PORTFOLIO

BY YASMIN BASHIR SHEIKH-MOHAMED

Data Scientist specialising in domain-knowledge–based ML modelling, with a complementary strength in human-centred integration.

I'm a data scientist who specialises in domain-knowledge–informed machine learning. I enjoy working at the intersection of domain expertise and technical modelling—turning expert insights into clear data preprocessing, thoughtful feature engineering, and model choices that reflect how systems actually behave.

The technical work is always my foundation, but I also care about presenting results in a way that supports the people who rely on them. My goal is to deliver solutions that are both technically solid and practically useful, communicated with clarity and respect for the domain they serve.

TABLE OF CONTENTS

NB: Some diagrams and examples in this section are simplified or generalized to illustrate the methods. They are based on my project work, but adapted for clarity, confidentiality, and cross-industry transferability.

GRAPH NEURAL NETWORKS FOR OPERATIONAL PLANNING

Organisation: SINTEF Energy Research

Role:
Researcher (ML & modelling)

Methods:
Graph Convolutional Networks, supervised learning, cost prediction, simulation-based datasets
Outcome: Developed a fast proxy model to predict operational cost under contingencies, reducing dependence on time-domain simulations.

Context & Problem:
Operational planning in power systems requires evaluating preventive strategies under many possible contingencies. Full time-domain simulations capture protection behaviour and dynamic response, but they are computationally expensive and limit how many strategies can be explored. The research question was whether ML could replace part of this simulation pipeline while preserving physically meaningful behaviour.

Author Accepted Manuscript version of the paper by Yasmin Bashir Sheikh-Mohamed et al.
In 2023 IEEE Belgrade PowerTech (2023), DOI: http://dx.doi.org/10.1109/PowerTech55446.2023.10202799

Graph Convolutional Networks for probabilistic power system operational planning

Yasmin Bashir Sheikh-Mohamed
Sigurd Hofsmo Jakobsen
Espen Flo Bødal
Fredrik Marinius Haugseth
Erlend Sandø Kiel
SINTEF Energy Research
Trondheim, Norway

Signe Riemer-Sørensen
SINTEF Digital
Oslo, Norway

***Abstract*—Probabilistic operational planning of power systems usually requires computationally intensive and time consuming simulations. The method presented in this paper provides a time efficient alternative to predict the socio-economic cost of system operational strategies using graph convolutional networks. It is intended for fast screening of operational strategies for the purpose of operational planning. It can also be used as a proxy for operational planning that can be used in long term development studies. The performance of the model is demonstrated on a network inspired by the Nordic power system.**

***Index Terms*—probabilistic operational planning, power system reliability, contingency analysis, machine learning, graph neural networks**

I. Introduction

A. Motivation

Most power systems today are operated and planned according to deterministic criteria, which are often socio-economical sub-optimal. In order to keep adequate reliability of supply while also minimizing the socio-economic costs, moving towards probabilistic criteria is recommended by the European FP7 project, GARPUR [1]. ACER, the European Union's Agency for the Cooperation of Energy Regulators, adopted a decision for transmission system operators (TSOs) to develop a methodology on probabilistic risk assessment [2]. The move towards probabilistic criteria requires efficient and accurate methods for decision support. This study explores the possibility of using machine learning for modeling decision support according to probabilistic criteria in system operational planning. The idea is that the model can be used as a screening method for operational planning, or for including a more accurate and fast representation of operations in long-term planning studies. There is a multitude of considerations when implementing probabilistic operational planning. The method should not only consider the risk of violating physical constraints but also the cost of operation [3]. Methods that attempt to minimize the cost of power system operation are therefore considered in this work.

A dynamic programming model was developed in [4] to quantify the socio-economic cost and detect the most favorable socio-economic operational strategy. Dynamic time-domain simulations were used to capture situations that normally go undetected by traditional static methods. However, this approach is computationally intensive. It is also necessary to predict consequences for a large sample space of possible contingencies, available corrective actions and other uncertainties, as steps toward identifying the optimal operational strategy.

As a response to these challenges, this paper proposes a two-step supervised learning model based on graph convolutional networks (GCNs) to rapidly predict the expected costs of simulated operational strategies, exemplified using data from [4].

B. Related works

Alternative methods to [4] for probabilistic operational planning have been presented in the literature [5]–[9]. In [5] a DC power flow was used to include power system response in a probabilistic operational planning model. Different cost-based criteria are compared in [6], where a transport model was used to model the power system response. An AC power flow and a linear approximation of frequency response were used in [7] to include frequency response in an operational planning model. More recent approaches use machine learning for generating proxy models of real-time operation to speed up probabilistic operational planning [8], [9]. In these papers, a machine learning model is trained to act as a DC-security constrained optimal power flow (SCOPF) and to predict the optimal corrective actions given a set of preventive actions. This is a promising approach, however, the use of a DC-SCOPF means that voltage, frequency and stability issues will not be captured. Moreover, the time domain characteristic of protection systems cannot be included. In the proposed approach, a GCN is trained to predict the result of a detailed time-domain simulation that calculates the cost of operating a power system given a set of preventive and corrective actions while considering the time-domain characteristics of protection systems [4].

Graph neural networks (GNNs) is a collective term describing neural networks that process data structured as graphs. Use

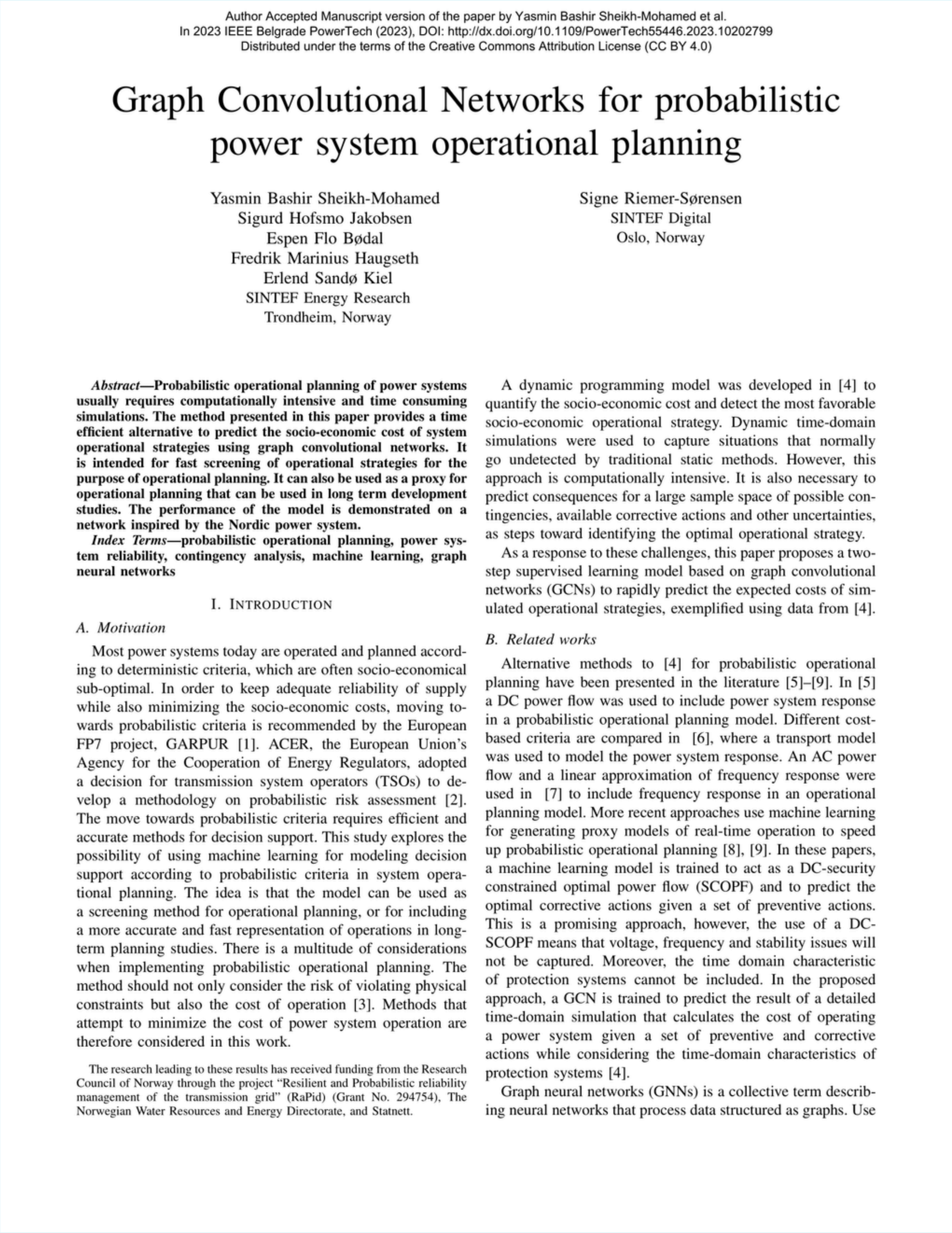
The research leading to these results has received funding from the Research Council of Norway through the project "Resilient and Probabilistic reliability management of the transmission grid" (RaPid) (Grant No. 294754), The Norwegian Water Resources and Energy Directorate, and Statnett.

