tinyML Multiobjective Optimization for Urban Infrastructure

Vipin Singh, Felix Biessmann
Cognitive Algorithms Lab (Calgo Lab)

MLaftermath Workshop
10.03.2026

BHT Berliner Hochschule für Technik

RIWWER UDE cognitive algorithms EINSTEIN CENTER Digital Future

Overview

1. Motivation
2. Data
3. Multi-Objective Optimization
   a. Neural Architecture Search
   b. Deployment on (simulated) FPGA
4. Conclusion
5. Demo

Challenges in Urban Water Management

Berliner Hochschule für Technik
Studiere Zukunft

BHT Berliner Hochschule für Technik

Urban Water Management

Fresh water supply

Wastewater management

Storm water drainage

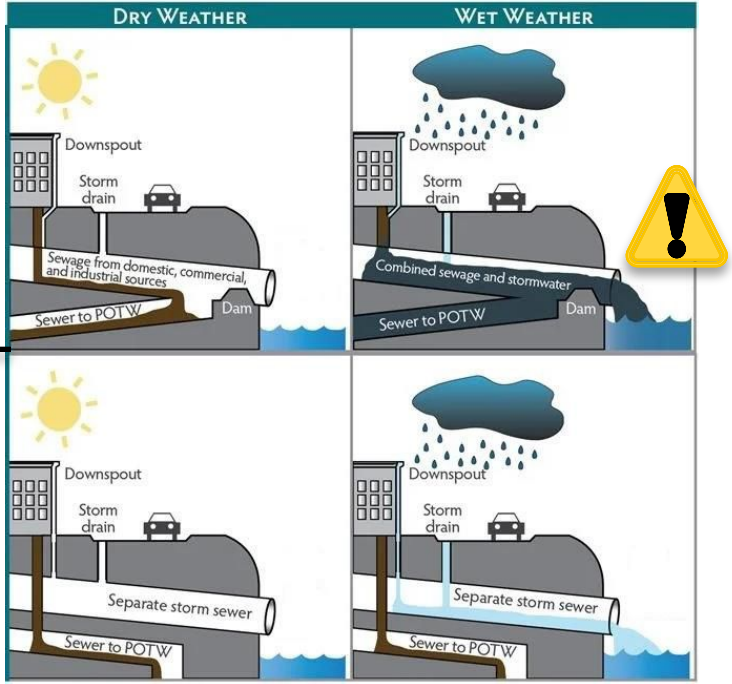
Today's Challenges

Combined Sewer System (CSS)

carries both stormwater and wastewater in one pipe.

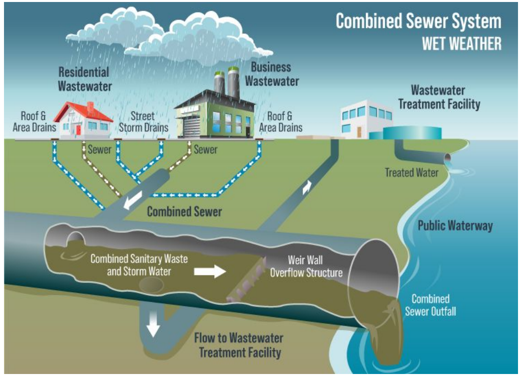
Separate Sewer System (SSS)

separates stormwater and wastewater into two different pipes.



Today's Challenges

- Rapid Urbanization → More people in cities
- Climate change → Heavy rainfall
- Increased strain on aging Combined Sewer Systems

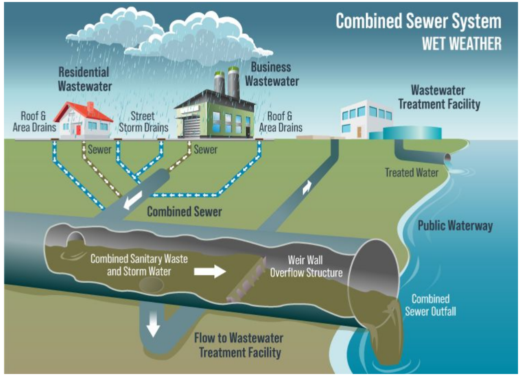


tinyML Multiobjective Optimization for Urban Infrastructure - Vipin Singh, Felix Biessmann

Berliner Hochschule für Technik
Studiere Zukunft

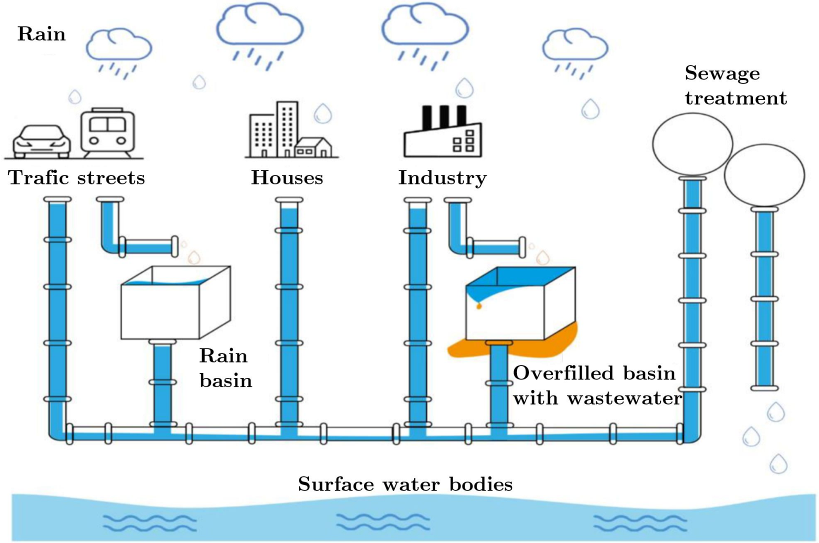
Today’s Challenges

- Rapid Urbanization → More people in cities
- Climate change → Heavy rainfall
- Increased strain on aging Combined Sewer Systems
- Combined Sewer Overflows



