# Enhancing Streamflow Prediction in Vulnerable Regions through Probabilistic Deep Learning and Satellite-Derived Data


Fatemeh Ghobadi[1], Amir Saman Tayerani Charmchi[1], JungMin Lee[2], Myeong In Kim[2], and Kichan jung[1]

[1] Department of Digital Convergence, Onpoom Co., LTD, Seoul, South Korea
[2] Land and Housing Research Institute, Daejeon, 34047, Korea


EGU General Assembly 2025

# Contents



# Problem statement

## Streamflow forecasting incorporates various parameters

Meteorological data, Hydrological models, and Historical data;

## Importance of streamflow forecasting

Flood and water resource management, environmental monitoring, climate change adaptation, and informed decision making;

## Challenges in poorly gauged basins

Data scarcity, climate variability and change, human intervention, and inadequate infrastructure;

## Innovative approaches to overcome data limitations and improve forecasting accuracy

Geo-spatiotemporal models, mesoscale data, Attention-based networks

EGU General Assembly 2025

# Improve forecasting accuracy

## Geo-spatiotemporal features

Meteorological data (such as temperature, precipitation, humidity, and wind speed);

Geographic locations, time stamps, and related attributes

## Data Preprocessing

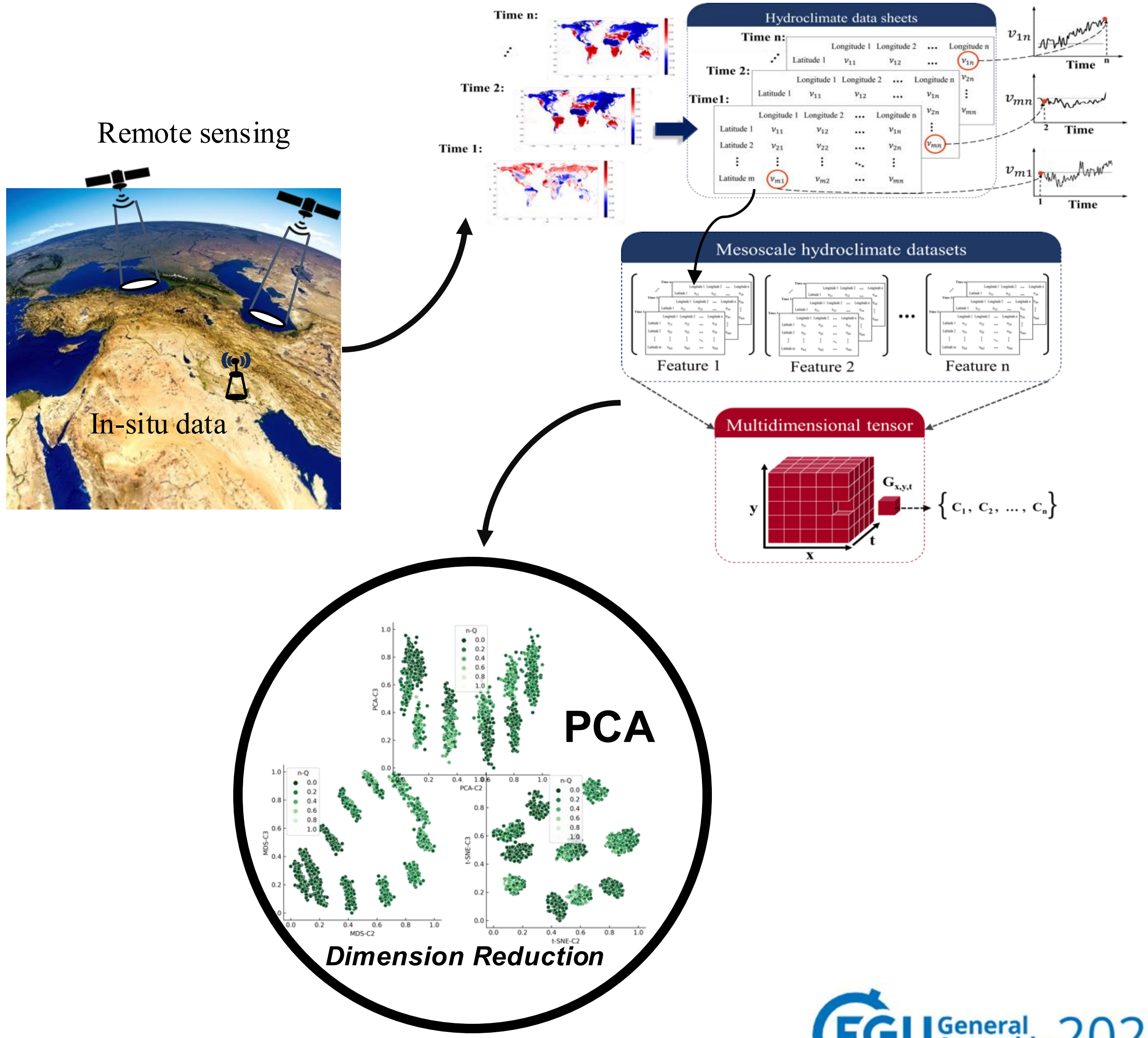
Cleaning the data, normalizing the data, and dealing with missing data

## Feature engineering

Cascade dimensionality reduction for feature extraction, Cross-correlation analysis, Feature selection

## Model development

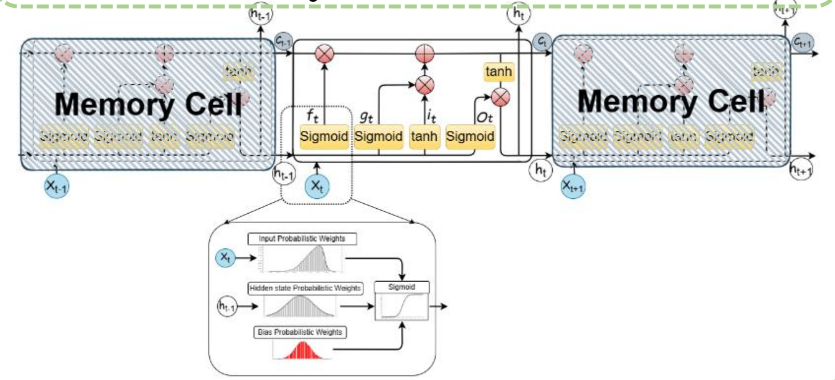
Advanced algorithms, optimization, and validation for geo-spatiotemporal modeling.