Source: Sheikh-Mohamed et al., 2023 (CC BY 4.0)

GRAPH NEURAL NETWORKS FOR OPERATIONAL PLANNING

Approach:
I developed a two-stage GCN framework trained on simulation output: a classification model to filter high-cost strategies and a regression model to predict cost for low-cost cases. GCNs were chosen because transmission systems form a non-Euclidean graph where interactions follow network topology rather than spatial coordinates; learning across edges preserves meaningful dependence patterns between areas, flows, and protection states.

Outcome:
The model enabled fast screening of preventive actions, providing near-optimal cost estimates while significantly reducing the number of full simulations required. This supports system operators and researchers in evaluating operational strategies at scale.



Author Accepted Manuscript version of the paper by Yasmin Bashir Sheikh-Mohamed et al.
In 2023 IEEE Belgrade PowerTech (2023), DOI: http://dx.doi.org/10.1109/PowerTech55446.2023.10202799

Graph Convolutional Networks for probabilistic power system operational planning

Yasmin Bashir Sheikh-Mohamed
Sigurd Hofsmo Jakobsen
Espen Flo Bødal
Fredrik Marinius Haugseth
Erlend Sandø Kiel
SINTEF Energy Research
Trondheim, Norway

Signe Riemer-Sørensen
SINTEF Digital
Oslo, Norway

***Abstract*—Probabilistic operational planning of power systems usually requires computationally intensive and time consuming simulations. The method presented in this paper provides a time efficient alternative to predict the socio-economic cost of system operational strategies using graph convolutional networks. It is intended for fast screening of operational strategies for the purpose of operational planning. It can also be used as a proxy for operational planning that can be used in long term development studies. The performance of the model is demonstrated on a network inspired by the Nordic power system.**

***Index Terms*—probabilistic operational planning, power system reliability, contingency analysis, machine learning, graph neural networks**

I. Introduction

A. Motivation

Most power systems today are operated and planned according to deterministic criteria, which are often socio-economical sub-optimal. In order to keep adequate reliability of supply while also minimizing the socio-economic costs, moving towards probabilistic criteria is recommended by the European FP7 project, GARPUR [1]. ACER, the European Union's Agency for the Cooperation of Energy Regulators, adopted a decision for transmission system operators (TSOs) to develop a methodology on probabilistic risk assessment [2]. The move towards probabilistic criteria requires efficient and accurate methods for decision support. This study explores the possibility of using machine learning for modeling decision support according to probabilistic criteria in system operational planning. The idea is that the model can be used as a screening method for operational planning, or for including a more accurate and fast representation of operations in long-term planning studies. There is a multitude of considerations when implementing probabilistic operational planning. The method should not only consider the risk of violating physical constraints but also the cost of operation [3]. Methods that attempt to minimize the cost of power system operation are therefore considered in this work.

The research leading to these results has received funding from the Research Council of Norway through the project "Resilient and Probabilistic reliability management of the transmission grid" (RaPid) (Grant No. 294754), The Norwegian Water Resources and Energy Directorate, and Statnett.

A dynamic programming model was developed in [4] to quantify the socio-economic cost and detect the most favorable socio-economic operational strategy. Dynamic time-domain simulations were used to capture situations that normally go undetected by traditional static methods. However, this approach is computationally intensive. It is also necessary to predict consequences for a large sample space of possible contingencies, available corrective actions and other uncertainties, as steps toward identifying the optimal operational strategy.

As a response to these challenges, this paper proposes a two-step supervised learning model based on graph convolutional networks (GCNs) to rapidly predict the expected costs of simulated operational strategies, exemplified using data from [4].

B. Related works

Alternative methods to [4] for probabilistic operational planning have been presented in the literature [5]–[9]. In [5] a DC power flow was used to include power system response in a probabilistic operational planning model. Different cost-based criteria are compared in [6], where a transport model was used to model the power system response. An AC power flow and a linear approximation of frequency response were used in [7] to include frequency response in an operational planning model. More recent approaches use machine learning for generating proxy models of real-time operation to speed up probabilistic operational planning [8], [9]. In these papers, a machine learning model is trained to act as a DC-security constrained optimal power flow (SCOPF) and to predict the optimal corrective actions given a set of preventive actions. This is a promising approach, however, the use of a DC-SCOPF means that voltage, frequency and stability issues will not be captured. Moreover, the time domain characteristic of protection systems cannot be included. In the proposed approach, a GCN is trained to predict the result of a detailed time-domain simulation that calculates the cost of operating a power system given a set of preventive and corrective actions while considering the time-domain characteristics of protection systems [4].

Graph neural networks (GNNs) is a collective term describing neural networks that process data structured as graphs. Use

Source: Sheikh-Mohamed et al., 2023 (CC BY 4.0)

