **Reverse Process:**With the help of conditioning data from the temporal variables and the static non-temporal basin properties that relate to the particular streamflow sample, the model learns to denoise the streamflow and forecast the original signal.
- **Sampling/Testing:** This step is completed after training. From a random noise sample, the trained model creates synthetic streamflow data. This is now a conditional generation that is dependent on the target basin's unique non-temporal features as well as the temporal data sequence that was provided.

Chapter 2

Review of Prior Works

2.1 Streamflow prediction in ungauged basins using conditional GAN

Sneha Bhattacharya's study on the use of Conditional GANs (CGANs) to time series forecasting identifies a number of significant drawbacks that reduced the model's efficiency. One significant problem was that the generative model had trouble interpreting the longer conditioning values, which went beyond basic label encodings. Training challenges, especially with Autoencoder variation 1's vanishing gradient issue, exacerbated this insufficiency. While a reduction in the size of the input vector brought some respite, it also highlighted the need for a more advanced model architecture that could handle complicated embeddings efficiently.

Even though a Wasserstein GAN (WGAN) loss function was used, signs of mode collapse showed up. This meant that the generator could take advantage of flaws in the discriminator and stop learning for real. The fact that there isn't much written about debugging mode collapse in this situation shows that we still don't fully understand it. Principal Component study (PCA) study also showed that the synthetic values were very spread out and didn't pick up on short-term patterns and spikes. Instead, they only picked up on long-term seasonal trends with some accuracy. These flaws show that the model wasn't strong enough to accurately reflect the changes in the original dataset.

The problems with managing many inputs were made worse by the fact that the dataset was too complicated, with 56 factors and time series inputs, which was more than what had been shown in previous research. When different loss functions, such as RMSE and the traditional GAN Jensen-Shannon distance, were tried to replace the WGAN loss function, they performed worse. This showed that the WGAN method, despite its problems, was still the best choice. Furthermore, several autoencoder configurations were tried, but only Variation 2 produced satisfactory outcomes. The fact that Variation 3, which used all the available time series data, couldn't make things better says that the model didn't make good use of the link between rainfall and streamflow. Because of these problems, it was decided to look into diffusion models, which might offer a stronger way to show complicated temporal changes in time series data.

Chapter 3

Data Description

CAMELS dataset: Catchment Attributes and MEteorology for Large-sample Studies.

It has information on hydrometeorological forces from 671 basins in the US. It records streamflow every day from January 1, 1980, to December 31, 2014. There are 18 close areas made up of the 671 catchments. There is only one gauge for each river. For this dataset, the words ”catchment,” ”river,” and ”gauge” can all mean the same thing. It has a lot of non-temporal catchment qualities that stay the same or stay about the same over time. Some of these properties are shown in Table 3.1. It also comes with time data, which is summed up in Table 3.2. There is no record of tributary and distributary statistics.

3.1 Non-temporal Attributes

Non-temporal characteristics, as shown in Table 3.1.

3.2 Temporal Attributes

Temporal data is given in Table 3.2.