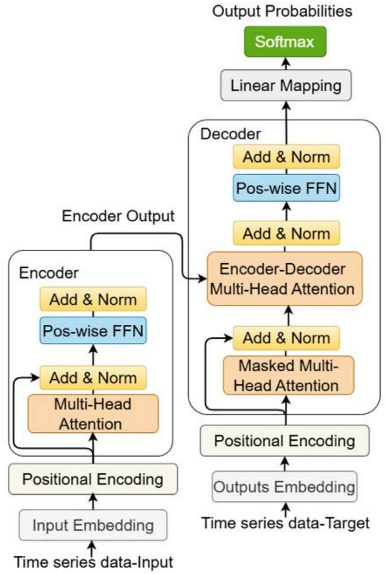
# Advanced probabilistic algorithms

## Attention-based architecture

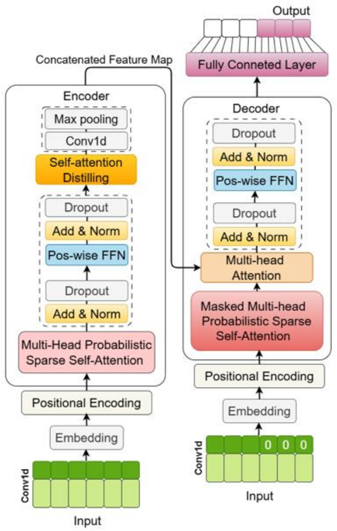
### Bayesian LSTM



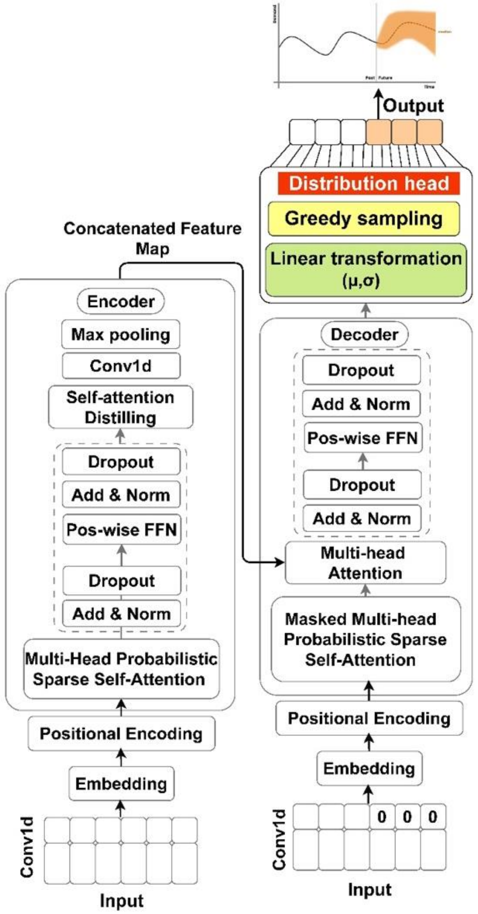

### Transformer




### Informer



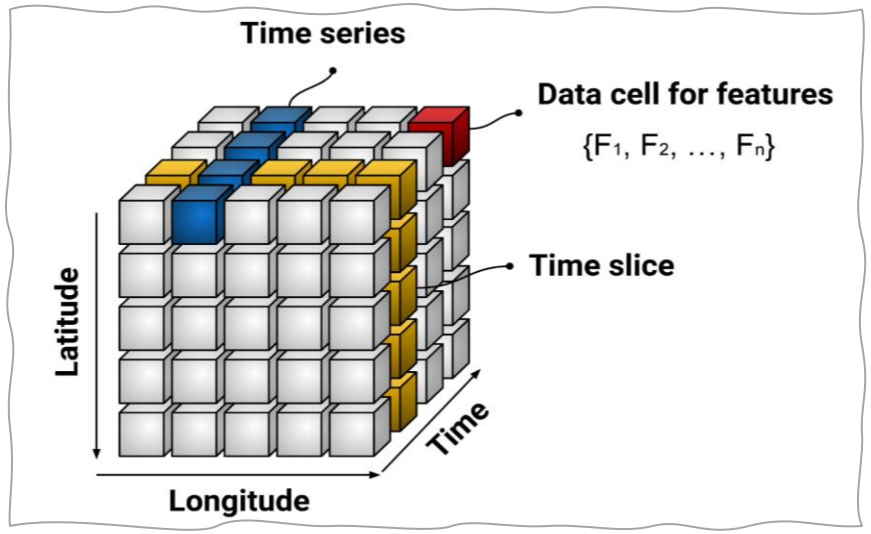

### Probabilistic Informer



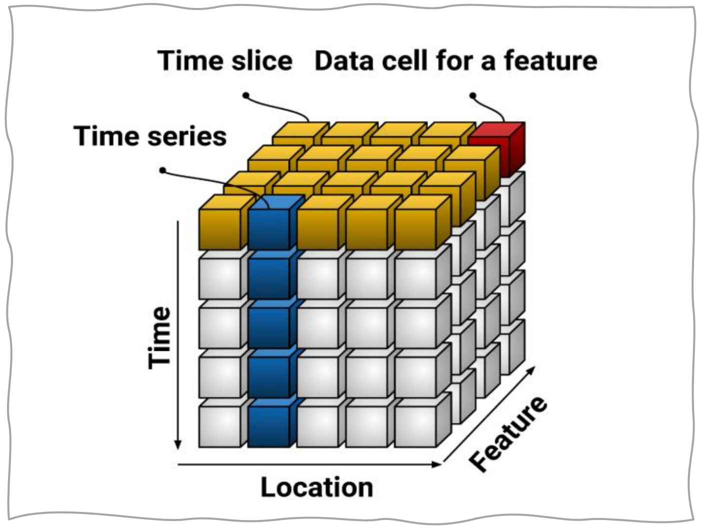

## Three attention-based networks

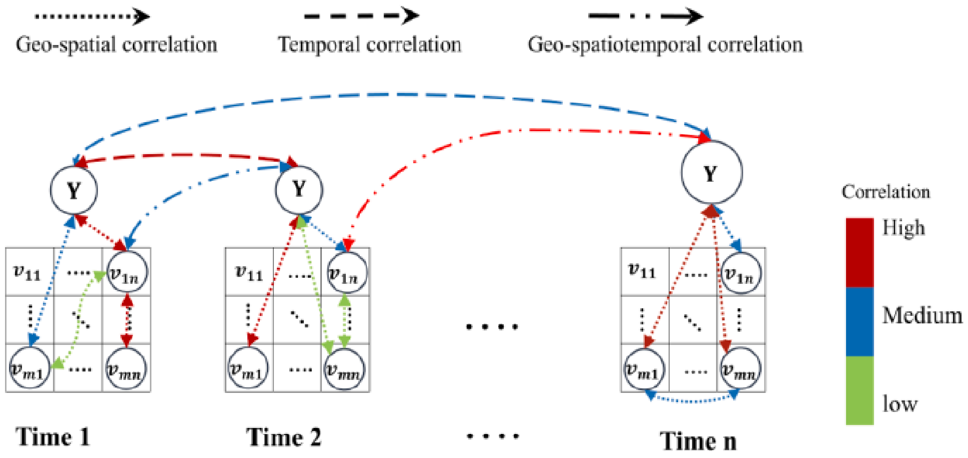
| Model | Input Encoding | Attention Style | Strengths | Limitations |
|---|---|---|---|---|
| **SageFormer** | GNN (label encoding) | Spatial + Temporal | Graph-based spatiotemporal learning | Needs graph structure, more complex |