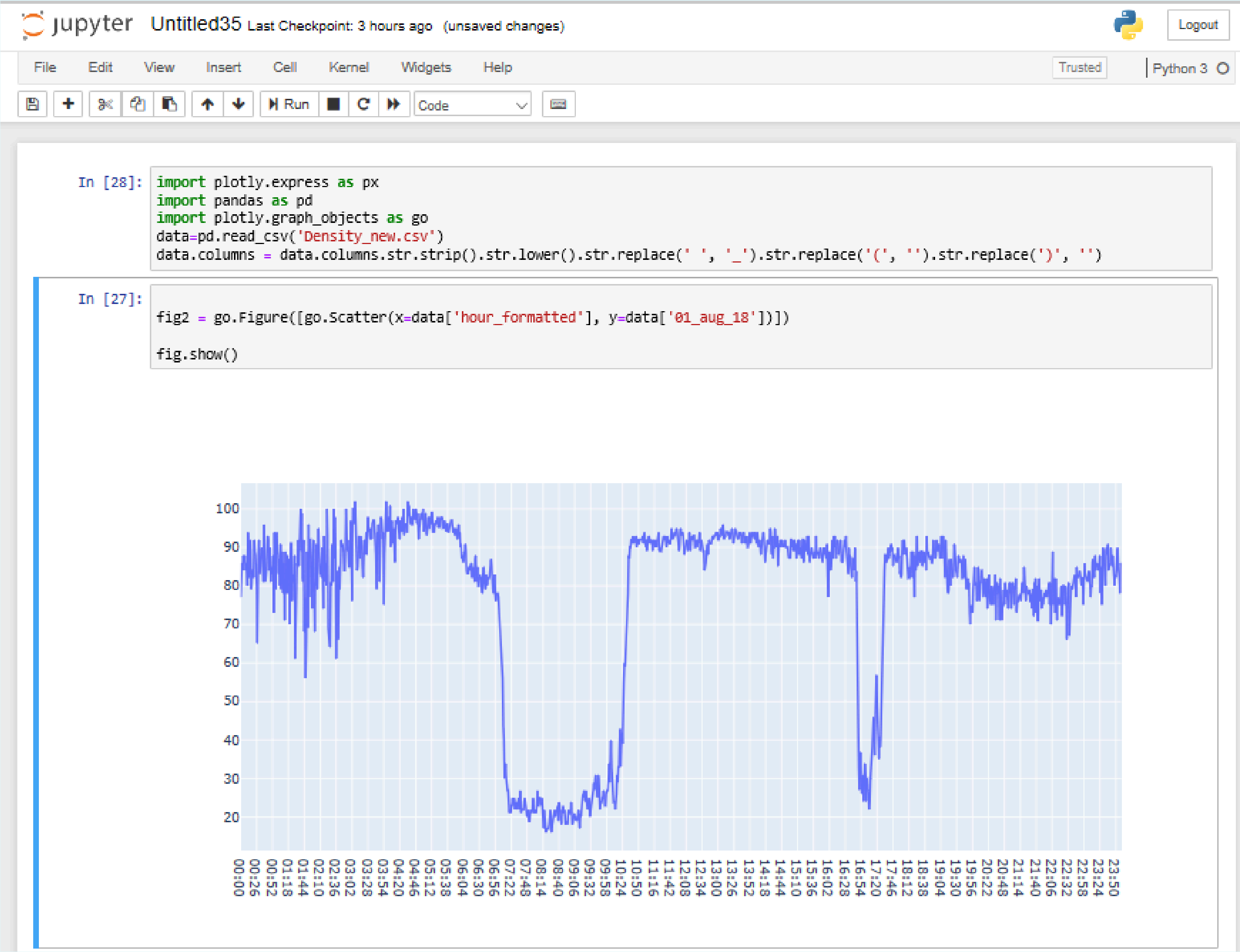
PREDICTIVE MAINTENANCE – TRANSFORMER COOLING SYSTEM MONITORING

Organisation: SINTEF Energy Research

Role:
ML Engineer / Research Assistant

Methods:
LSTMs, expert-informed feature engineering, anomaly detection, sensor-state filtering
Outcome: Identified domain-specific operational states and built preprocessing logic enabling meaningful condition monitoring.

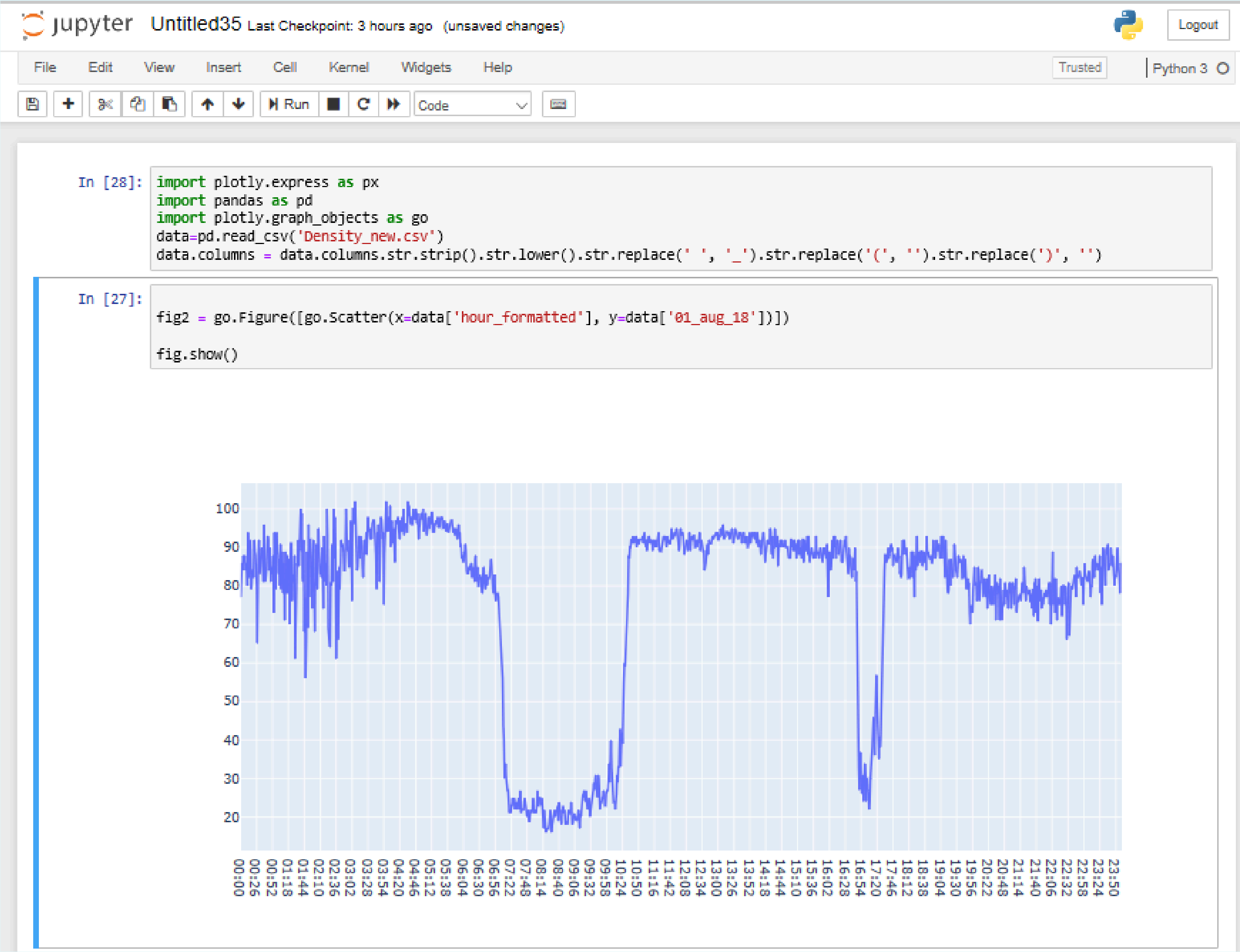
Context & Problem:
Transformer cooling behaviour reflects both thermal load and mechanical conditions, but sensor data includes periods where the cooling system is off or the transformer is not in operational use. Without filtering these states, anomaly models learn spurious patterns, resulting in false alerts and poor interpretability. The goal was not only to model behaviour, but to ensure the model was useful to engineers.



PREDICTIVE MAINTENANCE – TRANSFORMER COOLING SYSTEM MONITORING

Approach:
I worked with domain experts to map system states (e.g., load thresholds, thermal activation behaviour, sensor placement differences) and used these to construct rule-based filters before training. LSTMs were chosen due to thermal lag: temperature changes occur with delayed response to electrical load, which requires models capable of learning long-term dependencies rather than only point-in-time dynamics.

Outcome:
The project demonstrated that domain-aligned preprocessing and state filtering are prerequisites for meaningful anomaly detection. Instead of flagging every deviation, the model targeted anomalies only in periods where cooling should be active, making outputs actionable for operators.



ROOFTOP SOLAR INTEGRATION & PROSUMER ANALYSIS

Organisation: Fornybar Norge

Role:
Research Analyst

Methods:
Created a reusable Python script for whole analysis, Scenario modelling, PV production simulation analysis, market analysis, large datasets, result validation with expert interviews.