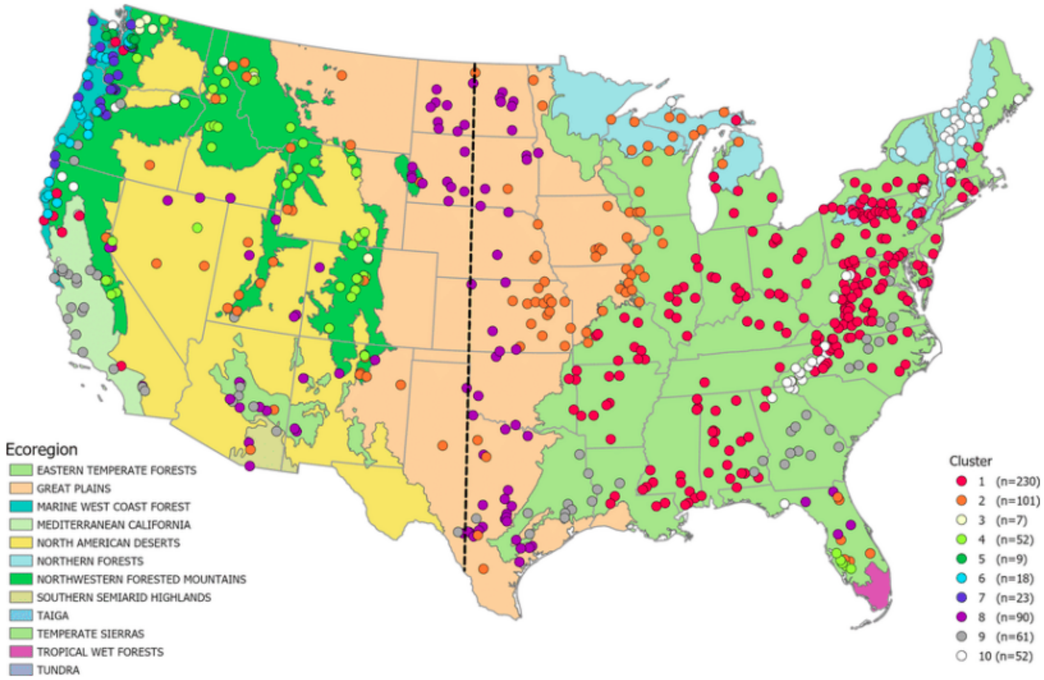

Fig. 3.1 CAMELS catchment basins

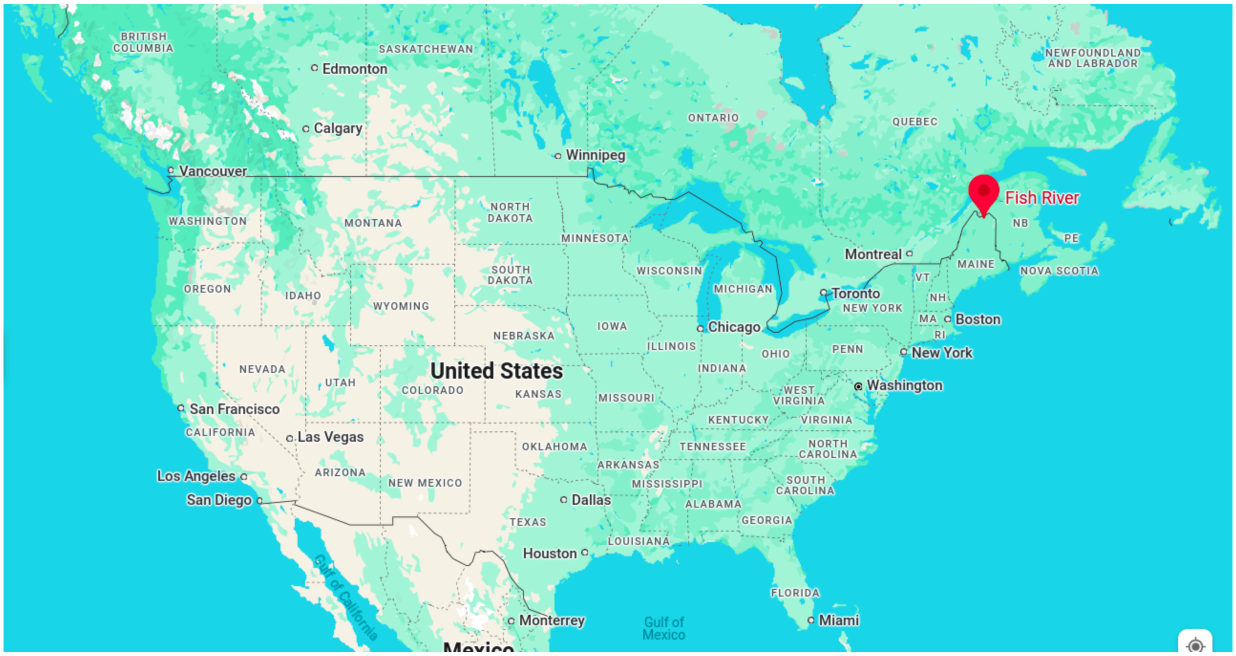

Fig. 3.2 Location of gauge used for training

Table 3.1 Description of Non-temporal attributes in CAMELS-US

Parameter	**Description**	**Unit**
gauge_lat	*gauge latitude* (ϕ)	° north
gauge_lon	*gauge longitude* (θ)	° east
elev_mean	catchment mean elevation	meter above sea level
slope_mean	catchment mean slope	m/km
area_gages2	catchment area (GAGESII estimate)	km^2
pet_mean	mean daily PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment]	mm/day
aridity	PET/P, ratio of mean PET to mean precipitation	-
p_seasonality	seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and preciptiation cycles	-
soil_depth_pelletier	depth to bedrock (maximum 50m)	m
soil_porosity	Saturated volumetric water content estimated using a multiple linear regression based on sand & clay fraction	-
silt_frac	silt fraction (of the soil material smaller than 2 mm, layers marked as oraganic material, water, bedrock and other were excluded)	%
water_frac	fraction of the top 1.5m marked as water	%
lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index (based on 12 monthly means)	-
gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction (based on 12 monthly means)	%
forest_frac	forest fraction	%
stream_elas	streamflow precipitation elasticity (sensitivity of streamflow to changes in precipitation at the annual time scale)	-
low_prec_timing	Season during which most dry days (<1 mm/day) occur	season
geol_permeability	subsurface permeability ($\log_{10}$)	m^2

Table 3.2 Description of temporal attributes in CAMELS-US

Parameter	**Description**	**Unit**