| **Informer** | CNN + Time features | ProbSparse Attention | Efficient, long-sequence forecasting | Less spatial modeling, heuristic pruning |
| **Transformer** | Linear + Positional | Full Self-Attention | General, deep temporal dependencies | $O(L^2)$ cost, no feature relation modeling |

# Adopted methodology




Convert high-dimensional geo-spatiotemporal data into lower-dimensional representations (Spatiotemporal Space ) that still capture essential variability.



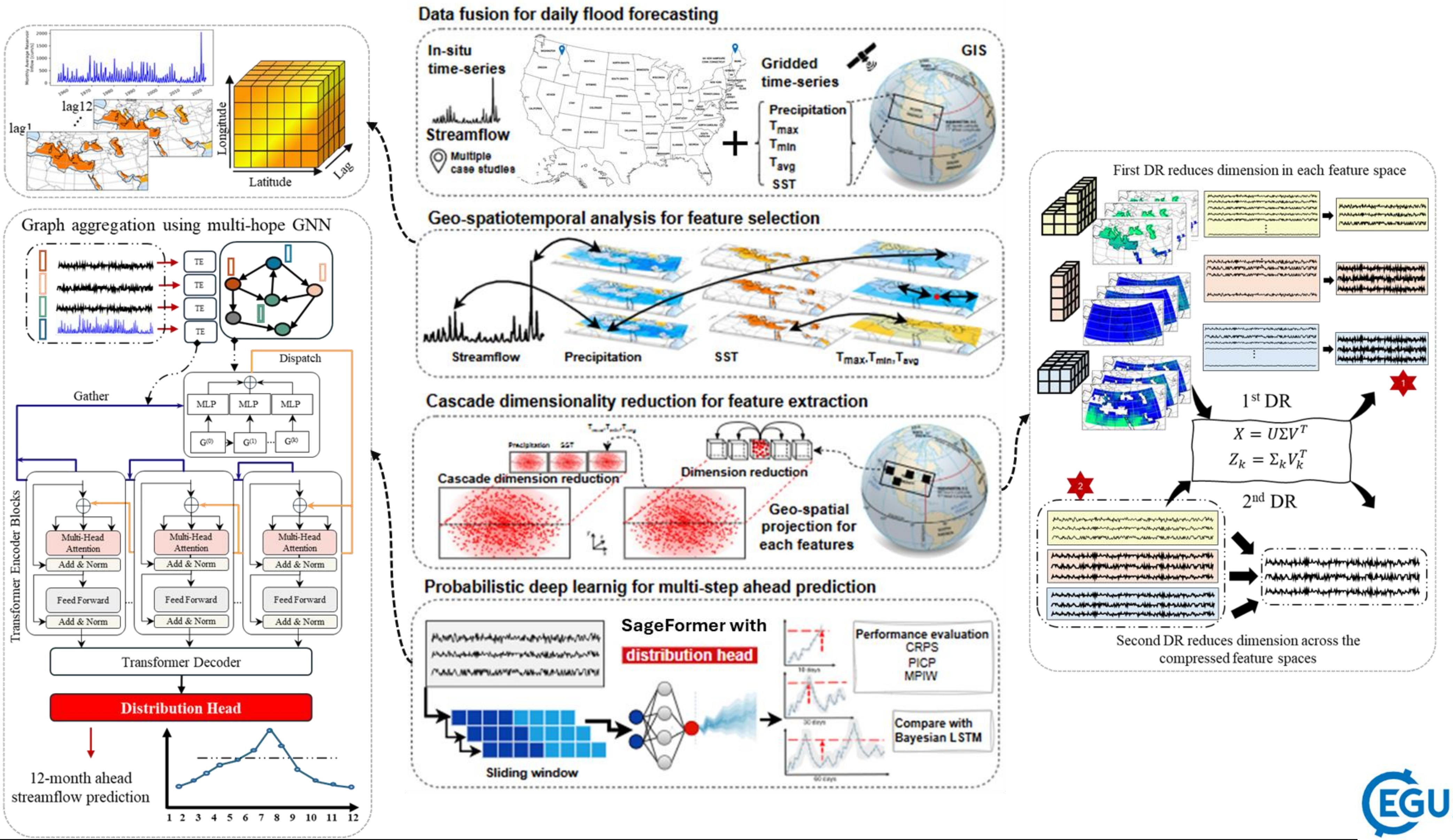




Visualization of interdependencies among geo-spatiotemporal data, highlighting the complex relationships and interactions between dimensions

