

# AI for detecting abnormal behavior in a sewerage network - Robustness against model evasion attacks

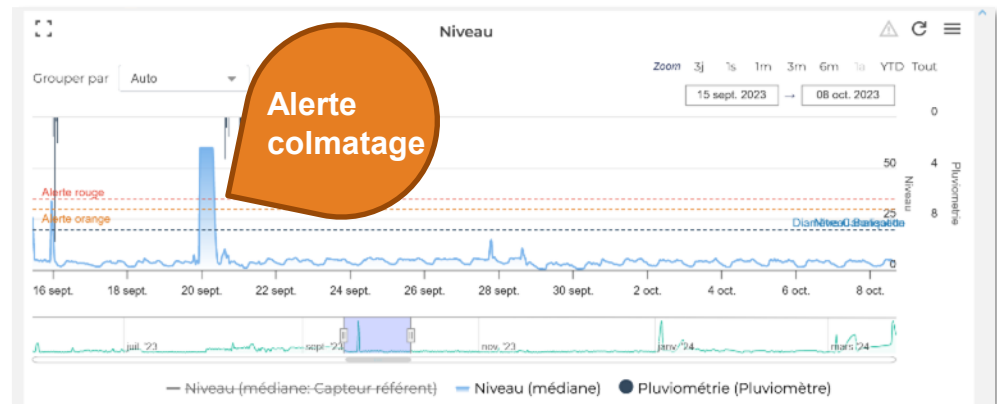
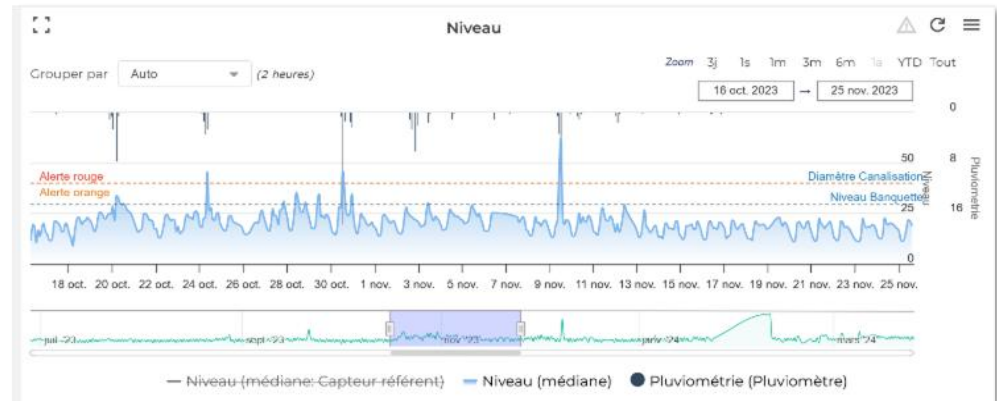
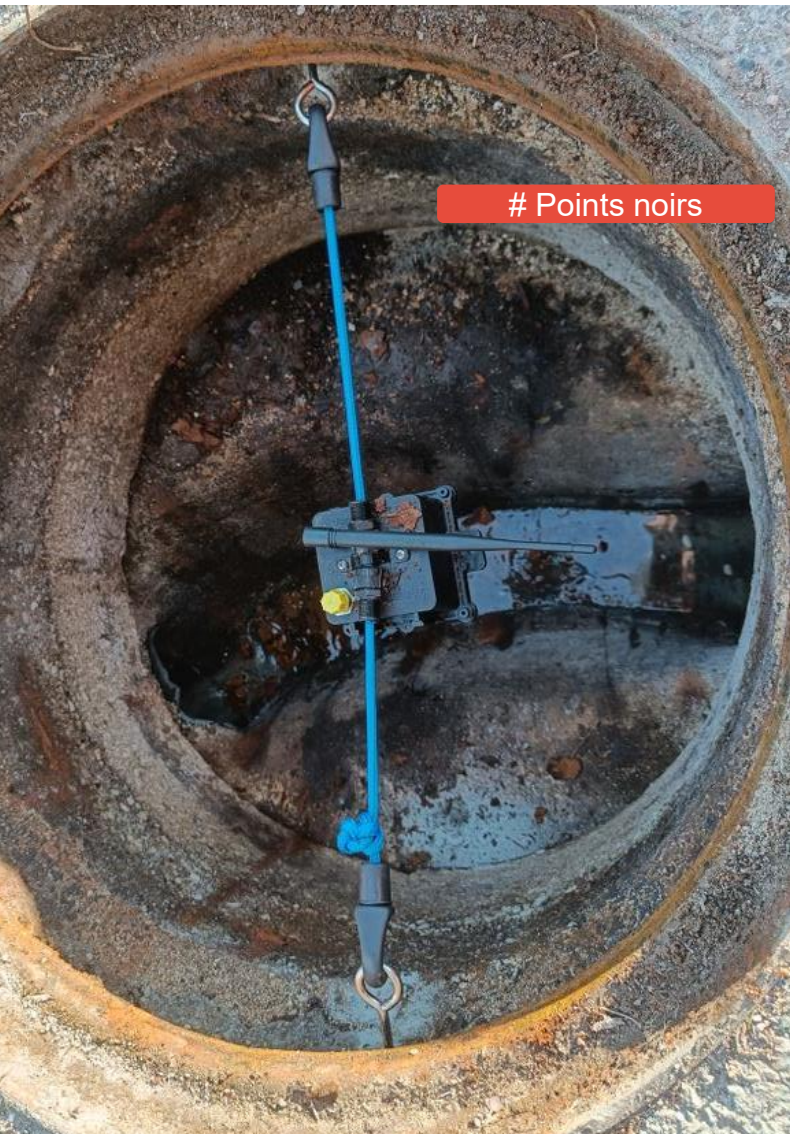
BOUSSAD ADDAD (CORTAIX LABS)  
FRANCOIS HAMON (GREENCITYZEN)