dayl	Daylength; the length of time between sunrise and sunset during a specific day	s
prcp	Precipitation; the amount of rainfall or snowfall over a day	mm
srad	Solar radiation, measured in watts per square meter; indicates the amount of solar energy reaching the Earth's surface	W m^{-2}
swe	Snow Water Equivalent; the amount of water contained within the snowpack, given in terms of the depth of liquid water that would result if the entire snowpack melted	mm
tmax	Maximum air temperature	°C
tmin	Minimum air temperature	°C
vp	Vapor pressure; the pressure exerted by water vapor in the air and is related to humidity	Pa

Chapter 4

Previous Implementation Details

By mid semester, we refined the conditional time series generation model by incorporating a larger dataset and modifying the model architecture. The improvements aimed to enhance generalization, reduce complexity, and improve prediction accuracy.

4.1 Training

4.1.1 Forward Process

Inputs

- **533 Gauge Data Points for Training**–A much larger dataset in order to enhance generalization.

Noise Addition

- Gaussian noise is added to the temporal data to gradually corrupt it and make it noisier.

4.1.2 Reverse Process

Inputs

- **Streamflow Data** –Streamflow data is now directly given to the decoder.

Key Modification

Seasonality Block Removed–The seasonality block has been removed, in contrast to the previous implementation, which lowers model complexity and enhances predictive performance by focusing on learned temporal relationships.

4.2 Testing Phase

- **30 Gauge Data Points Used for Testing**–The generalization capacity of the model was assessed by evaluating its performance on unseen gauge locations.