Context & Problem
Norway has increasing interest in rooftop solar despite historically low electricity prices and hydropower dominance. The challenge was to assess whether large-scale prosumer adoption would meaningfully affect grid flows, system costs, or market design—and to present findings in a way industry actors could use in policy discussions.

Source: Fornybar Norge (2023), "Hvordan få solkraft fra Norges hustak inn i kraftsystemet"

ROOFTOP SOLAR INTEGRATION & PROSUMER ANALYSIS

Approach:
I worked with both grid operators and solar suppliers to validate assumptions about production profiles, household demand, seasonal variation, and grid integration challenges. Rather than treating the dataset as neutral, I supplemented it with domain knowledge (e.g., seasonal hydro patterns, regional grid constraints) and removed periods where solar has no systemic relevance (e.g., winter months with negligible production).

Outcome:
Provided evidence-based insights on how distributed solar could impact the Norwegian grid and market structure.

The analysis showed that the main system effects come not from average production but from localized congestion and peak-time behaviour, shifting the conversation from volume-based arguments to infrastructure planning and regulatory design.

Source: Fornybar Norge (2023), "Hvordan få solkraft fra Norges hustak inn i kraftsystemet"

DOMAIN-ADAPTIVE VISUALISATION & STAKEHOLDER COMMUNICATION

Good visualization is a shared language. When results are communicated in formats people intuitively understand, it strengthens collaboration, speeds up validation, and turns analysis into real-world action.

EXAMPLE I - VISUALS THAT SUPPORT COST-AWARE OPERATIONAL DECISIONS

- A false high just means buying extra backup early.
- A false low means paying huge real-time costs.
- That’s why operators need the confusion matrix.

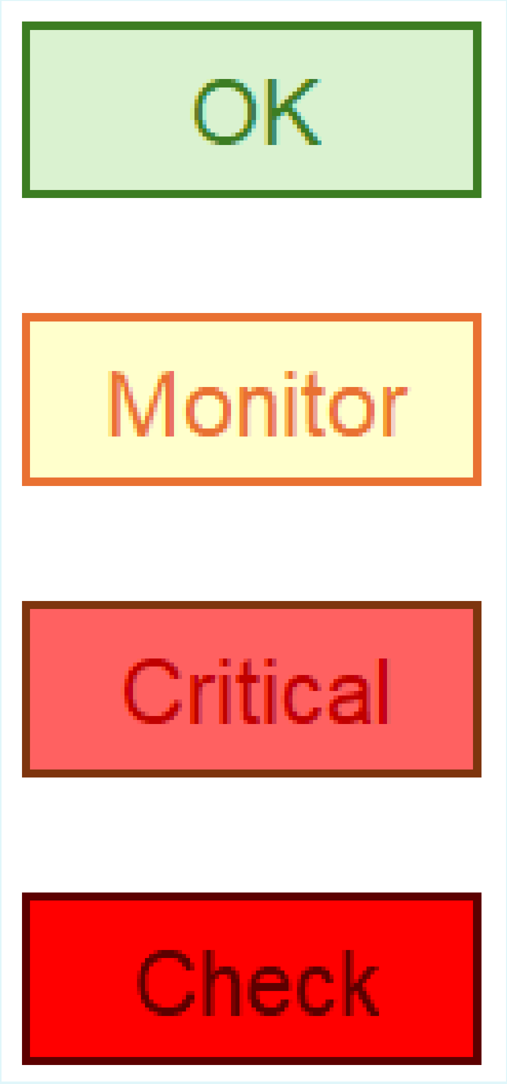
Accuracy=0.981
Precision=0.984
Recall=0.966

True low cost 519 99%	False low cost 5 1%
True high cost 11 3%	False high cost 314 97%

From **information** to **comprehensive knowledge**

EXAMPLE II - 2 AM ALARM CALL

- Imagine a transformer shows an anomaly score of 0.74.
- That number means nothing to an operator at 02:00 AM during an alarm call.
- The model becomes valuable only when it speaks the language of real operations.



From **information** to **comprehensive knowledge**

EXAMPLE III - IT'S NOT ABOUT THE FUEL. IT'S ABOUT THE TRAFFIC

I realised one of the common misconception that was difficult for power grid owners to convey was why at-home-installations could cause challenges for the grid although people were consuming more energy than they were producing:

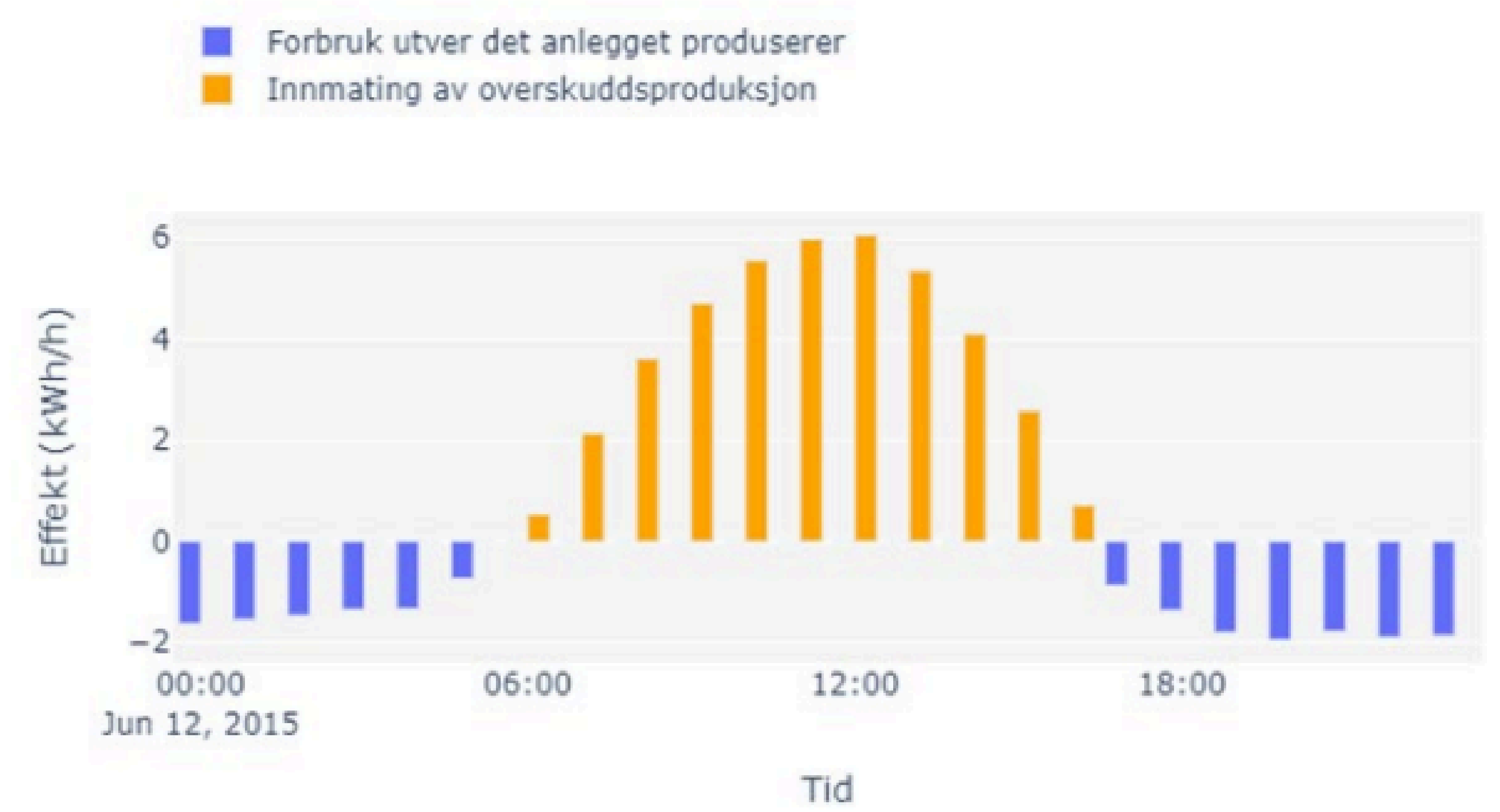
In other words:

- Net Production Per Year ≠ Net Production Per Minute

This plot shows net flow for one household.