## Context: Reduce operating costs at critical points in the sewer network



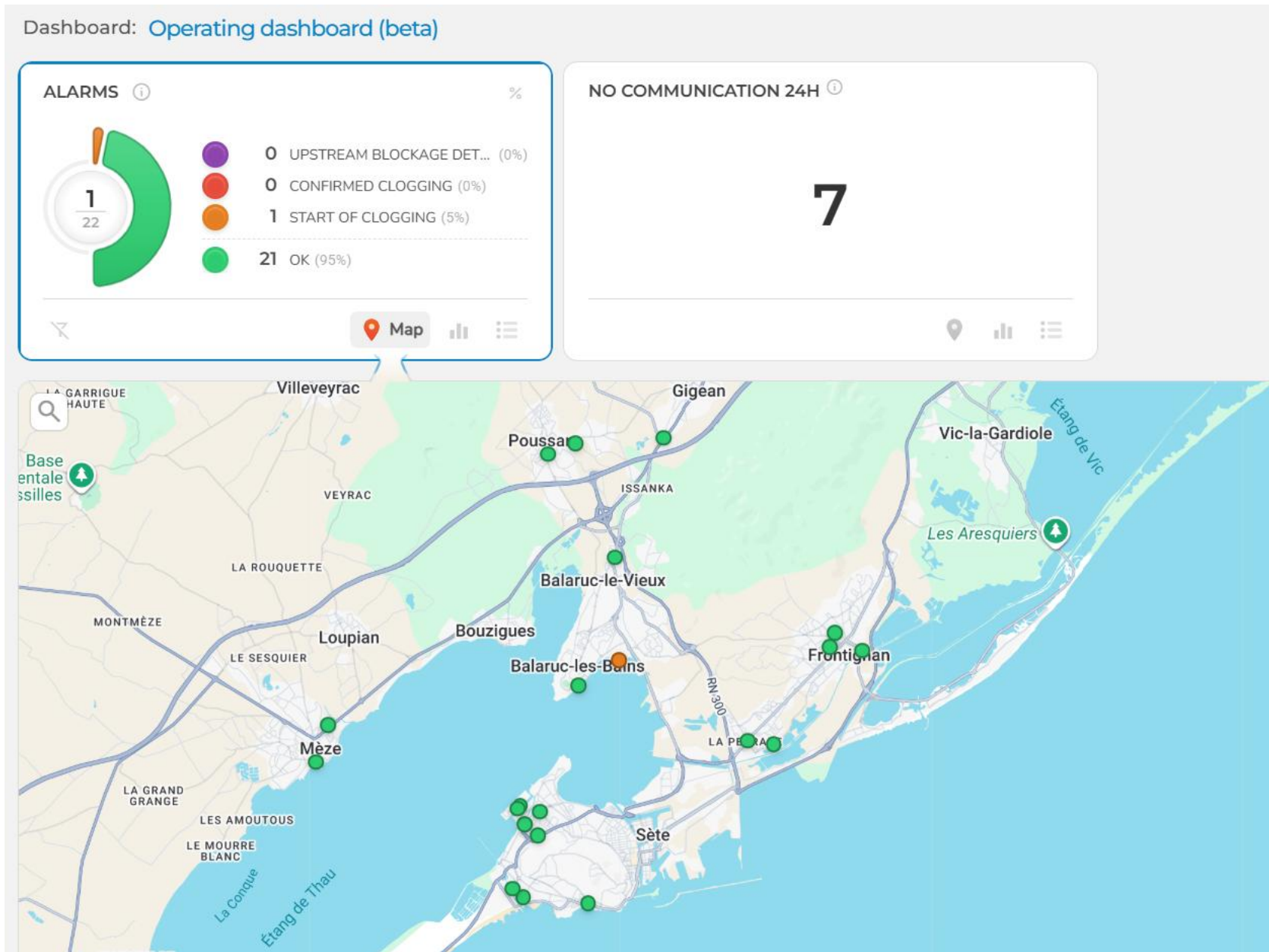
# Typical hot spot with a typical network load increase pattern