Countermeasures for Combined Sewer Overflows

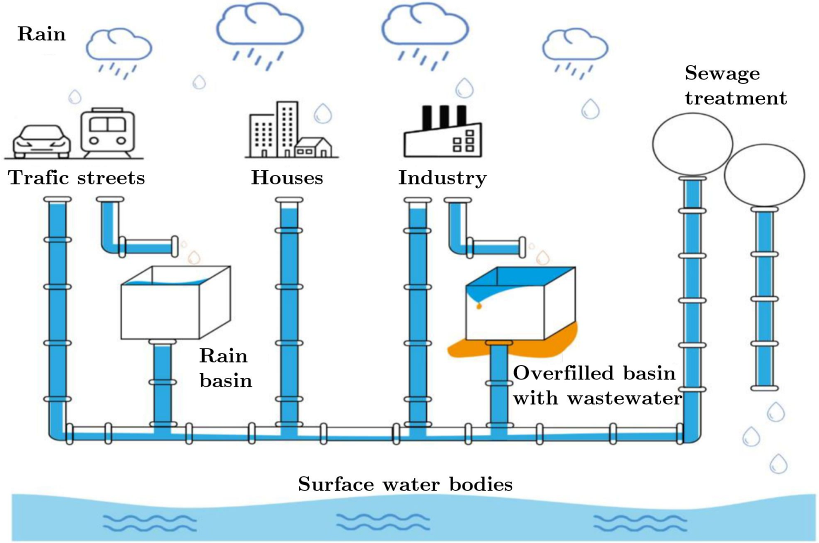
- Expand wastewater / stormwater systems:
  - Rebuild to today's requirements
  - Increase sewer & treatment capacity



Countermeasures for Combined Sewer Overflows

- Expand wastewater / stormwater systems:
  - Rebuild to today's requirements
  - Increase sewer & treatment capacity

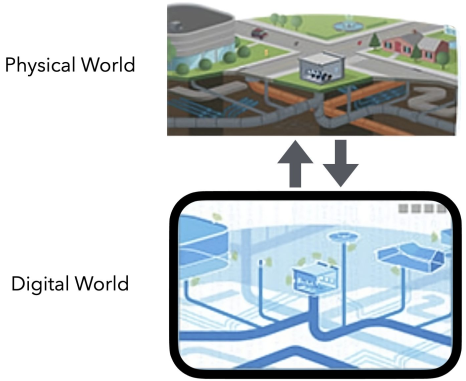
- Improve existing system efficiency:
  - Optimize load distribution
  - ➡ Monitor & forecast system dynamics



Current Modeling Approaches & Limitations

Traditional Physics-Based Models:

- Costly to implement & maintain
- Difficult to adapt to evolving dynamics



Machine Learning Alternatives:

- Greater adaptability
- However, challenges remain:
  - Limited interpretability
  - Robustness to real-world data issues

Data

Understanding Time Series Forecasting

Time Series Forecasting:

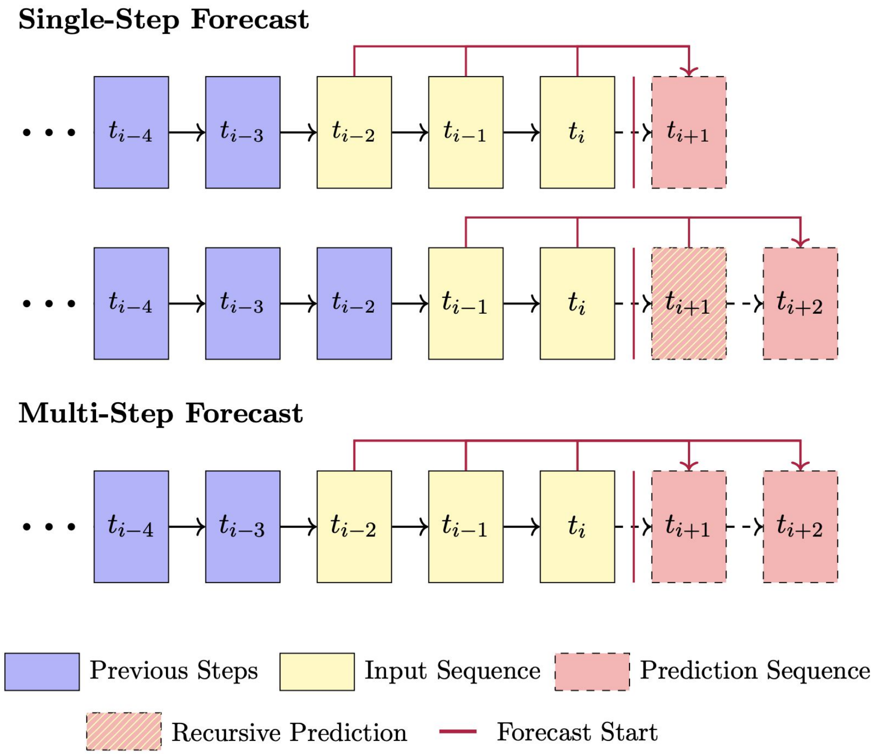
- Past Data → Future Values
- Models learn Temporal Patterns

Covariates:

- Past-known or Future-known

Study goal:

Forecast wastewater levels - 12h ahead



Time Series Data of a Combined Sewer System

Provided by *Wirtschaftsbetriebe Duisburg*

- 3 years
- 35 sensors at 6 locations, including:
  - Rainfall measurements
  - Pump activity
  - Filling levels (target)

Inclusion of historical rain forecasts as future covariate

District: Vierlinden, Duisburg, Germany

Multiobjective Optimization

Neural Architecture Search

Modeling Concepts

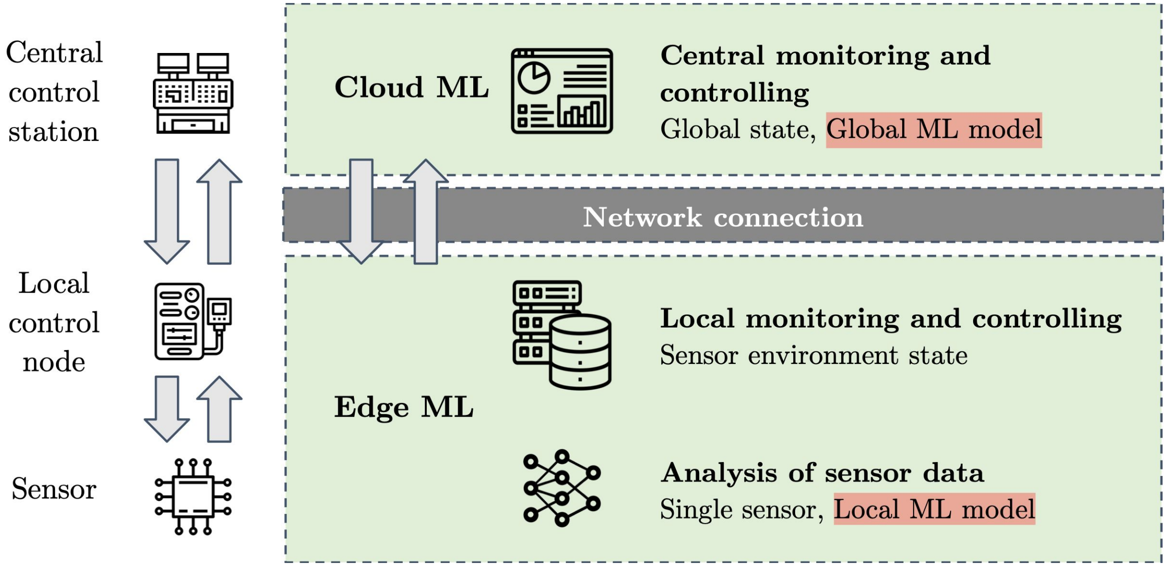
- Comparison of six neural network architectures