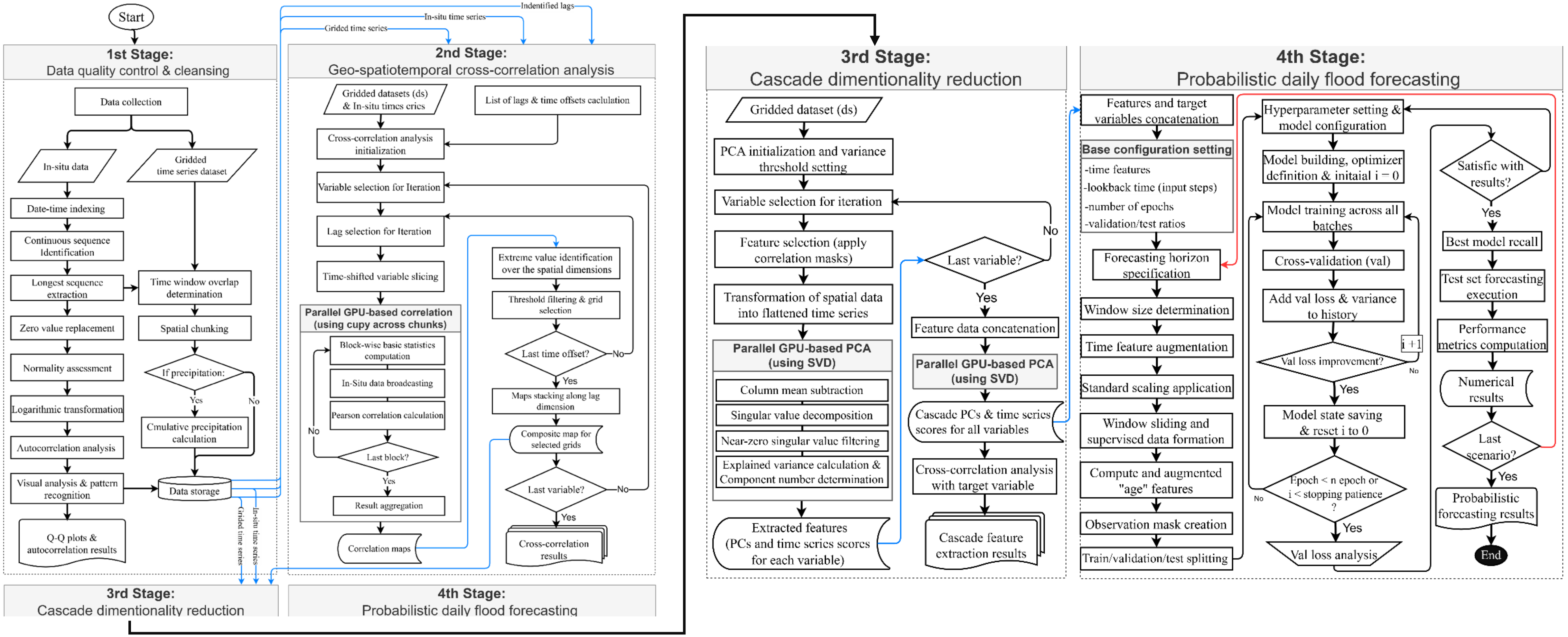
Why High Dimensionality is Problematic?

- Computational Complexity:
  High-dimensional datasets significantly increase computation cost and processing time.
- Overfitting Risk:
  With many features (spatial grid points over time), machine learning models can become prone to overfitting, degrading prediction accuracy.
- Redundant Information:
  Satellite-derived data often contains redundant information (spatial correlation) across adjacent grid points.
- Curse of Dimensionality:
  As dimensions grow, data points become sparse, causing performance deterioration and reducing statistical significance