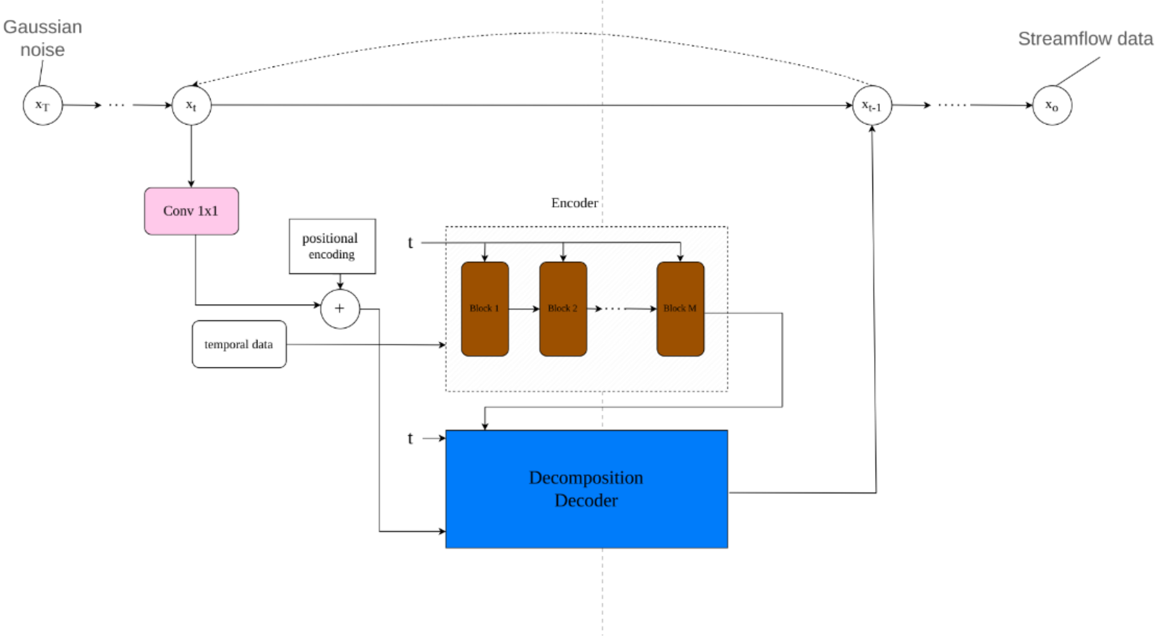

Fig. 4.1 Diffusion TS model

Chapter 5

Previous Results and Evaluation

The graph below represents the new gauge point, tested exclusively with temporal data, achieving an RMSE of 0.181.

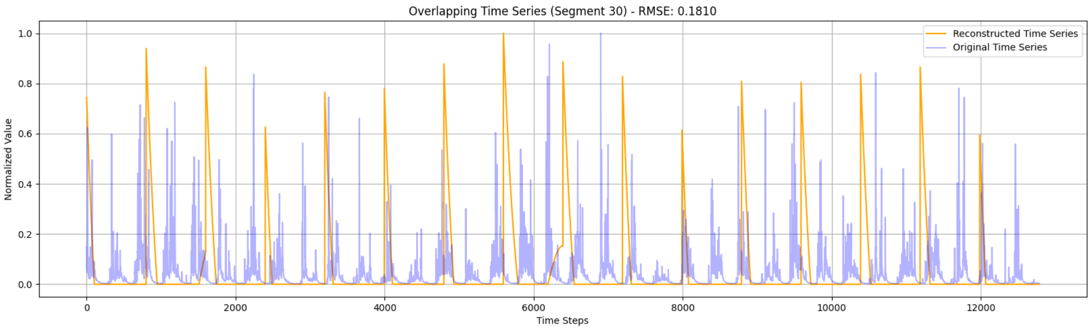

Fig. 5.1 result of new architechture (blue: original, orange: constructed)