Mesure de niveau & pluviométrie

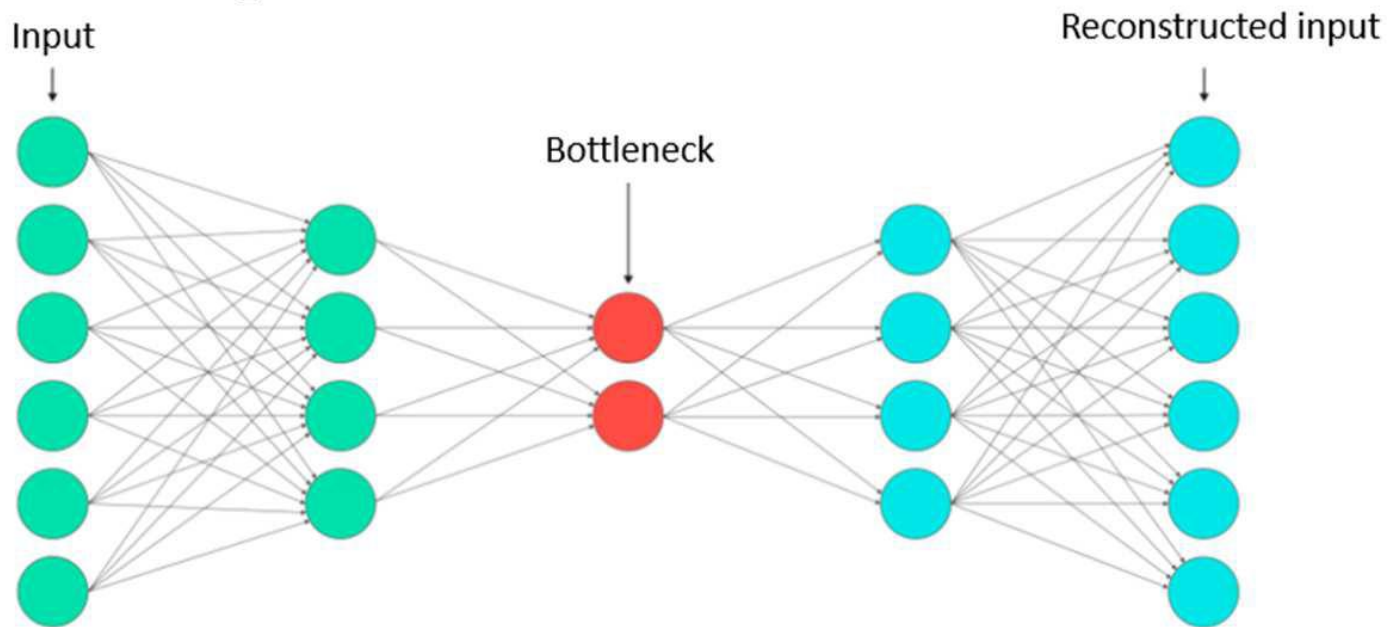


# Automatically detect abnormal behavior over a large number of sites



## Tried two different architectures

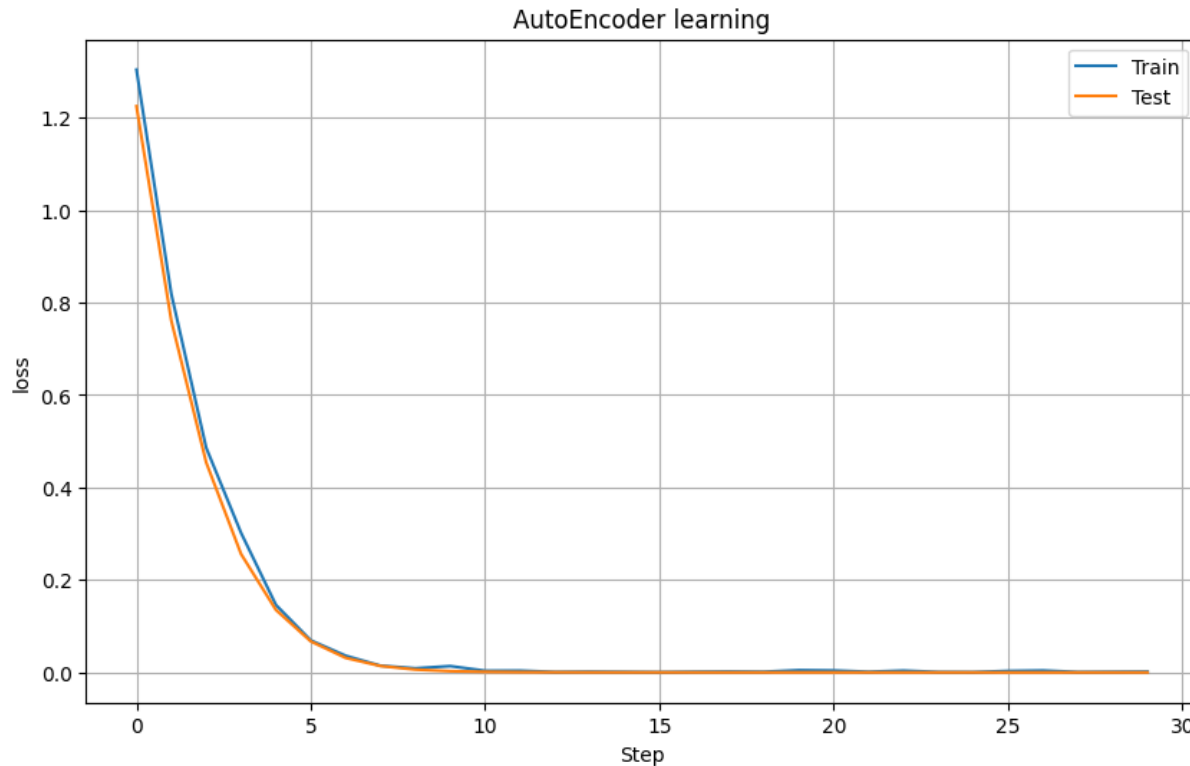
- Encoder-Decoder RNN → not stable
- **Convolutional auto-Encoder** → much better performance
- The training is made on normal data only (waste water level and rainfall)