# Proposed Framework



# Solution Framework

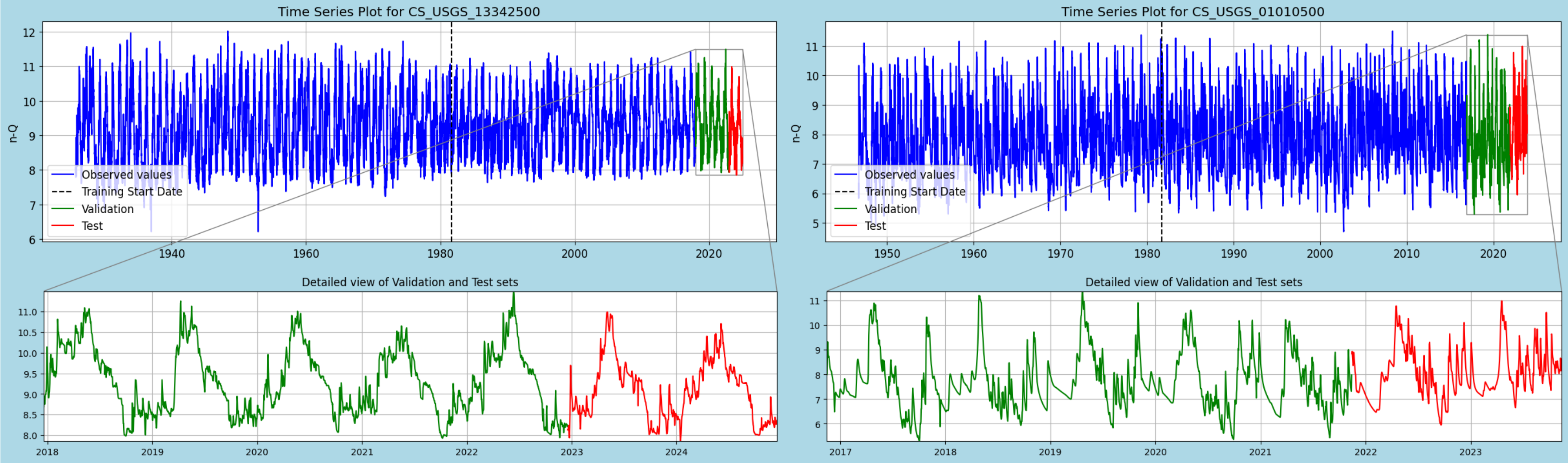


# Case study

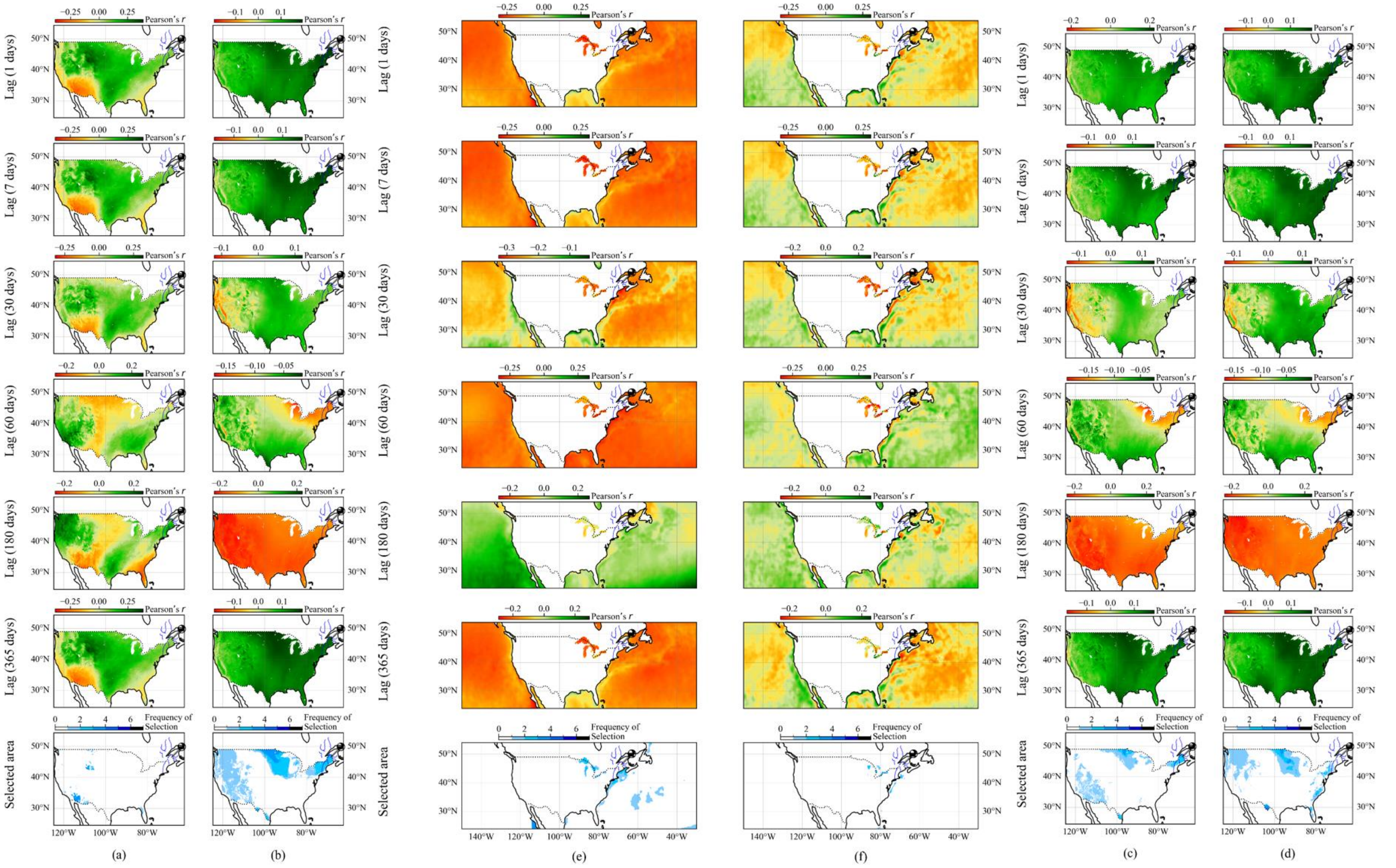
**In Situ timeseries**

**Mesoscale hydroclimate datasets**

- **Optimum Interpolation Sea Surface Temperature**
- **NOAA nClimGrid-Daily Version 1**



# Geo-spatiotemporal correlation analysis



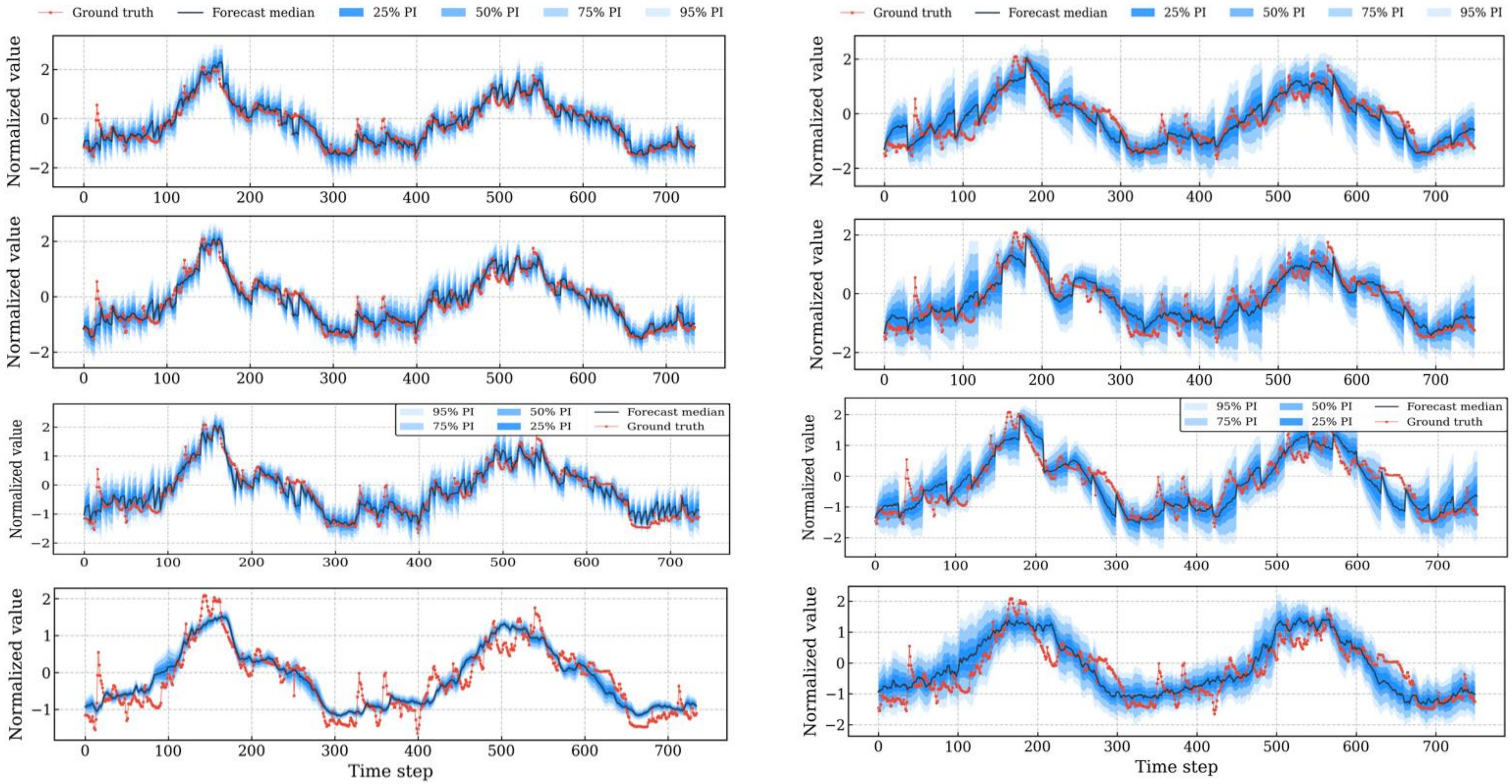

Geo-spatiotemporal correlation between streamflow and (a) Precip, (b) Tavg, (c) Tmin, (d) Tmax, (e) sea surface temperature (SST), and (f) anomaly SST.
The red areas represent a negative correlation, and the green areas represent a positive correlation.