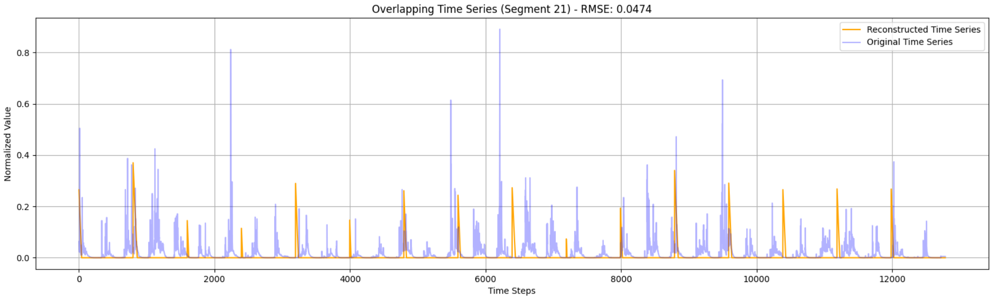

Fig. 5.2 Test basin 21

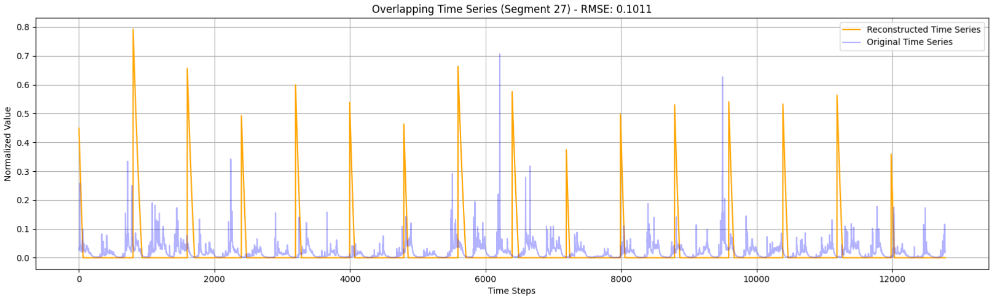

Fig. 5.3 Test basin 27

Chapter 6

Current Implementation Details

This semester, adding static non-temporal basin features to the larger temporal dataset from many basins greatly improved the conditional time series generation model. With an objective of better generalization and basin-specific streamflow production, architectural changes were made to the Transformer model to make efficient use of this extra static information.

6.1 Training

6.1.1 Forward Process

Inputs

- **Multi-Basin Temporal Data:** Time series data (such as Temp data) from 533 different gauge basins are used for training. The dynamic influences of the environment are captured by this temporal data.

  Dimension (per batch): (Batch Size, Seq Length, TEMP_DIM) → (16, 799, 7)

- **Aligned Non-Temporal Basin Attributes:** Static attributes (such as height, soil type, and area) are given for each of the 533 basins and correlated with temporal sequences.

Dimension (per basin): (NON_TEMP_DIM,) → (17,)

Dimension (input to model per batch): (Batch Size, 1, NON_TEMP_DIM) → (16, 1, 17)

- **Streamflow Data:** Streamflow time series corresponding to the 533 basins.

  Dimension (per batch): (Batch Size, Seq Length, SF_DIM) → (16, 799, 1)

Noise Addition

- **Gaussian Noise Addition:** Gaussian noise is gradually added to pure streamflow samples over discrete timesteps according to a defined schedule (e.g., cosine), creating noisy samples x_t for the training.

6.2 Reverse Process

Inputs

- **Temporal Conditioning Data:** Time series data corresponding to the batch.

  Dimension: (16, 799, 7) → Embedded to (16, 799, d_model=96)

- **Non-Temporal Conditioning Data:** Static basin attributes for each sequence.

  Dimension: (16, 1, 17)

- **Noisy Streamflow Data:** Noise-corrupted streamflow sample x_t at timestep t.

  Dimension: (16, 799, 1) → Embedded to (16, 799, d_model=96)

- **Diffusion Timestep t:** The noise level index for the present sample.

Key Modification

Integration of Non-Temporal Data: Static non-temporal features were incorporated into the decoder:

- **Encoder Processing:** Temporal data is processed by the Encoder yielding a latent representation.

  Dimension: (16, 799, d_model=96)

- **Encoder Output Projection:** Encoder output projected.

  Dimension: (16, 799, enc_out_proj_dim=79)

- **Non-Temporal Expansion:** Non-temporal data dimension along the sequence length.

  Dimension: (16, 799, NON_TEMP_DIM=17)

- **Conditioning Fusion:** Projected encoder output concatenated with expanded non-temporal data.

  Dimension: (16, 799, 79+17=96)

Seasonality Block Removed: By removing the explicit seasonality block, the model's ability to learn complex temporal relationships directly from the data and conditioning variables was improved.