- C1) Global and Local Models
- C2) Peak Event Analysis
- C3) Error Models for Robustness Analysis

C1) Global and Local Models

Two scenarios for continuous operation:

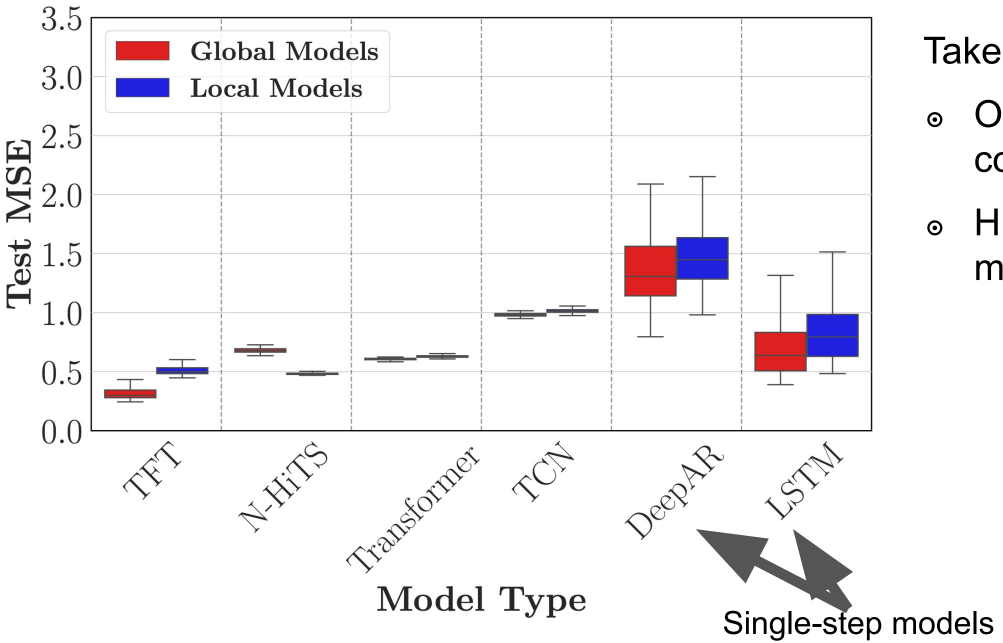
- Global:
  - All features
- Local:
  - Only target



•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2024). *Data-driven modeling of combined sewer systems for urban sustainability: An empirical evaluation. The 2nd Workshop on Public Interest AI at 47th German Conference on AI (KI 2024)*
•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2025). *Evaluating Time Series Models for Urban Wastewater Management: Predictive Performance, Model Complexity and Resilience. 10th International Conference on Smart and Sustainable Technologies (SpliTech) 2025*

Predictive Performance: Overall MSE

C1
C2
C3



Takeaways:

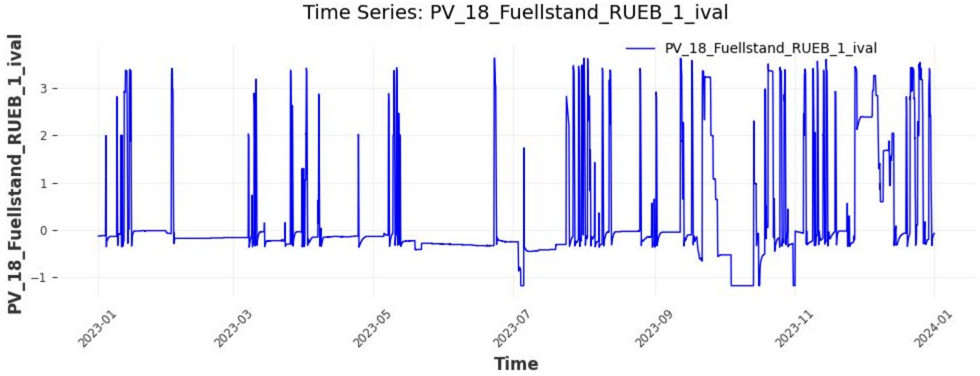
- Overall MSE: Local models are comparable to Global models
- High Variability: DeepAR and LSTM models (Single-step models)

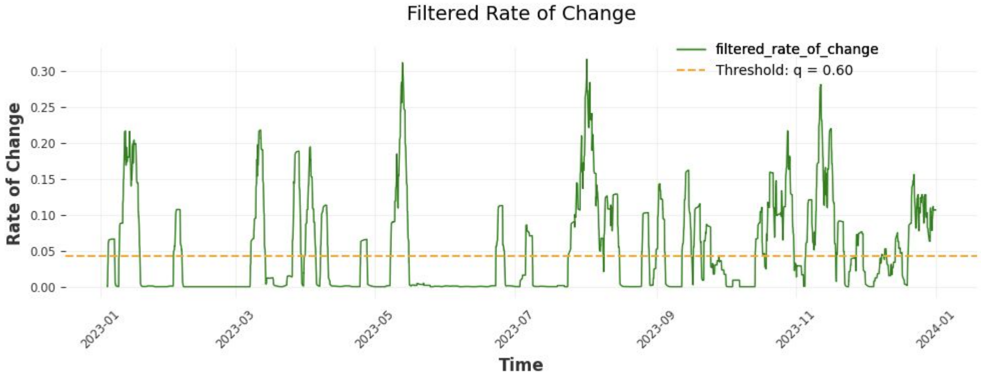
(smaller is better)

C2) Peak Event Analysis

Identifying Peak Events:

- Smoothed Rate of Change
- Evaluation on events above a threshold



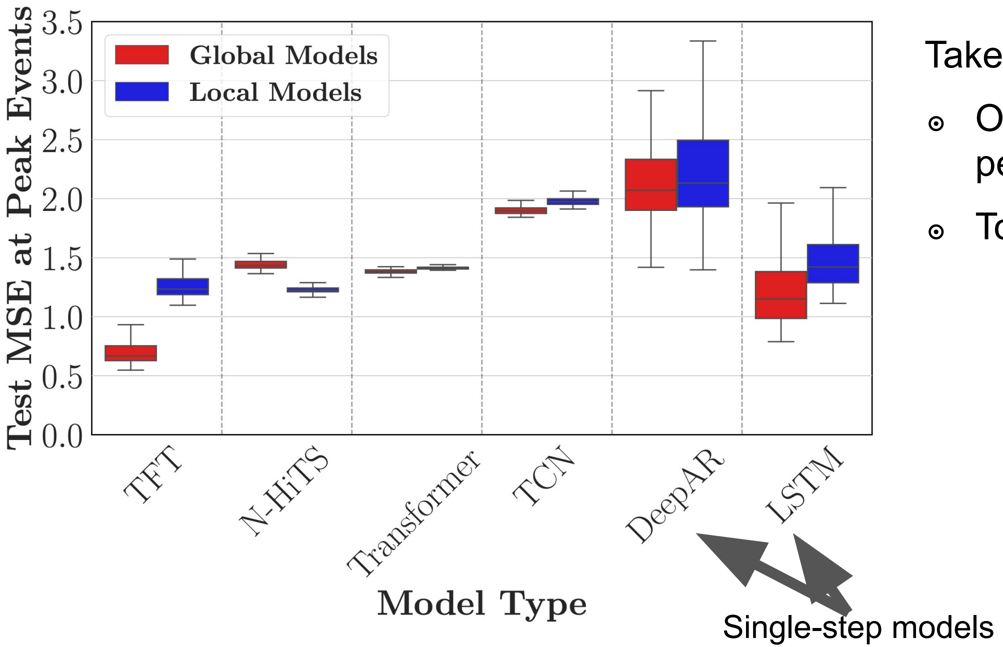


C1
C2
C3

•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2025). *Evaluating Time Series Models for Urban Wastewater Management: Predictive Performance, Model Complexity and Resilience. 10th International Conference on Smart and Sustainable Technologies (SpliTech) 2025*

Predictive Performance: Peak Events

C1
C2
C3



Takeaways:

- On Peaks: Global models generally perform better
- Top Performer: Global TFT models

(smaller is better)

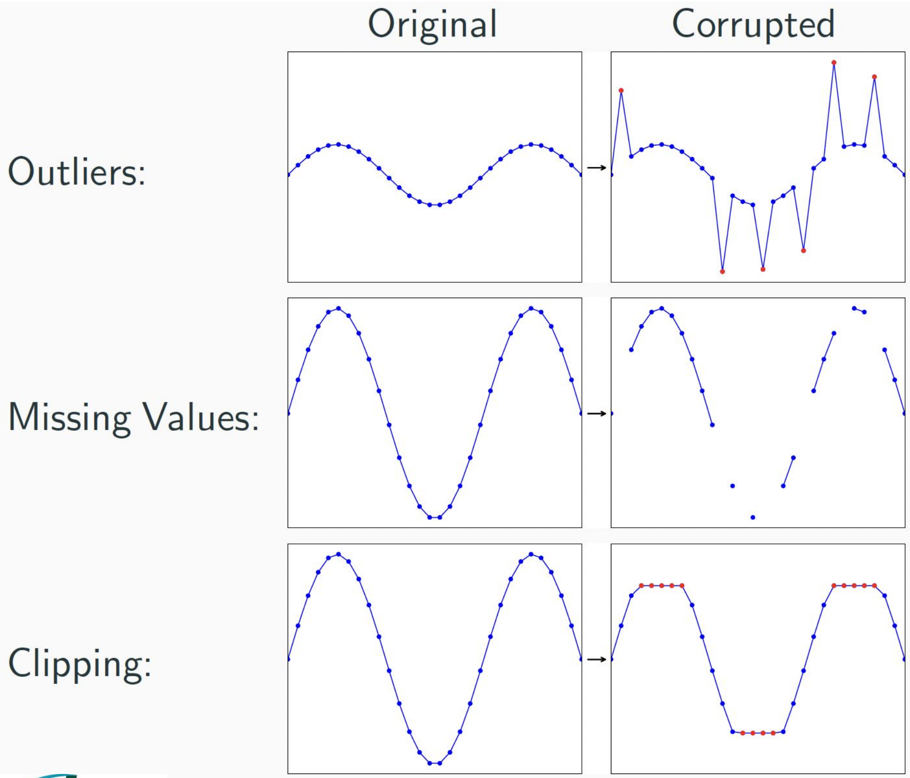
C3) Error Models for Robustness Analysis

C1
C2
C3

Simulation of real-world data quality issues

- Noise, Sensor malfunctions
- System maintenances
- Physical Boundaries

- Evaluated for global models



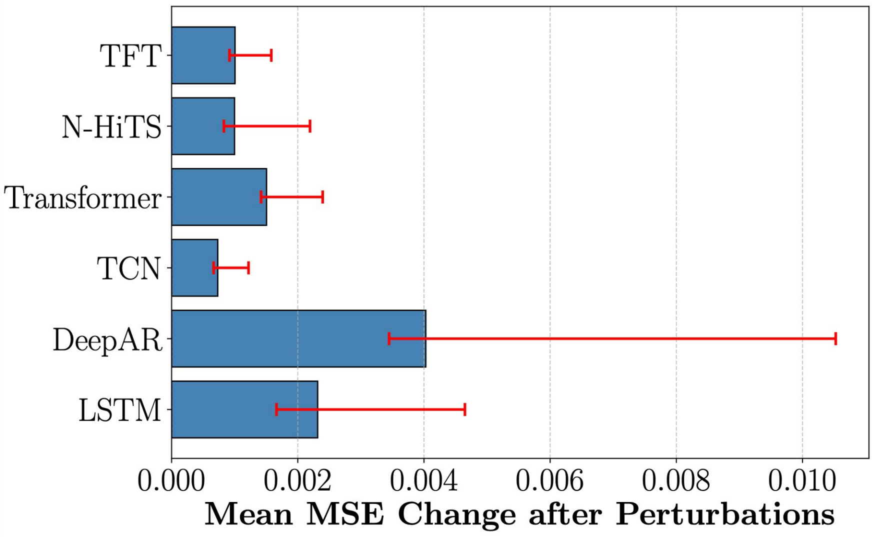
•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2025). *Evaluating Time Series Models for Urban Wastewater Management: Predictive Performance, Model Complexity and Resilience. 10th International Conference on Smart and Sustainable Technologies (SpliTech) 2025*

Robustness to Perturbations (Error Models)

C1

C2

C3



Takeaways:

- Most Robust: TCN, followed by TFT and N-HiTS
- Most Sensitive: DeepAR and LSTM models (Single-step models)