- Above zero = production surplus exported to the grid.
- Below zero = consumption higher than production, so power is imported.

Solar homes flip between the two many times a day.



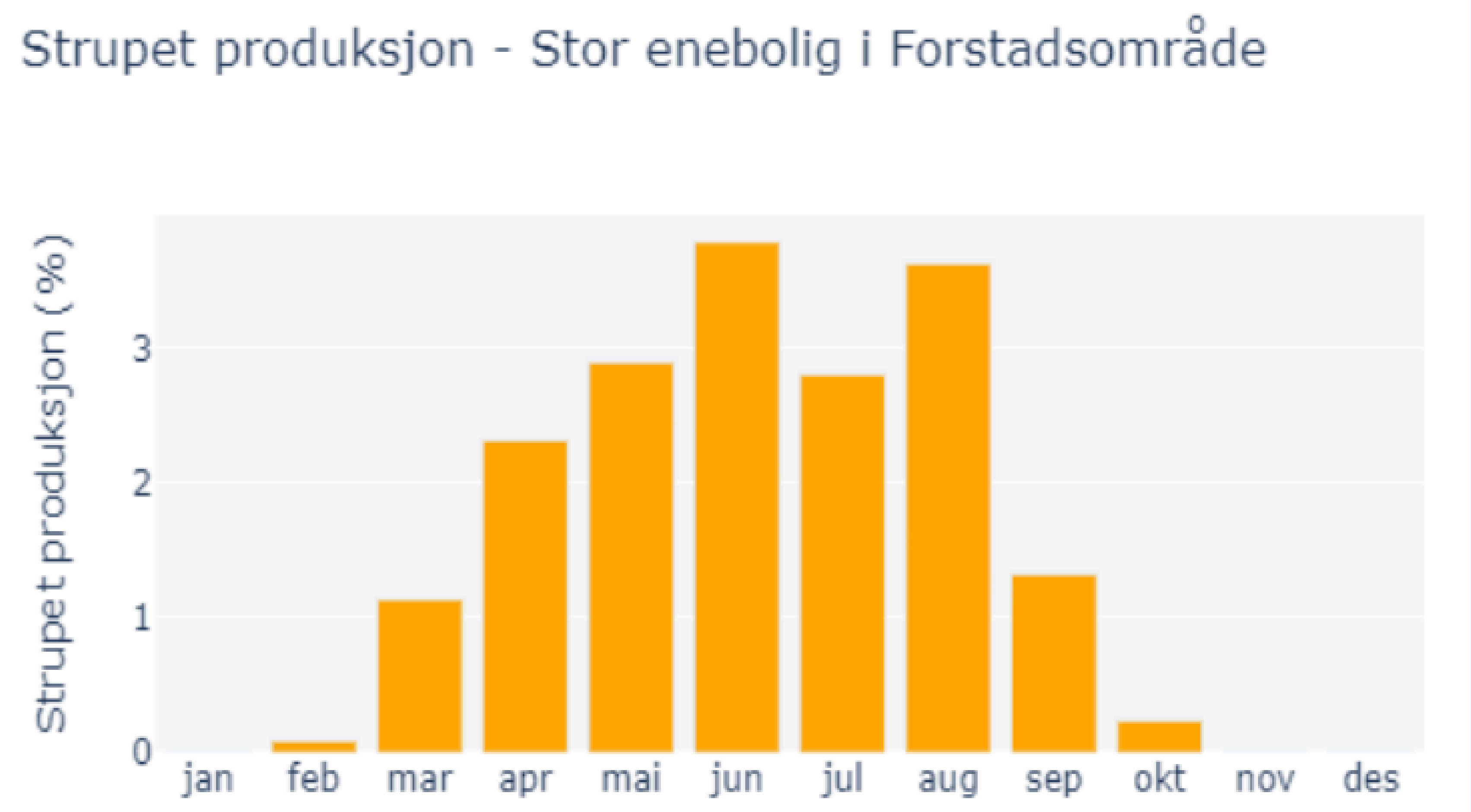
Source: Fornybar Norge (2023), "Hvordan få solkraft fra Norges hustak inn i kraftsystemet"

EXAMPLE IV- HELPING GRID OWNERS INTERPRET CURTAILMENT AND REGULATORY LIMITS DIFFERENTLY

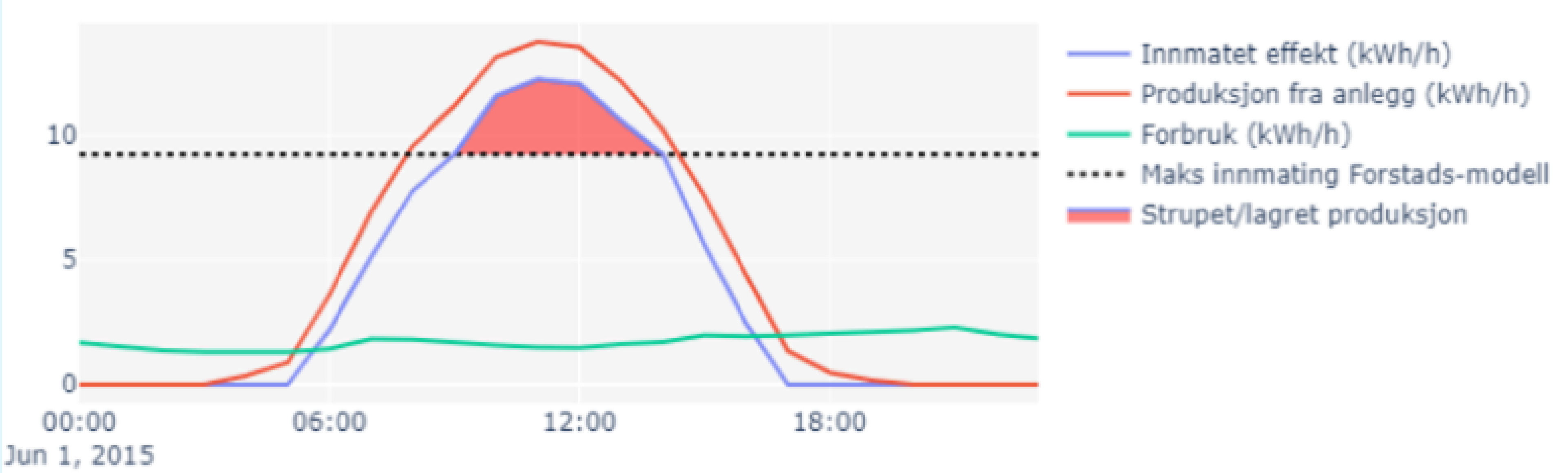
The concept of curtailment is uncommon in Norway, therefore when presenting it, intuitive visuals were crucial.

The bottom graph shows production from a household solar PV-system on the highest production day of the year. I found out that the industry is well aware that peak production days are few and far between, so this visual helped communicate how little energy was actually going to ‘waste’.

Curtailment doesn’t mean smaller solar systems.
It’s like cars:
You don’t ban cars because roads get crowded a few days a year.
You just can’t drive during those rare peak hours — less than 3% of the time.



Source: Fornybar Norge (2023), “Hvordan få solkraft fra Norges hustak inn i kraftsystemet”



Source: Fornybar Norge (2023), “Hvordan få solkraft fra Norges hustak inn i kraftsystemet”

DOMAIN-INFORMED DATA QUALITY & FEATURE ENGINEERING

Using domain logic — expert heuristics, physical constraints, operational rules — to clean, filter, and structure data before training any ML model.