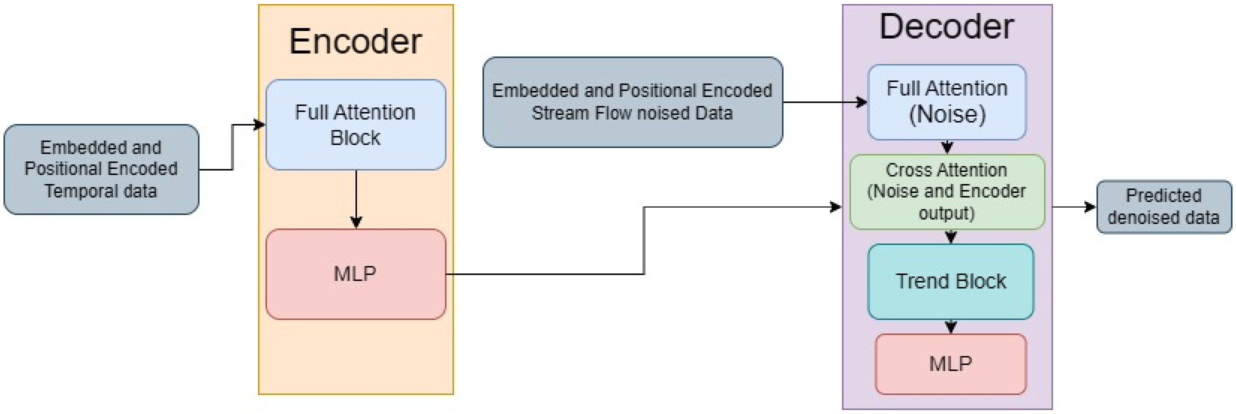
Embedded and Positional Encoded Temporal data
Encoder
Full Attention Block
MLP
Embedded and Positional Encoded Stream Flow noised Data
Decoder
Full Attention (Noise)
Cross Attention (Noise and Encoder output)
Trend Block
MLP
Predicted denoised data

Fig. 6.1 Encoder-Decoder Architecture

6.3 Testing Phase

- **30 Unknown Gauge Basins Used for Testing:** 30 basins that were not included in training were used to assess the model's generalization.
- **Conditional Generation Inputs:** For every unseen basin, temporal sequences and related non-temporal properties are provided to the model during testing.
- **Output:** The model creates simulated streamflow sequences that are dependent on static non-temporal basin properties.

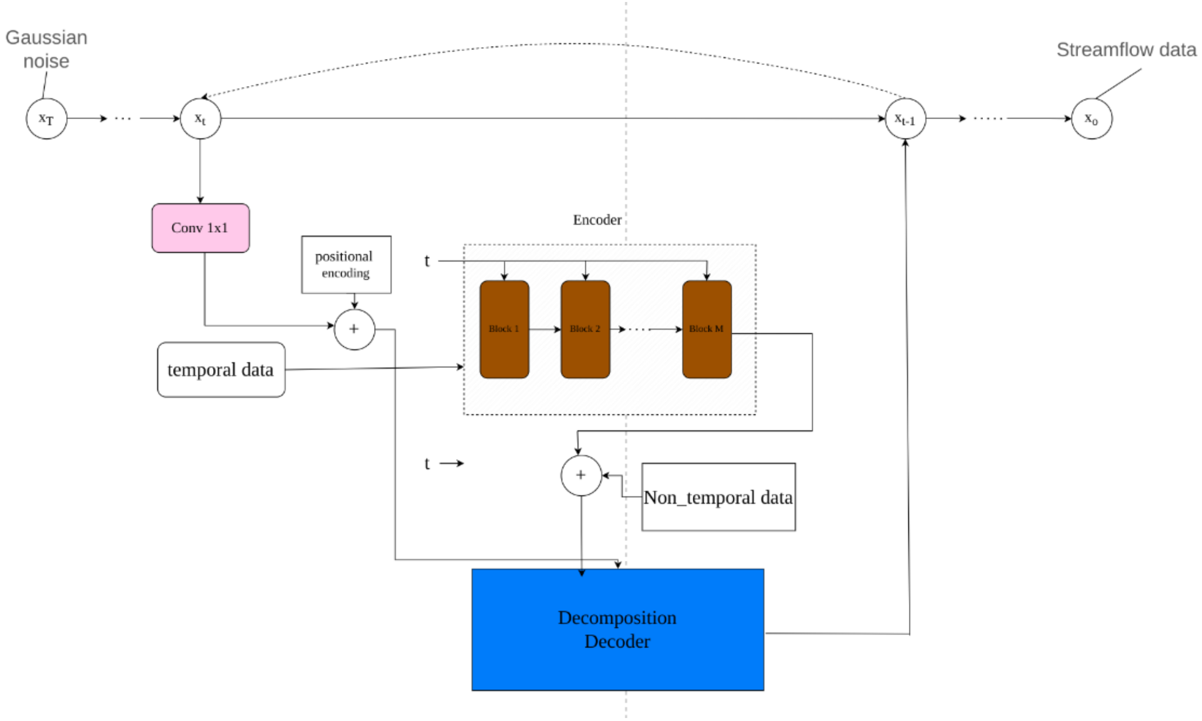

Fig. 6.2 Updated Diffusion model architecture

Chapter 7

Results

The graph below represents the new gauge point, tested with temporal data encorporating with non temporal data.