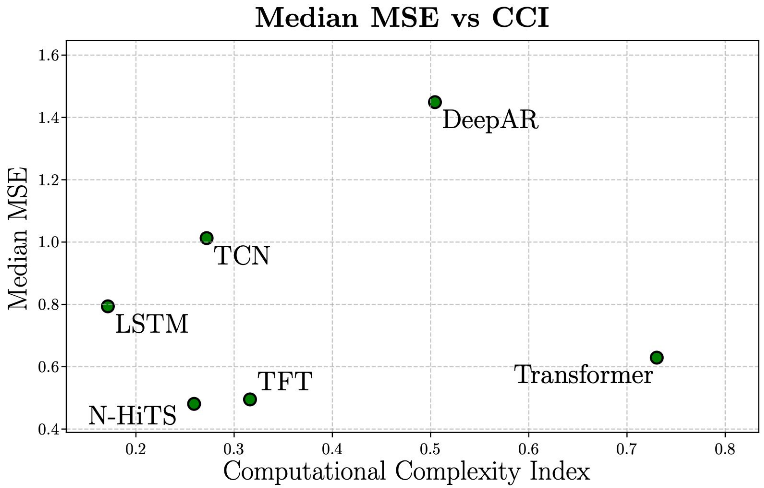
(smaller is better)

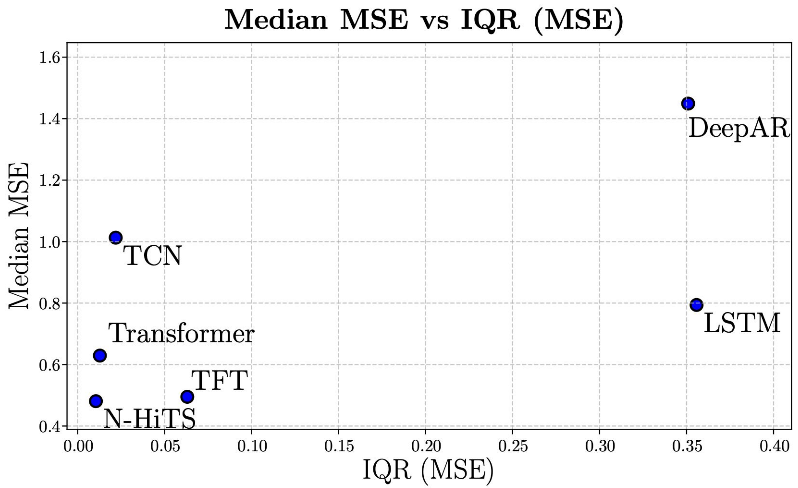
Summarized comparison of the models

- Precision of the forecast: MSE
- Computational complexity: Inference Time and Model Size
- Robustness
  - Local: Spread of the forecasts
  - Global: Sensitivity against data perturbations

Trade-offs: Precision, Complexity, Robustness (Local Models)

MSE, Computational Complexity Index (CCI) and IQR of MSE



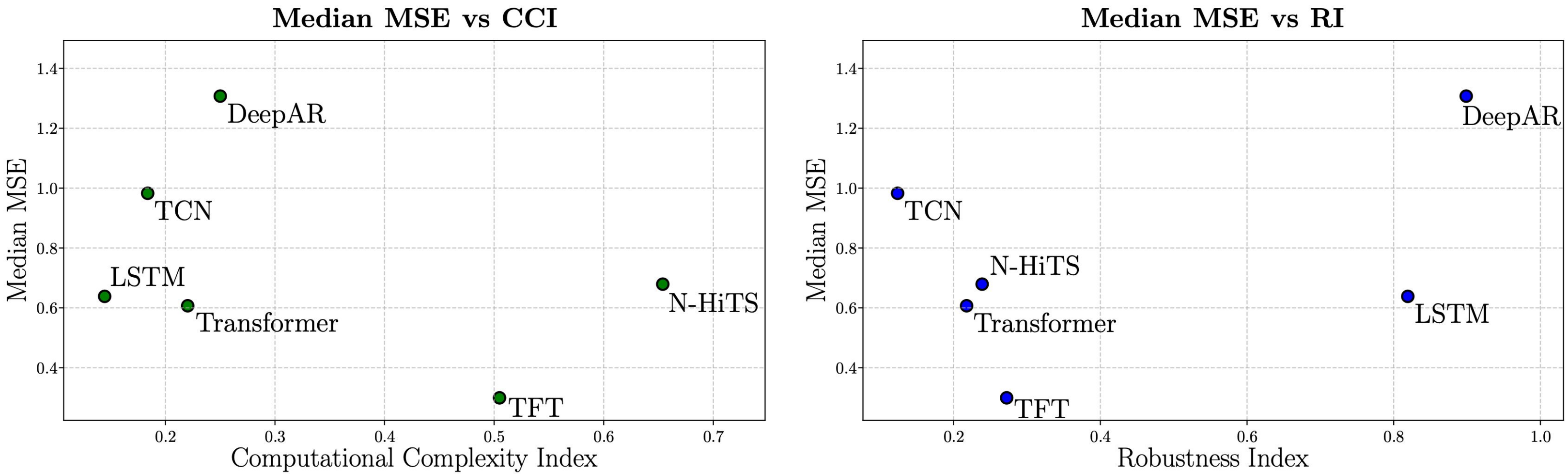


- N-HiTS and TFT exhibit the best trade-off

(smaller is better)

Trade-offs: Precision, Complexity, Robustness (Global Models)

MSE, Computational Complexity Index (CCI) and Robustness Index (RI)



- TFT: Poor complexity, but trade-off + Best in Peak Events
- Transformer: Best balance but poorer predictive performance

(smaller is better)