# Model building: training results

## Training on normal data only

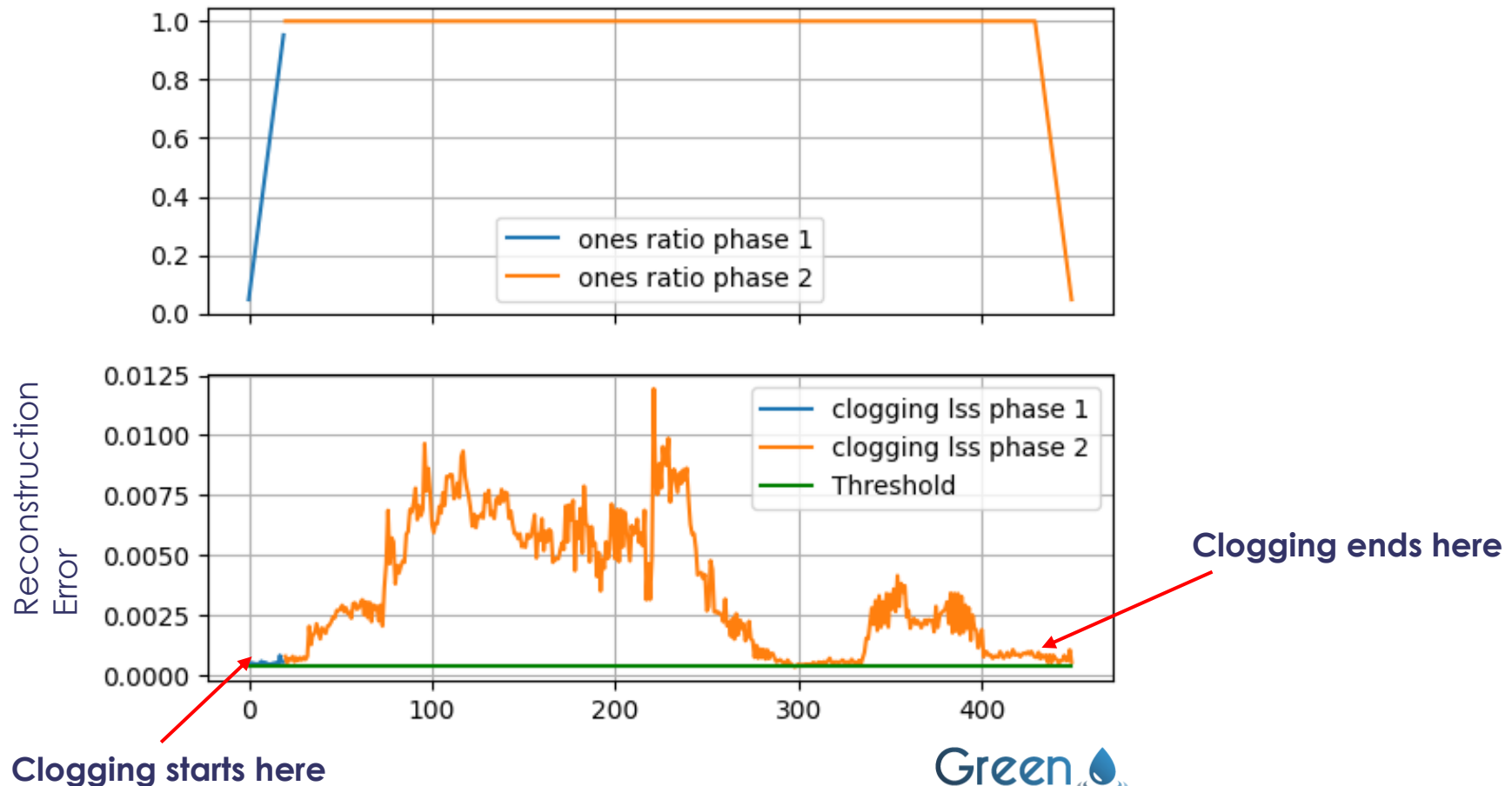
➤ Train and test losses are good



# Model building: the model usage for clogging detection