EXAMPLE I – CLEANING TIME-SERIES FOR TRANSFORMER COOLING SYSTEM MONITORING

Step 1 — Apply ML Preprocessing Methods

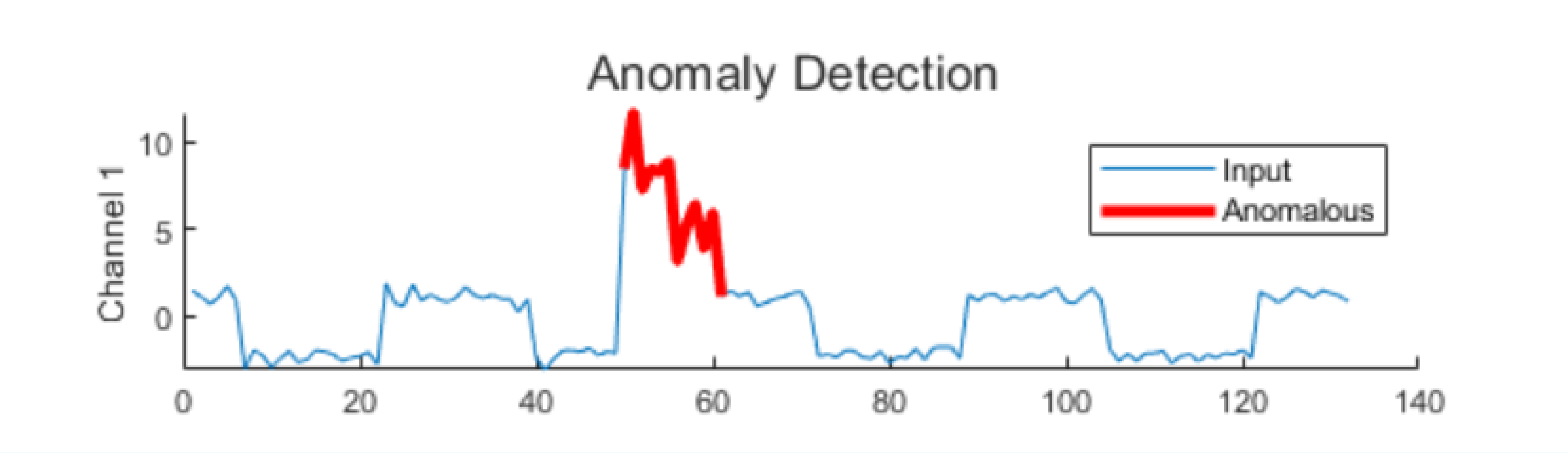
- Initial cleaning
- Basic filtering
- Preliminary anomaly detection
- Baseline feature extraction

Step 2 — Share Results with Domain Experts (Visual + Analytical Review)

- Use visualised data (heatmaps, correlation plots, sensor-state timelines)
- Identify explainable anomalies and domain-driven patterns

Examples of expert-driven insights:

- Sensor location vs. thermal zones explains correlations/mismatches
- Identifying irrelevant system states (remove from dataset). Pseudocode to the right exemplifies this.




```python
def classify_operating_state(row):
    """
    Classify transformer and cooling system state for a single time step.

    row: an object with at least:
        - row["load"]
        - row["temp_sensor_16"]
    """
    load = row["load"]
    t16  = row["temp_sensor_16"]

    # Default assumptions
    transformer_in_use = True
    cooling_system_active = True

    # Rule 1: transformer not in operational use
    if load <= LOAD_OFF_THRESHOLD:
        transformer_in_use = False

    # Rule 2: cooling system inactive
    if t16 <= TEMP_SENSOR16_INACTIVE_THRESHOLD:
        cooling_system_active = False

    return {
        "transformer_in_use": transformer_in_use,
        "cooling_system_active": cooling_system_active,
    }
```

Step 3 — Translate Expert Insights into Data Rules

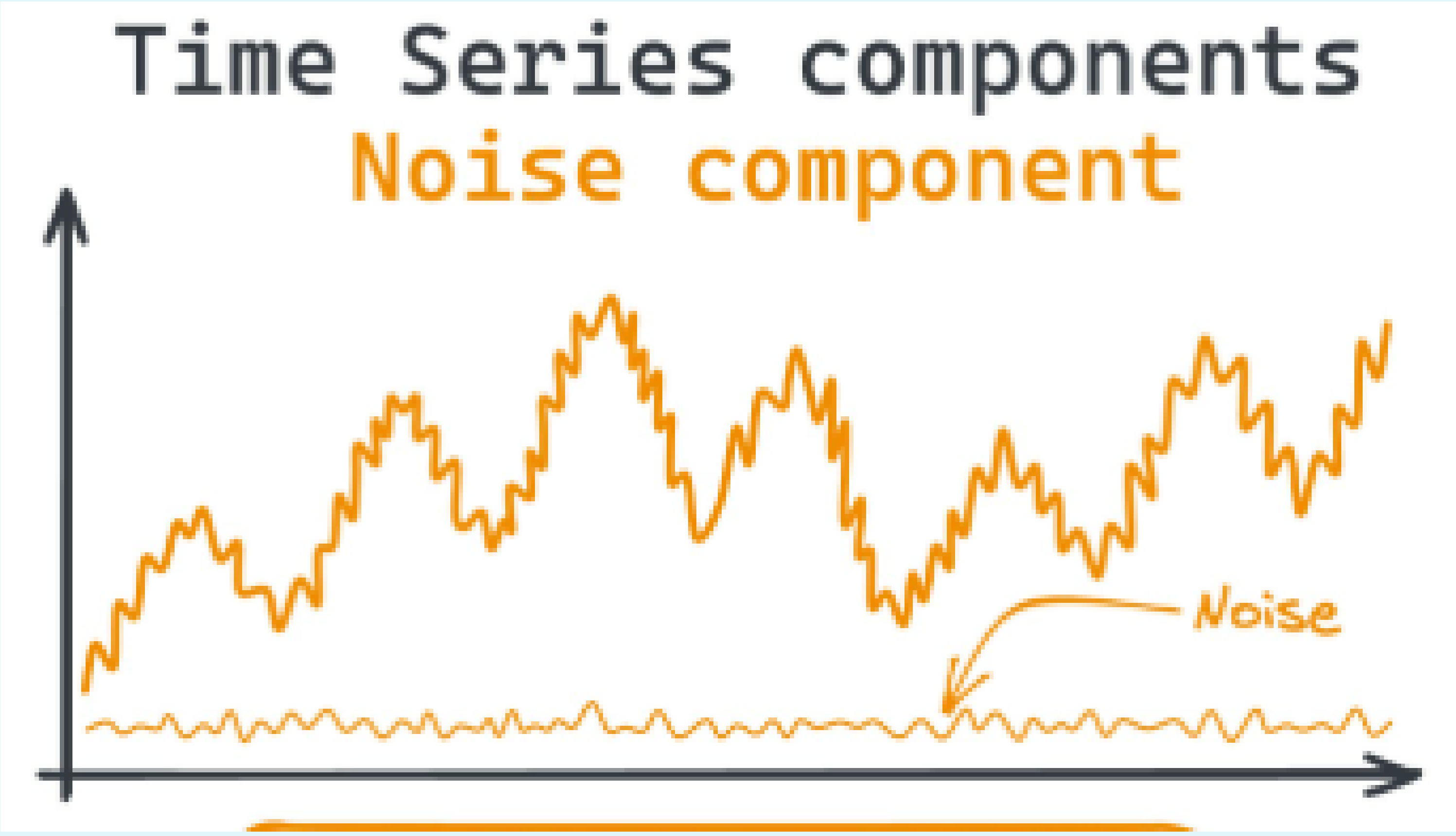
- Turn domain explanations into filtering constraints
- Add conditional logic to remove misleading samples. See example in pseudo code.