Multiobjective Optimization
Deployment on FPGA

BHT
Berliner Hochschule für Technik

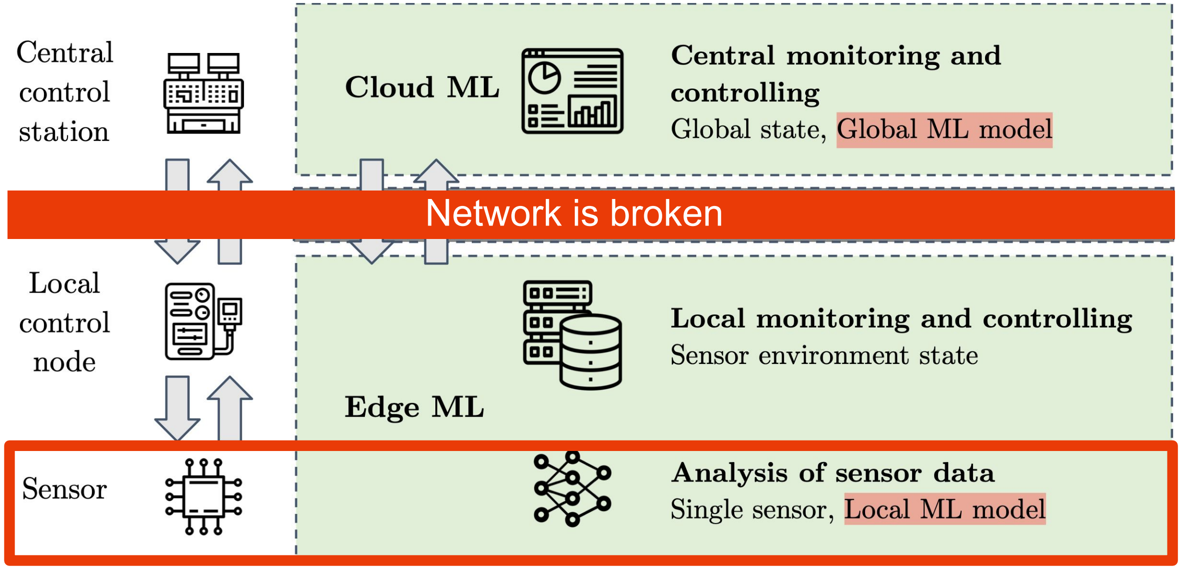
cognitive algorithms

EINSTEIN CENTER Digital Future

Hardware-Aware Model Optimization

Two scenarios for continuous operation:

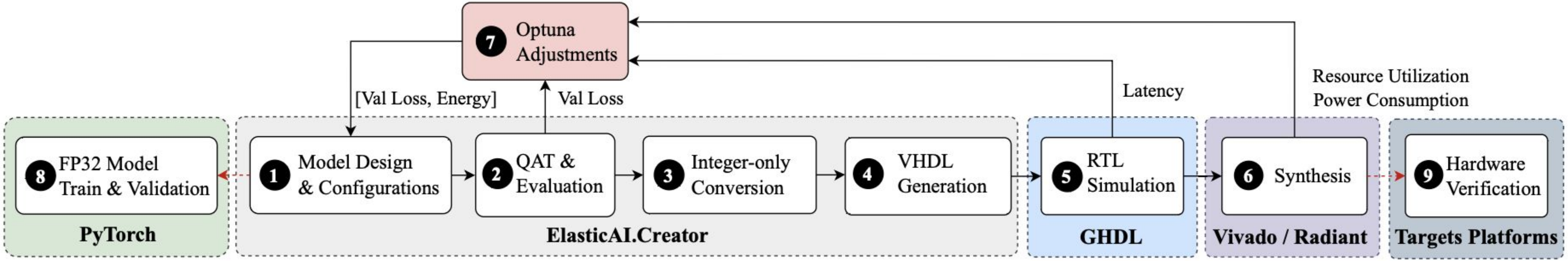
- Global:
  - All features
- Local:
  - Only target



•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2024). *Data-driven modeling of combined sewer systems for urban sustainability: An empirical evaluation. The 2nd Workshop on Public Interest AI at 47th German Conference on AI (KI 2024)*

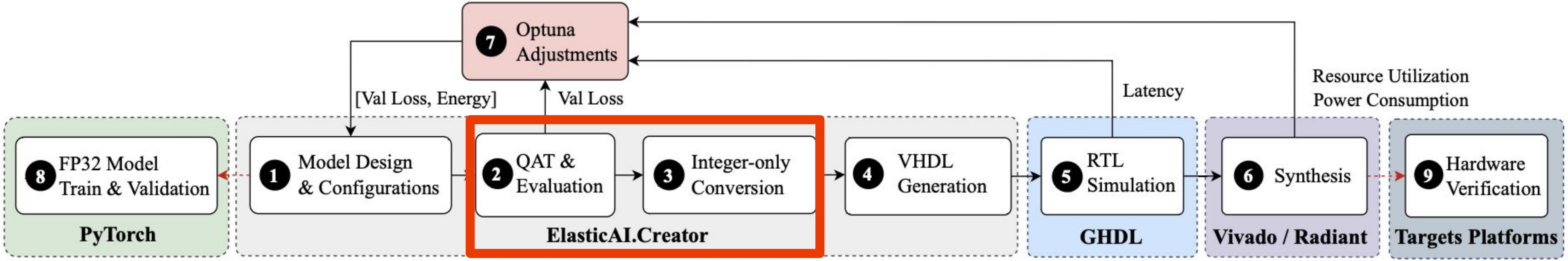
•Singh, V., Ling, T., Chiaburu, T., & Biessmann, F. (2025). *Evaluating Time Series Models for Urban Wastewater Management: Predictive Performance, Model Complexity and Resilience. 10th International Conference on Smart and Sustainable Technologies (SpliTech) 2025*

Hardware-Aware Model Optimization



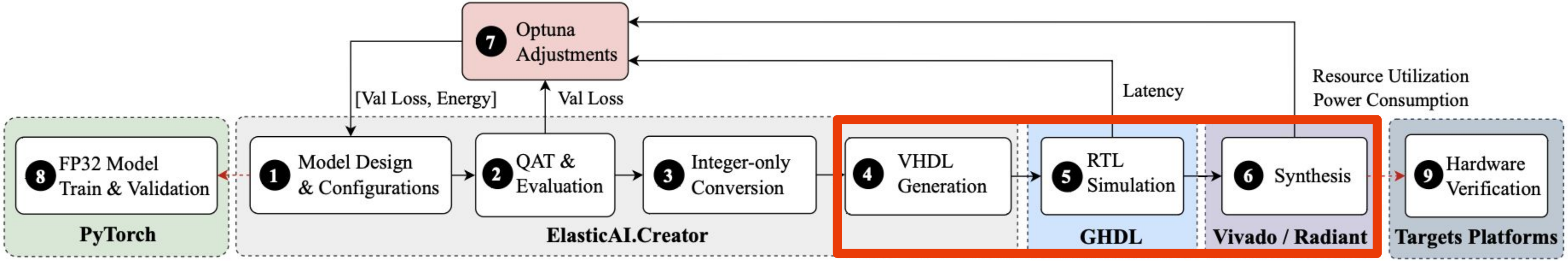
- *Ling, T., Qian, C., & Schiele, G. (2024, July). Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 38-44). IEEE*
- *Ling, T., Singh, V., Qian, C., Biessmann, F., & Schiele, G. (2025). Automated Energy-Aware Time-Series Model Deployment on Embedded FPGAs for Resilient Combined Sewer Overflow Management. In 2025 IEEE International Smart Cities Conference (in press). IEEE*