## Reconstruction error is used to detect the clogging

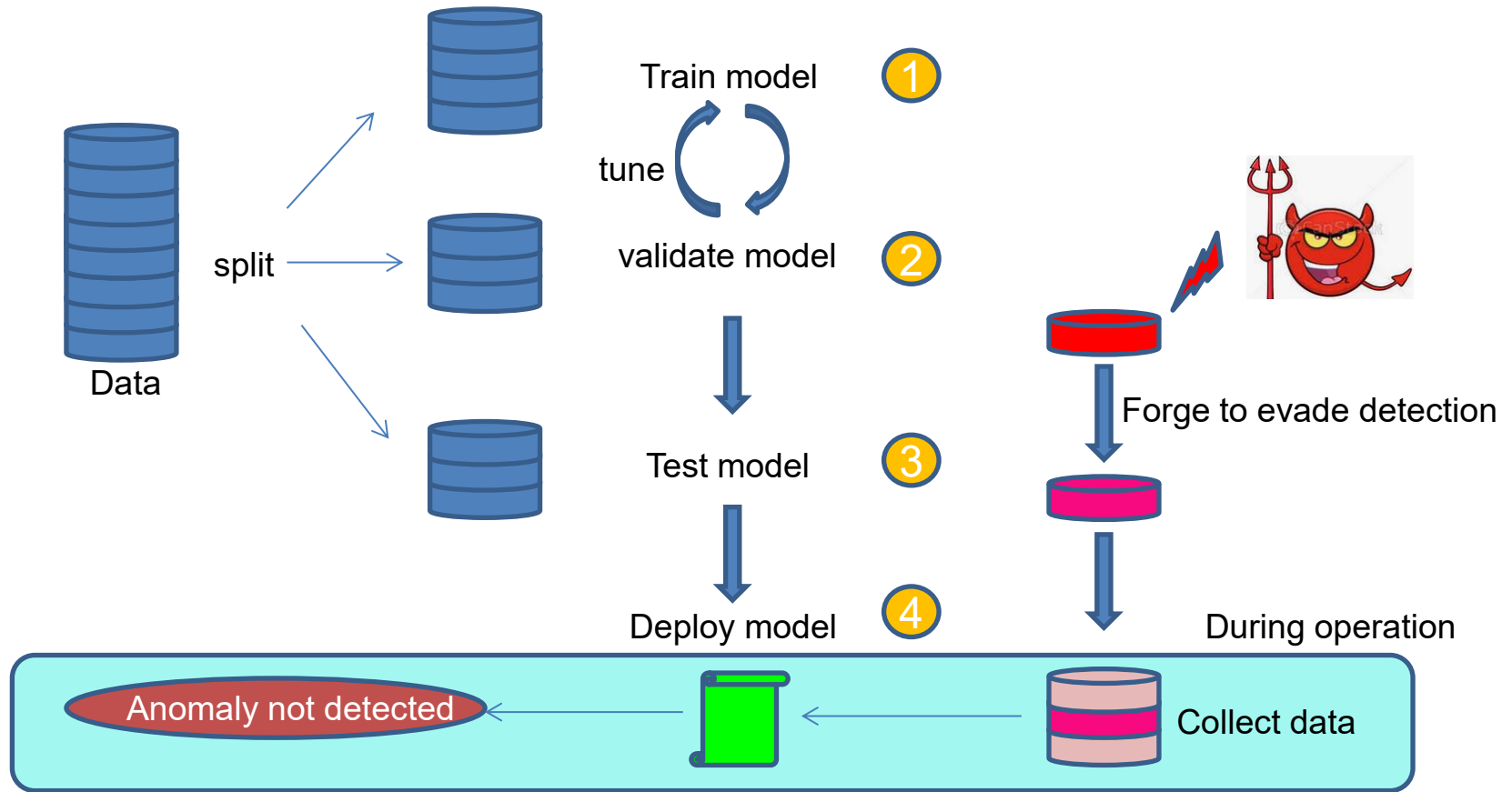
- Reconstruction error > Threshold → Clogging
- More than 90% TP detection rate, less than 3% FP rate