Step 4 — Update Data Processing Pipeline

- Add new cleaning constraints
- Update feature engineering logic
- Re-define imputation decisions based on domain knowledge

Step 5 — Re-run Preprocessing + Return to Experts

- New visualisations
- New anomalies to explain
- Repeat cycle until modelling dataset converges




```python
def preprocessing_pipeline(data):
    # Step 1: Compute derived state
    data["transformer_off"] = data["load"] <= LOAD_THRESHOLD
    data["cooling_active"] = data["cooling_state"] == "on"

    # Step 2: Filter out irrelevant periods
    mask = (~data["transformer_off"]) & (data["cooling_active"])
    filtered = data[mask]

    return filtered
```

CROSS-INDUSTRY EXAMPLES

Domain-informed preprocessing ensures the model learns meaningful behaviour by using domain constraints, system states, physical limits, and expert insight to shape filtering, feature engineering, and data selection.

💳 Finance & Fraud Detection

- Apply seasonal/holiday segmentation before training
- Normalize behavior per-user rather than global
- Treat travel mode as separate behavioural baseline
- Use regulatory thresholds to create cost-sensitive labels

🩺 Healthcare & Biometric Signals

- Segment data by physiological state (rest, sleep, exertion)
- Normalize vitals per patient baseline instead of global scaling
- Remove samples when sensor contact is lost
- Engineer features from known circadian or pharmacological cycles

📦 Retail / Demand Forecasting

- Remove periods of stockouts to avoid misinterpreting demand
- Condition demand features on promotions, holidays, weather
- Engineer cross-category demand elasticity features
- Use physical supply-chain constraints (lead times, batch deliveries)

📡 Telecom & Network Operations

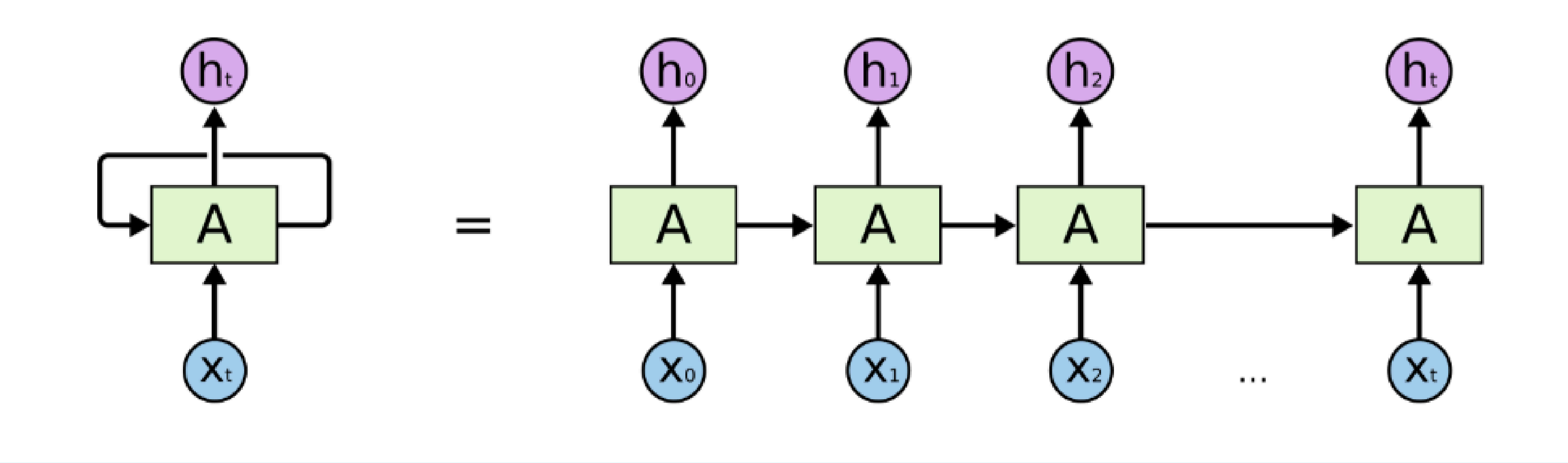
- Remove planned maintenance outages from fault datasets
- Engineer load ratios: traffic vs capacity
- Use time-of-day + routing topology as contextual features
- Filter synthetic anomalies from failover testing

DOMAIN-AWARE ML MODELLING & MODEL SELECTION

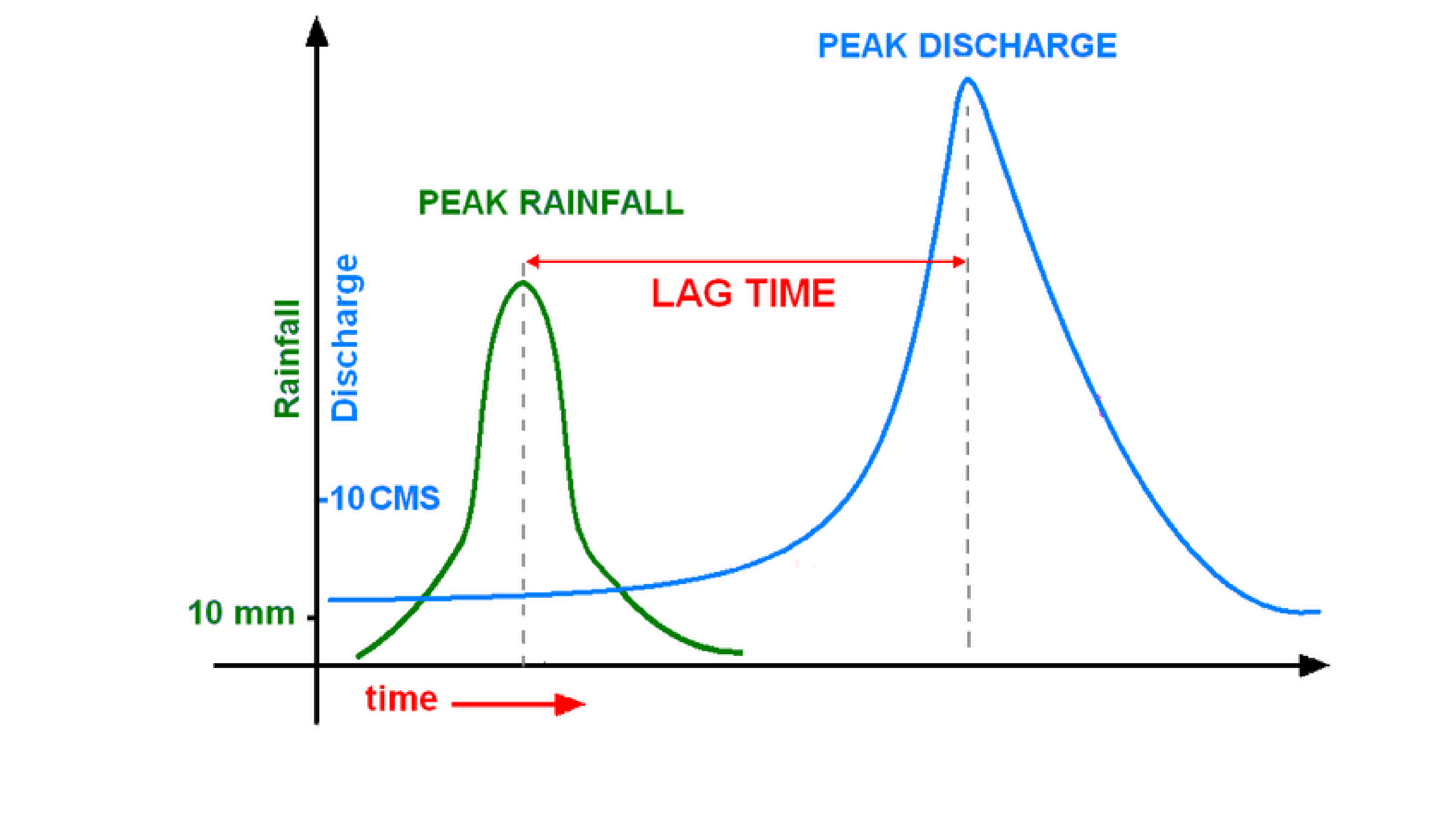
Choosing the right model is not purely about the data. It’s about the real system and processes the model is supposed to simulate. It’s about the intended use.