Hardware-Aware Model Optimization



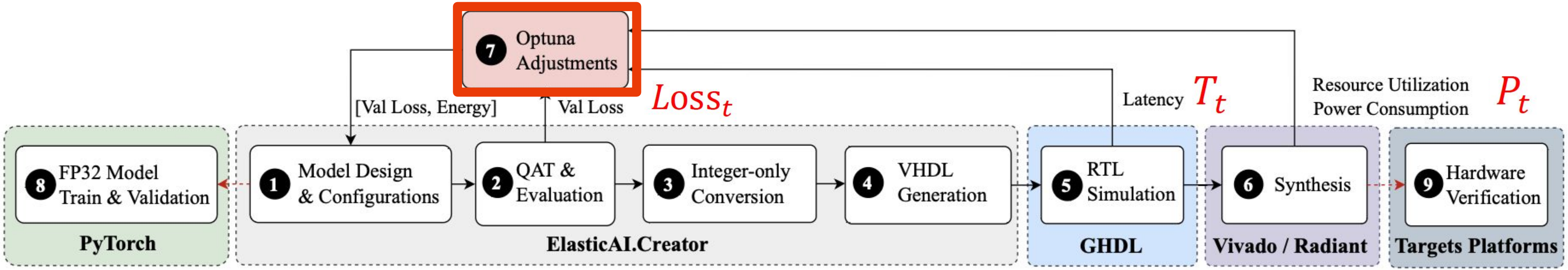
- Ling, T., Qian, C., & Schiele, G. (2024, July). *Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 38-44). IEEE*
- Ling, T., Singh, V., Qian, C., Biessmann, F., & Schiele, G. (2025). *Automated Energy-Aware Time-Series Model Deployment on Embedded FPGAs for Resilient Combined Sewer Overflow Management. In 2025 IEEE International Smart Cities Conference (in press). IEEE*

Hardware-Aware Model Optimization



- *Ling, T., Qian, C., & Schiele, G. (2024, July). Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 38-44). IEEE*
- *Ling, T., Singh, V., Qian, C., Biessmann, F., & Schiele, G. (2025). Automated Energy-Aware Time-Series Model Deployment on Embedded FPGAs for Resilient Combined Sewer Overflow Management. In 2025 IEEE International Smart Cities Conference (in press). IEEE*

Hardware-Aware Model Optimization



- Multiobjective Optimization: Optuna
  - Objective 1: Minimize loss
  - Objective 2: Minimize energy consumption

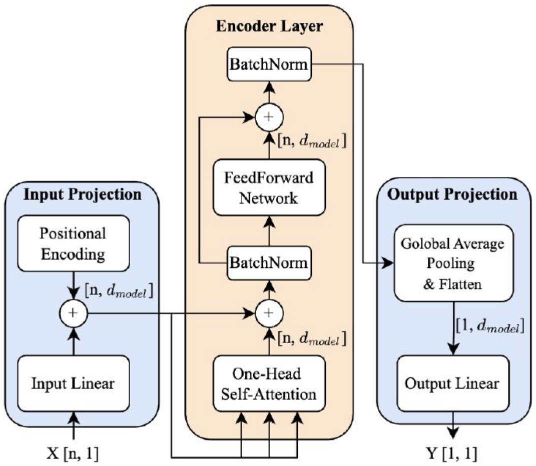
$Loss_t$

$E_t = P_t \times T_t$

- Ling, T., Qian, C., & Schiele, G. (2024, July). *Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 38-44). IEEE*
- Ling, T., Singh, V., Qian, C., Biessmann, F., & Schiele, G. (2025). *Automated Energy-Aware Time-Series Model Deployment on Embedded FPGAs for Resilient Combined Sewer Overflow Management. In 2025 IEEE International Smart Cities Conference (in press). IEEE*

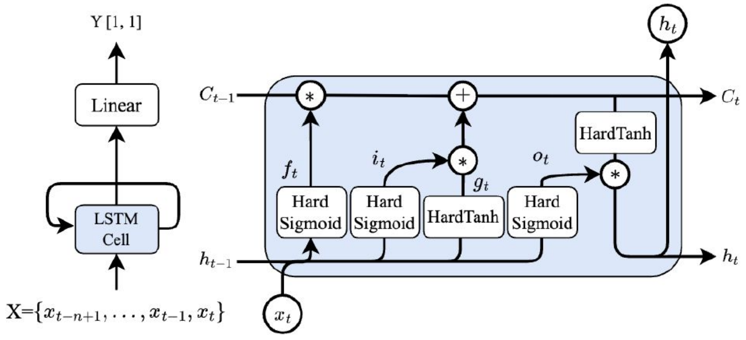
The Lightweight Models for Edge Deployment

Encoder-only Transformer model



- *Ling, T., Qian, C., & Schiele, G. (2024, July). Integer-only Quantized Transformers for Embedded FPGA-based Time-series Forecasting in AIoT. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 38-44). IEEE*

Single-layer LSTM model



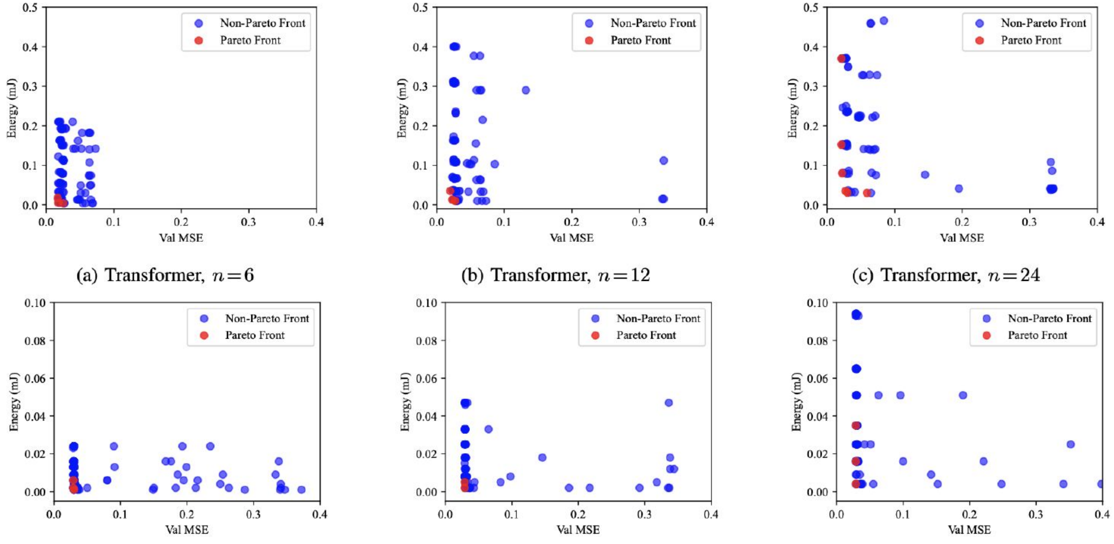
- *Qian, C., Ling, T., & Schiele, G. (2024). Exploring energy efficiency of LSTM accelerators: A parameterized architecture design for embedded FPGAs. Journal of Systems Architecture, 152, 103181.*