## Evasion attacks against anomaly detectors



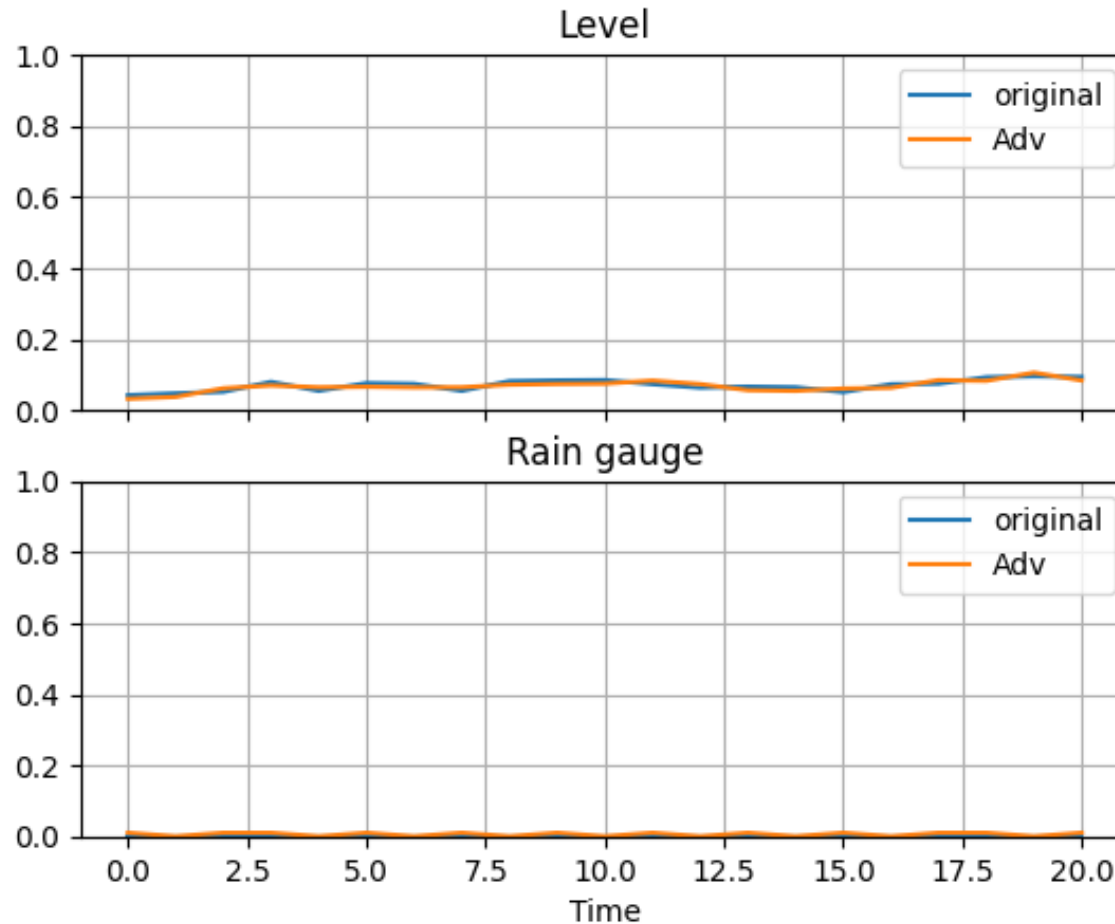
# Recall: Model evasion attacks on ML (after deployment of the model)