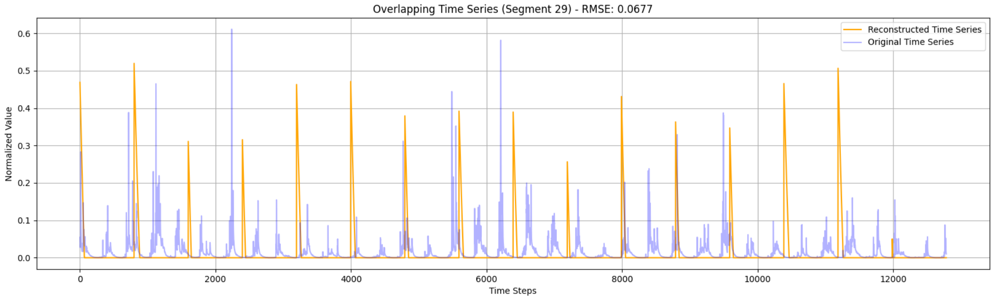

Fig. 7.1 Streamflow plot of Gauge 29

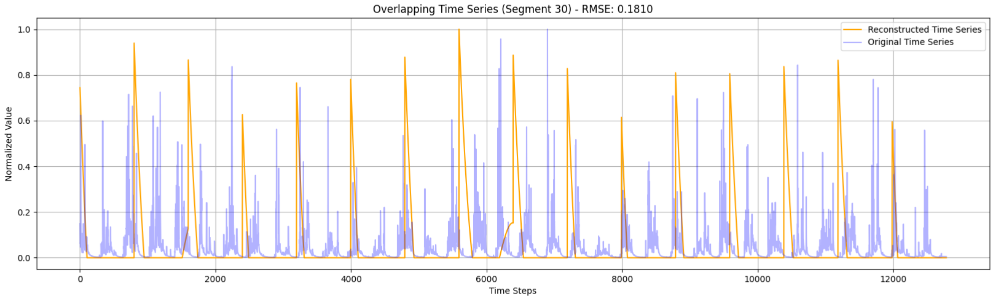

Fig. 7.2 Streamflow plot of Gauge 30

Chapter 8

Conclusion

In this study, a diffusion-based model (Diffusion-TS) was developed to enhance a conditional time series generation framework for streamflow prediction. Expanding the training dataset to 533 different basins and combining static basin attributes like elevation and soil type with dynamic temporal variables like temperature were two significant improvements over earlier work. The model's ability to learn intricate, data-driven temporal correlations was made possible by architectural improvements, most notably the elimination of an explicit seasonality block. Using 30 previously unexplored basins, the model demonstrated strong generalization by effectively generating basin-specific streamflow sequences by conditioning the reverse diffusion process on both temporal and non-temporal inputs.

Although there is still difficulty in accurately representing the range of hydrological variability, especially extreme events, the model demonstrated promise, particularly for ungauged or data-scarce scenarios. Further improvement of the integration of static features, the addition of attention mechanisms designed for non-temporal data, and the expansion of evaluation with more thorough hydrological measures are possible future paths. Despite these obstacles, the improved conditional diffusion model shows great promise for improving hydrological modeling and water resource management and is a major step toward more precise streamflow prediction.

8.1 Contribution

Sasidhar: Model Design and Methodology,Results Analysis and Evaluation

Karthikeya: Data Collection and Preprocessing,Model Training and Experimentation.

References

[1] X. Yuan and Y. Qiao, “Diffusion-ts: Interpretable diffusion for general time series generation,” *Hefei University of Technology*, 2023.

[2] S. Bhattacharjee, “Streamflow prediction in ungauged basins,” 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*. NeurIPS, 2017.