Hardware-Aware Model Optimization Results

Transformers* vs LSTMs*:

- LSTMs show greater variance in MSE
- LSTMs are more energy efficient than Transformer

*Lightweight variations, with evaluation on one-step forecasts only



(a) Transformer, $n=6$ (b) Transformer, $n=12$ (c) Transformer, $n=24$

(d) LSTM, $n=6$ (e) LSTM, $n=12$ (f) LSTM, $n=24$

Hardware-Aware Model Optimization Results

Model	n	Configuration				Test MSE			LUTs (%)	BRAMs (%)	DSPs (%)	Energy (mJ)	Power* (mW)	Latency** (ms)
		b	bs	lr ($\times 10^{-4}$)	$d_{\text{model}}/h_{\text{size}}$	FP32	Quantized	Variance (%)						
Transformer	6	6	80	3.653	8	0.0415	0.0427	↑2.89	43.56	15	95	0.004	49.0	0.091
	12	8	144	1.410	16	0.0394	0.0394	−0.00	55.88	100	100	0.036	67.0	0.532
	24	8	80	7.551	40	0.0339	0.0376	↑10.91	84.91	100	100	0.370	72.0	5.134
LSTM	6	8	112	1.210	16	0.0432	0.0433	↑0.23	32.84	0	55	0.002	48.0	0.046
	12	8	144	1.325	8	0.0433	0.0436	↑0.69	30.14	15	55	0.002	50.0	0.039
	24	8	256	9.803	16	0.0434	0.0432	↑0.46	32.53	5	55	0.009	49.0	0.182

Implementation Details

Implementation Details - Packages for Python

- For data exploration:
  - pip install ydata-profiling
- For Time Series models and data loading:
  - pip install darts
- For generating errors in data*:
  - pip install tab-err
- For experiments tracking:
  - pip install wandb
- For hyperparameter optimization:
  - pip install optuna
- For building demonstrators:
  - pip install streamlit

Conclusion

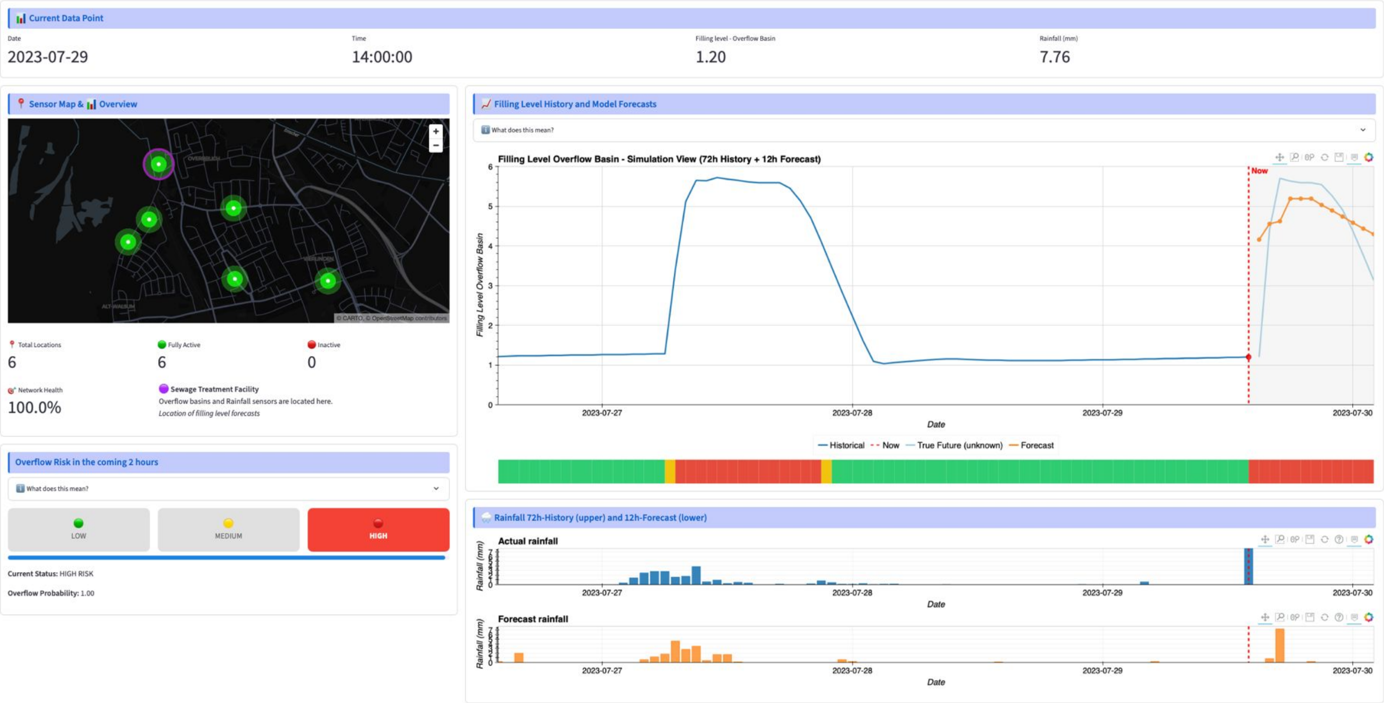
Key Findings

- Global models achieve superior accuracy, particularly for peak events
- Local models demonstrate resilience and offer a solution as fail-safes
- End-to-end local forecasting system for sewer overflow monitoring
- Cloud-free inference balancing performance, robustness and complexity

Practical Implications

- Machine learning models lower the cost of overflow management, benefitting public health
- Global models during normal operations
- Local models as backup during failure
- Deployment of local models with optimized energy consumption

Demonstrator: Showcasing Model Forecasts



riwwer.demo.calgo-lab.de

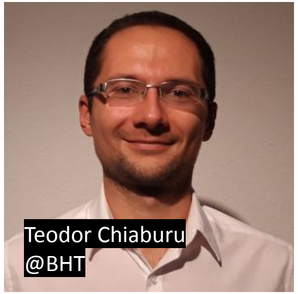
Future Work

- Generalize Findings: Data from other cities

- Complexer perturbation dynamics, e. g. Grouped feature perturbations


Vipin Singh
@BHT


Tianheng Ling
@UDE


Teodor Chiaburu
@BHT


Chao Qian
@UDE


Prof. Dr. Gregor Schiele
@UDE


Prof. Dr. Felix Biessmann
@BHT, ECDF

Thank You! Questions?

- Contact: Vipin Singh - vipin.singh@bht-berlin.de; Felix Biessmann - felix.biessmann@bht-berlin.de
- Code & Further Details:

Github Repository