# Adversarial evasion attacks

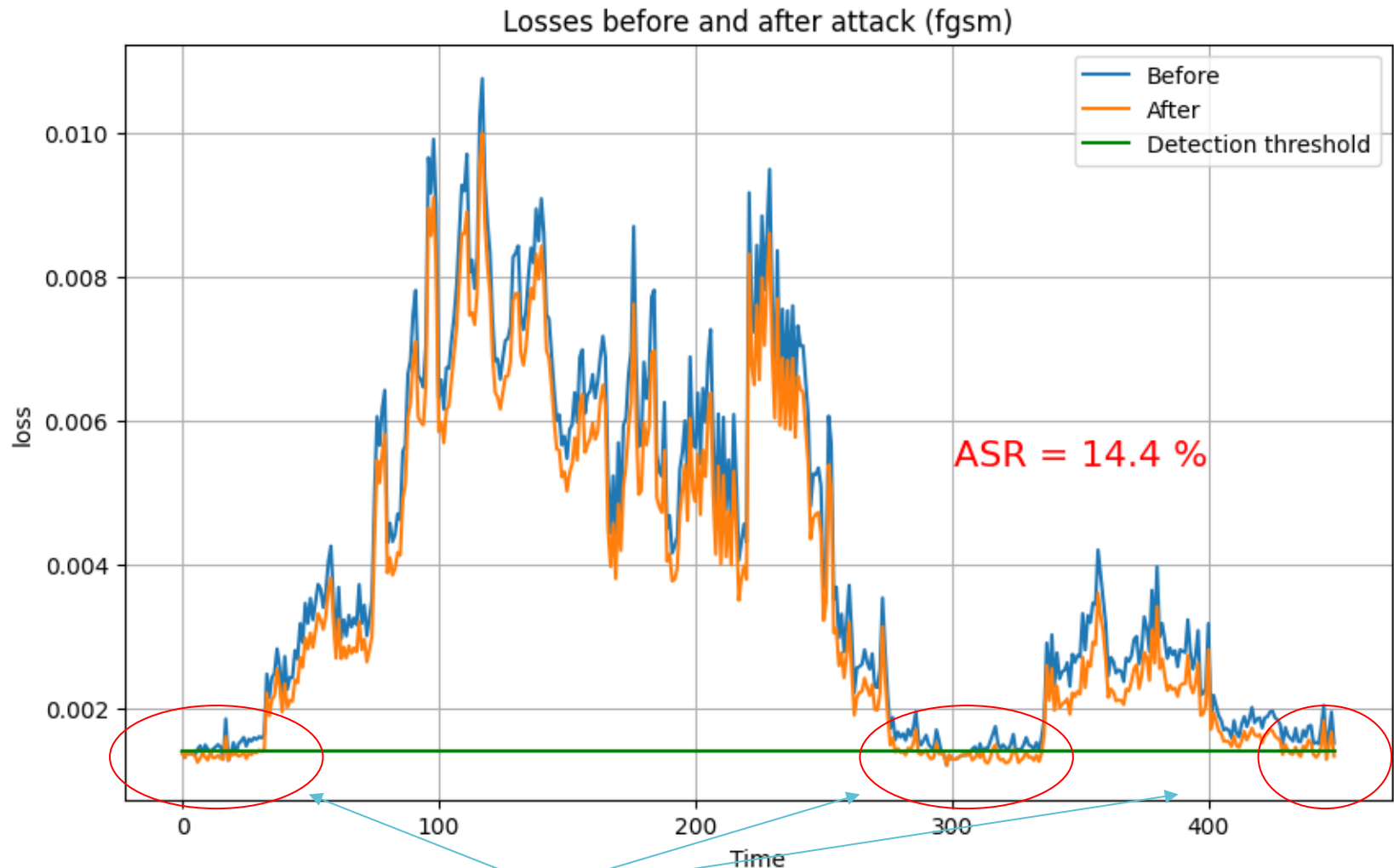
- FGSM attack for non supervised ML

The perturbations added to an input (level and rainfall) are very light



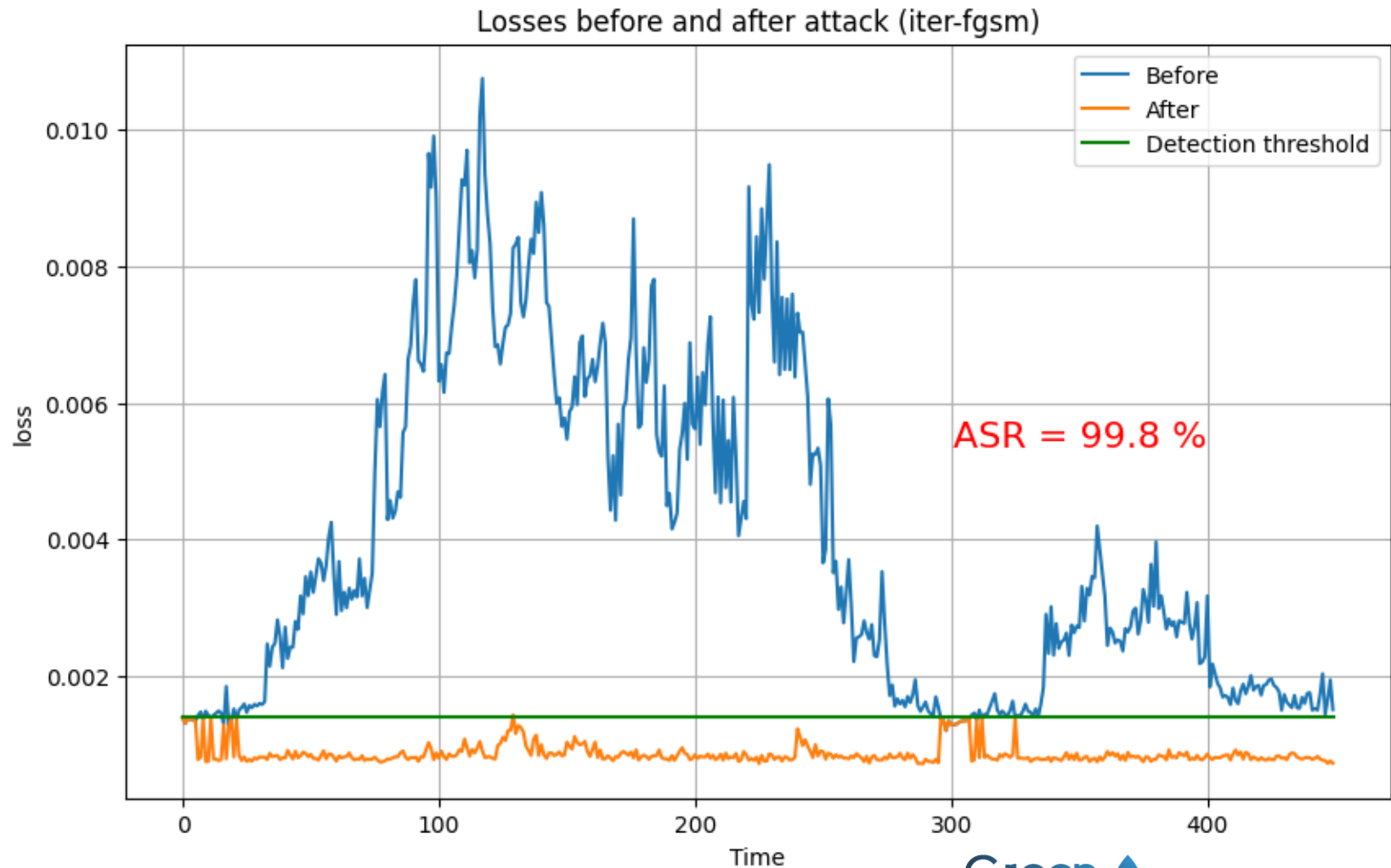
# Adversarial evasion attacks

## FGSM attack for non supervised ML



# Adversarial evasion attacks

## Iter-FGSM attack

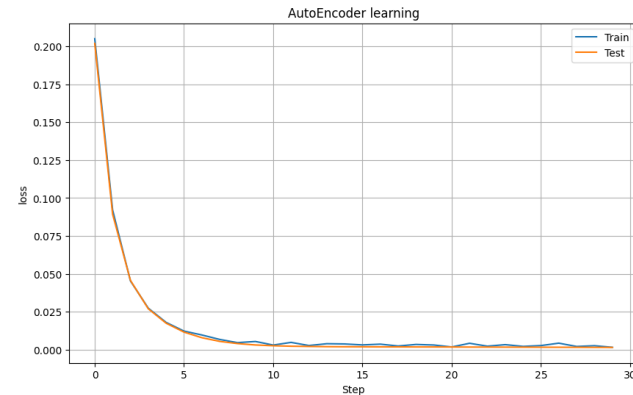