EXAMPLE I - LSTMS FOR SYSTEMS WITH LAG TIME (E.G. THERMAL LAG)

LSTMs are a class of Recurrent Neural Networks (RNNs) designed to model sequential data using feedback loops that retain information from earlier time steps. Unlike feed-forward models, which treat each sample independently, LSTMs keep a memory of past values, making them suitable for systems where the current state depends on historical behaviour.



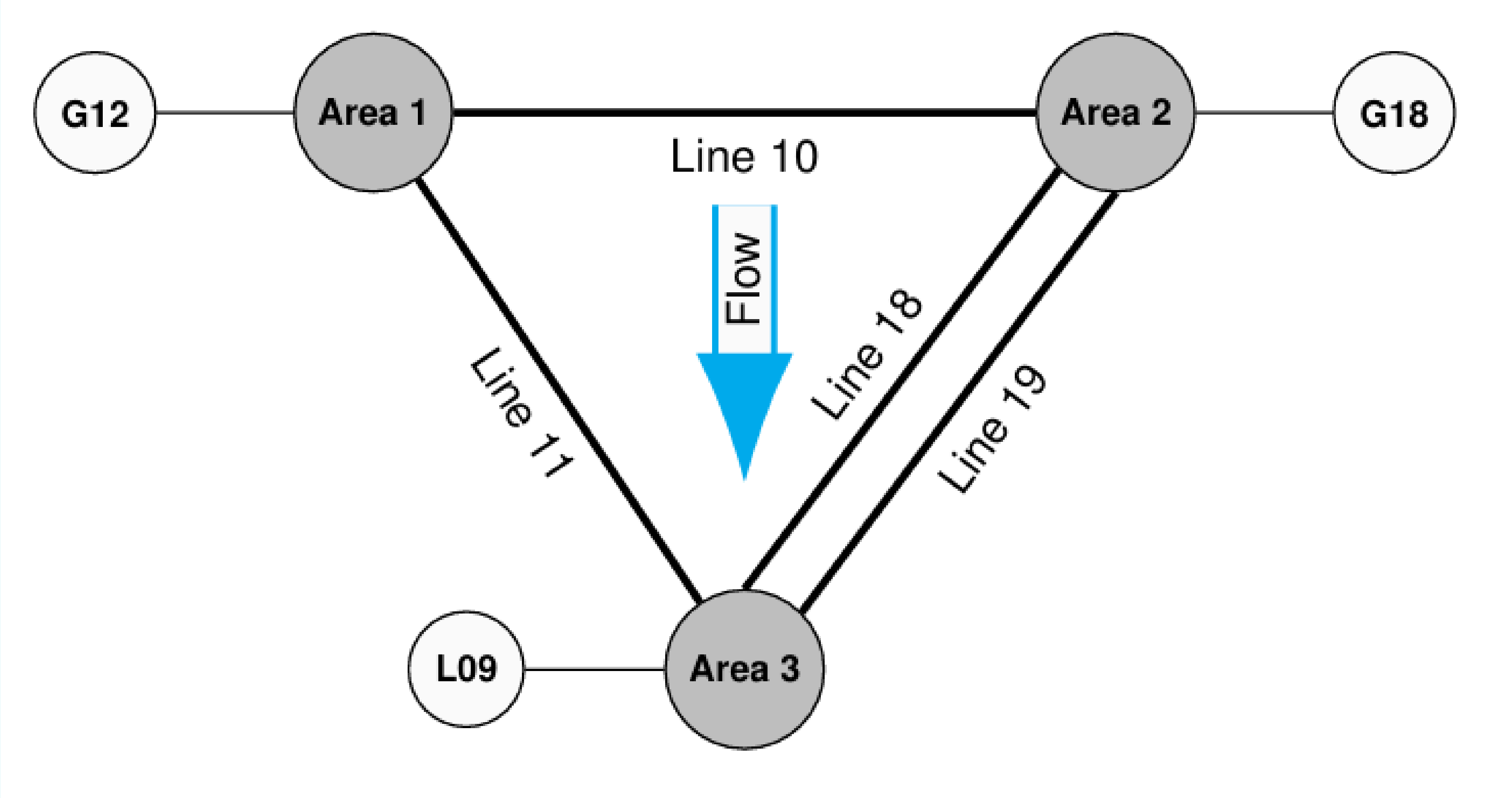
RNNs in general are a strong fit for time-series prediction, but standard RNNs struggle with vanishing gradients, limiting their ability to capture long-term dependencies. LSTMs address this through gating mechanisms (input, forget, output gates) that control how information is stored or discarded over time. This allows them to learn slow, accumulated effects, rather than only short-term fluctuations.



In systems with lag—such as thermal dynamics, where temperature responds gradually to load changes—this memory capability is crucial. LSTMs can model delayed responses caused by heat capacity, inertia, and cooling processes more effectively than models that assume immediate cause–effect relationships.

EXAMPLE II - GCNS FOR SYSTEMS WITH GRAPH TOPOLOGY

Although several model architectures were tested, Graph Convolutional Networks (GCNs) were ultimately the best fit because their mathematical foundations align with systems that have a non-Euclidean structure. Unlike classical neural networks, which assume data lies in a regular grid (e.g., images or sequences), GCNs operate directly on graphs, where relationships between components are represented as edges rather than fixed positions.



Source: Sheikh-Mohamed et al., 2023 (CC BY 4.0)

This makes GCNs particularly suited for power systems, where network topology, line connections, and flow constraints define behaviour.

In a graph-structured system, nodes represent entities and edges represent relationships between them. What makes this fundamentally different from Euclidean data is that these relationships are irregular—each node can connect to any number of other nodes, and those connections carry meaning.

GCNs learn not only from node attributes, but from how nodes are connected (the relationship between them is represented by edges). The mathemathical formula is given below.

$$G = \{\xi, V, E, \eta\}$$
$$\text{where } \eta : E \rightarrow \{\{i,j\} : i,j \in V\}$$

CROSS-INDUSTRY EXAMPLES

A model needs to perform well in the real world, and not only on a dataset. Realising that requires aligning algorithms with domain mechanics.

📈 Systems Where Outcomes Must Optimize a Cost or Utility Function

Real-world examples:

- Robot path planning to minimize energy + collision risk
- Dynamic inventory ordering to minimize stockouts vs warehousing cost
- Adaptive traffic signal control to minimize delays
- Power dispatch to minimize cost while satisfying security constraints

Task framing:

- Not just prediction → action selection
- Rewards encode business/operational cost

Good model types:

- Reinforcement learning (RL)
- Dynamic programming
- Monte Carlo optimization
- Constrained optimization models

📸Systems With High-Dimensional Signals (e.g., Images, Audio)

Real-world examples:

- Crack detection in infrastructure images
- Audio analysis for machine noise / bearing faults
- Medical radiography (MRI, CT scans)
- Log-based event embeddings for cybersecurity

Task framing:

- Classification, regression, or embedding learning

Good model types:

- CNNs (vision, spatial structure)
- Transformers (long-range context)
- Contrastive self-supervised learning

THANK YOU FOR YOUR ATTENTION