Number of selected grids for each feature in each case study, along with the number of features extracted for each variable for initial and cascaded PCA

| Case study | Features | No. of selected grids | No. of principal components | |
|---|---|---|---|---|
| | | | Initial PCA | Cascaded PCA |
| USGS-1010500 | Precip | 120 | 3 | |
| | $T_{avg}$ | 63 | 3 | |
| | $T_{min}$ | 100 | 3 | 21 |
| | $T_{max}$ | 50 | 3 | |
| | SST | 230 | 3 | |
| | Anomaly SST | 19 | 19 | |
| USGS-13342500 | Precip | 400 | 3 | |
| | $T_{avg}$ | 100 | 4 | |
| | $T_{min}$ | 200 | 4 | 22 |
| | $T_{max}$ | 120 | 5 | |
| | SST | 150 | 5 | |
| | Anomaly SST | 130 | 15 | |

EGU General Assembly 2025

# Two-year test dataset, Case study 1




SageFormer

Transformer

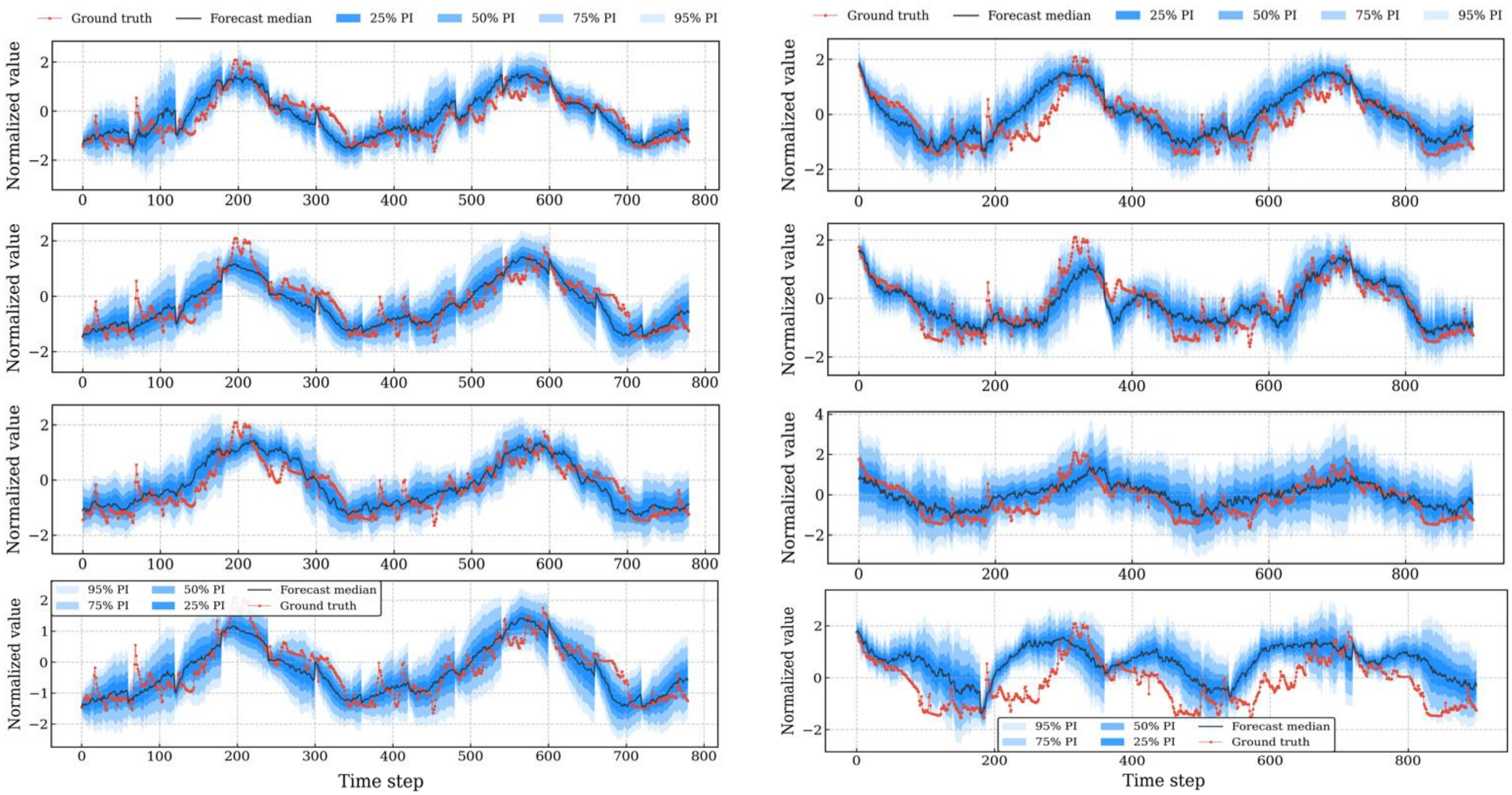
Informer

PLSTM

7-day forecasting

30-day forecasting

# Two-year test dataset, Case study 1




SageFormer

Transformer

Informer

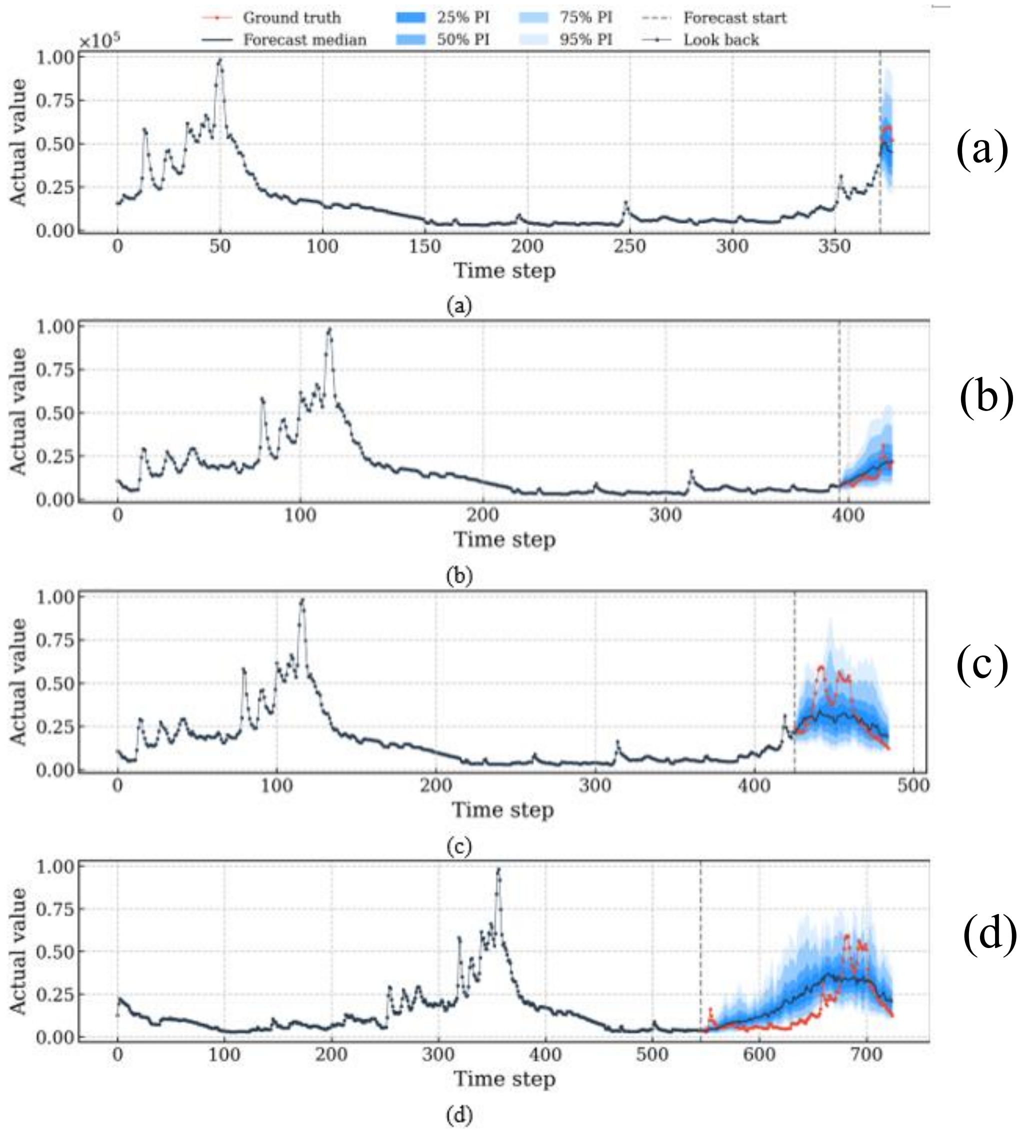
PLSTM

60-day forecasting

180-day forecasting

EGU General Assembly 2025

# Back-tested probabilistic daily forecasting




Back-tested probabilistic daily flood forecasting results for separately trained models of **SageFormer network** with different look-back periods corresponding to forecast horizons:

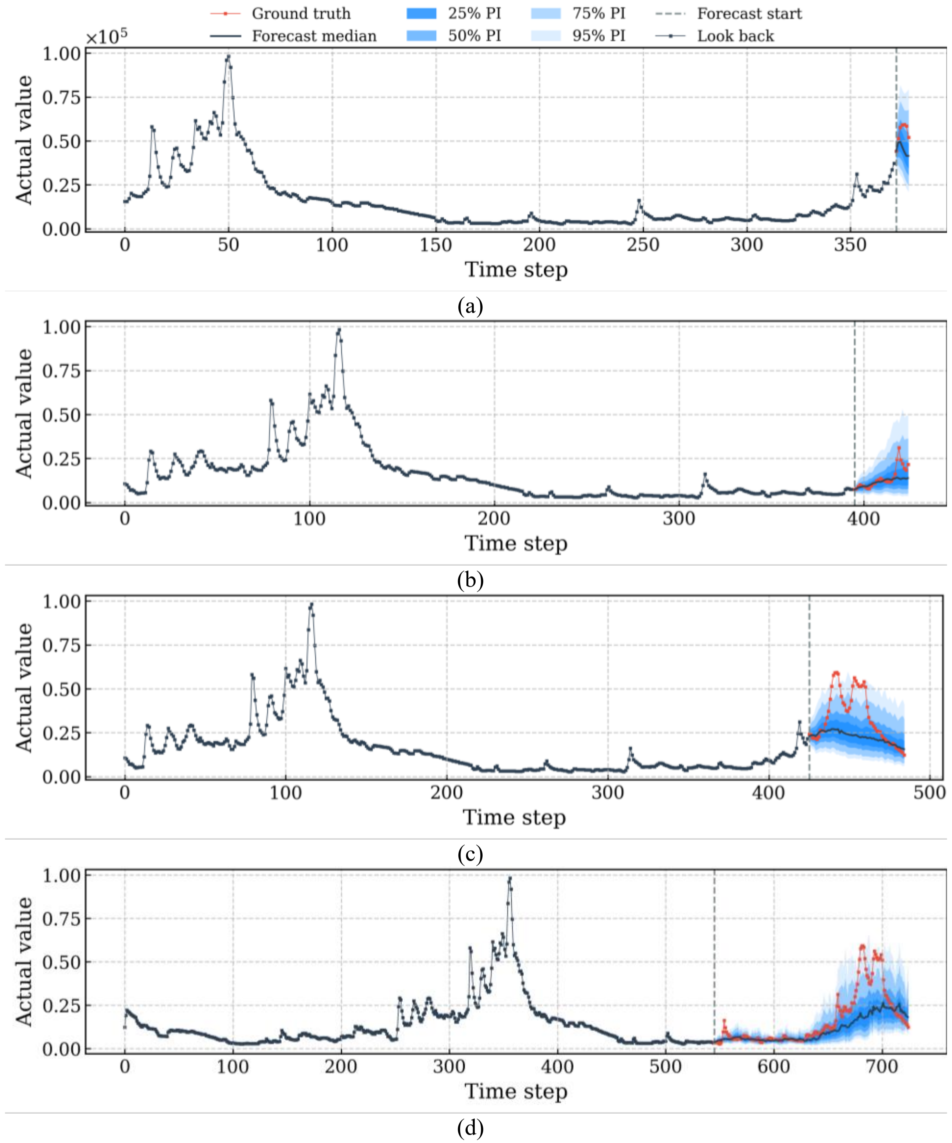
7-day forecast (372-day look-back)