# Implemented robustification techniques

## Adversarial training

➤ Generate many adversarial examples before adding them to the training dataset

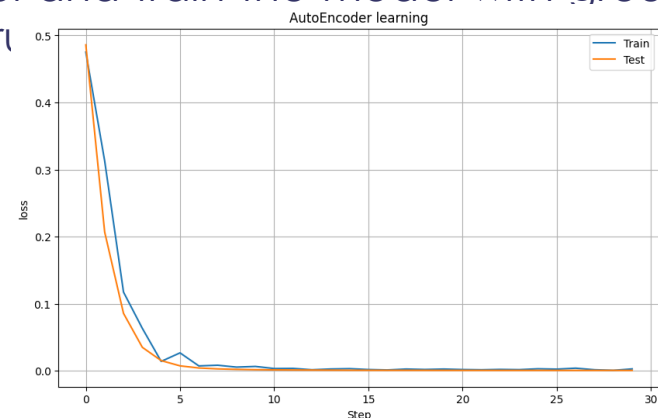
- TPR: 95%
- FPR: 2%
- **ASR(fgsm): 18,7%**



## Noising/denoising

➤ Add some noise at the input of the model and train the model with ground truth labels (signals without noise to be reconstructed)

- TPR: 99%
- FPR: 10%
- **ASR (fgsm): 14,2%**



# Thanks