30-day forecast (395-day look-back)

60-day forecast (425-day look-back)

180-day forecast (545-day look-back)

# Back-tested probabilistic daily forecasting




Back-tested probabilistic daily flood forecasting results for separately trained models of **Transformer network** with different look-back periods corresponding to forecast horizons:

(a) 7-day forecast (372-day look-back)

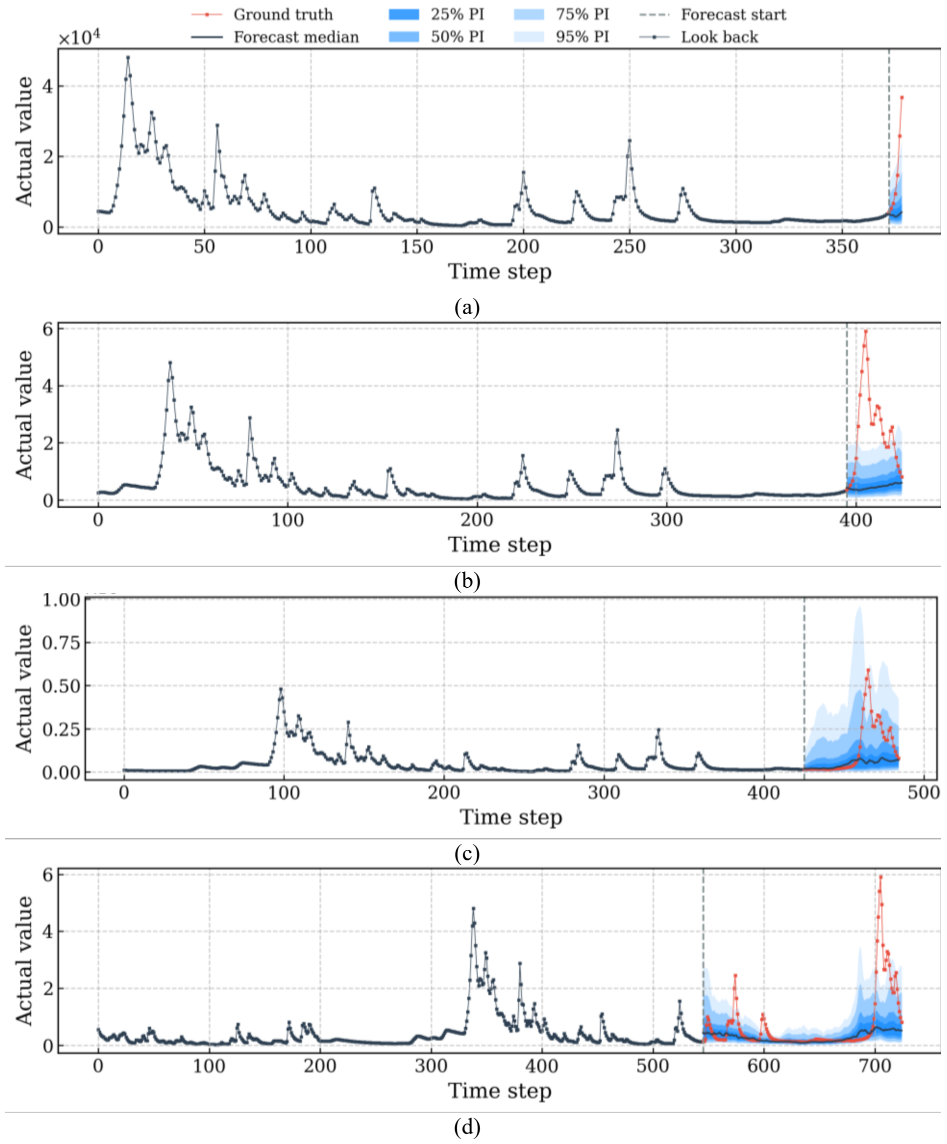
(b) 30-day forecast (395-day look-back)

(c) 60-day forecast (425-day look-back)

(d) 180-day forecast (545-day look-back)

EGU General Assembly 2025

# Back-tested probabilistic daily forecasting




(a)

(b)

(c)

(d)

Back-tested probabilistic daily flood forecasting results for separately trained models of
**PLSTM network**
with different look-back periods corresponding to forecast horizons:

7-day forecast (372-day look-back)

30-day forecast (395-day look-back)

60-day forecast (425-day look-back)

180-day forecast (545-day look-back)

EGU General Assembly 2025

# Conclusion

- **SageFormer** exhibited superior performance in terms of forecast **sharpness** (lowest MPIW) and **overall accuracy** (lowest CRPS) across all forecasting horizons, emphasizing its capability to reliably predict flood events even months in advance.

- Informer offered significant computational advantages, achieving accuracy comparable to Transformer but with reduced complexity, making it particularly suitable for resource-constrained operational contexts.

- Attention-based models effectively calibrated their uncertainty estimates, as indicated by PICP values closely matching nominal prediction intervals (75–95%), thus offering **valuable decision-making support during extreme hydrological events**.

- PLSTM, despite generating wide prediction intervals, consistently underperformed in capturing critical peaks and failed to provide precise uncertainty quantification, demonstrating inherent limitations of recurrent architectures in extended probabilistic forecasting scenarios.

- SageFormer, Transformer and Informer maintained stable performance across increasing forecast horizons, likely benefiting from their autoregressive architectures and progressively enriched historical context, underscoring the importance of attention mechanisms for managing long-range hydrological dependencies.

EGU General Assembly 2025

# 감사합니다

# THANK YOU FOR YOUR ATTENTION

Fatemeh Ghobadi ✉ *Ghobadi@khu.ac.kr*

ONPOOM

EGU General Assembly 2025