LobithNN - Pioneering towards Accountable AI

Robustness - Reliability - Explainability

- Some context before we start
- What is LobthNN?
- Robustness: handling missing data
- Reliability: weather ensembles & model uncertainty combined
- Explainability: going beyond SHAP with causality
- Wrap-up & Questions

Hans Korving, Jurian Beunk & Jing Deng
Deltares

Deltares

Before We Start

- **Official work starts in Q3 2025**
  - But I will already share insights from related research and preparatory work

- **What I can already share with you**
  - A preview of the LobithNN model from 2024
  - Applications of the method in other projects regarding wildfires and waves
  - Preliminary findings from exercises with LobithNN

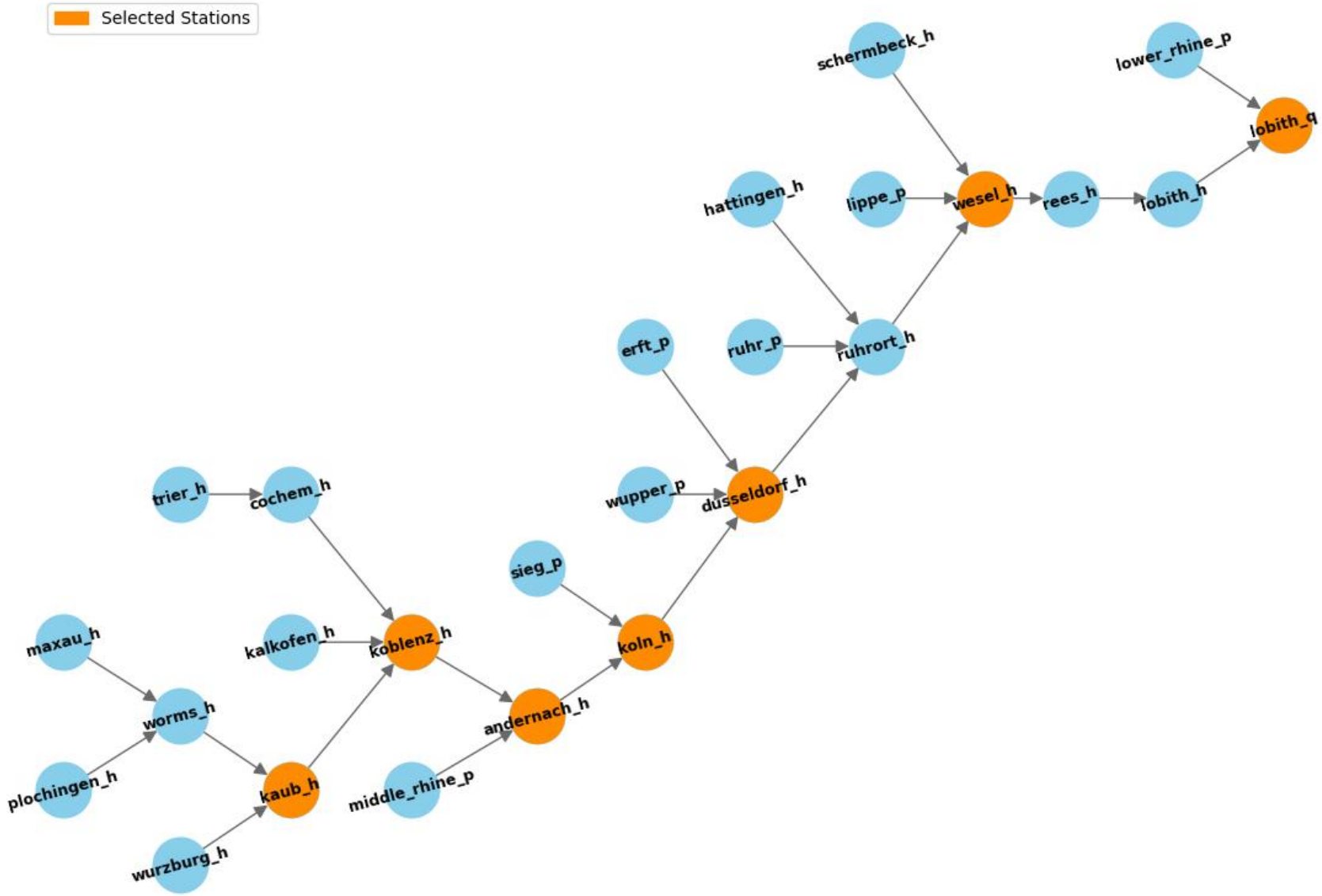
Accountable AI from ethical principles

Principle	Relevant	Supporting Methods	Our approach
Transparancy (aka Explainability)	✅	SHAP + Causal methods	Using **insights into model behavior and data relations**, for example via visualizations of feature contributions, making **causal relationships** visible and explaining **model choices in understandable terms**.
Accountability (aka Reliability)	✅	Conformal Prediction (CP)	Using **uncertainty explicitation** (for example, via Conformal Prediction), linking uncertainty to underlying model logic to help users **understand which decision patterns are more reliable or less**.
Fairness	✅	Fairness-aware AI	Understanding **who benefits or might be disadvantaged by model outcomes**, supporting fair trade-offs among stakeholders, sectors, and regions.
Privacy & Security	❌	Not directly applicable in this context	Water and meteorological data are **mostly publicly available**, and **our models don't use sensitive personal data**. Security is important, but usually not a primary ethical challenge here.
Beneficence	❌	No direct impact expected	Our models don't autonomously control water management but serve as **decision support**. The goal **is efficiency and sustainability**, not direct human impact such as in healthcare.

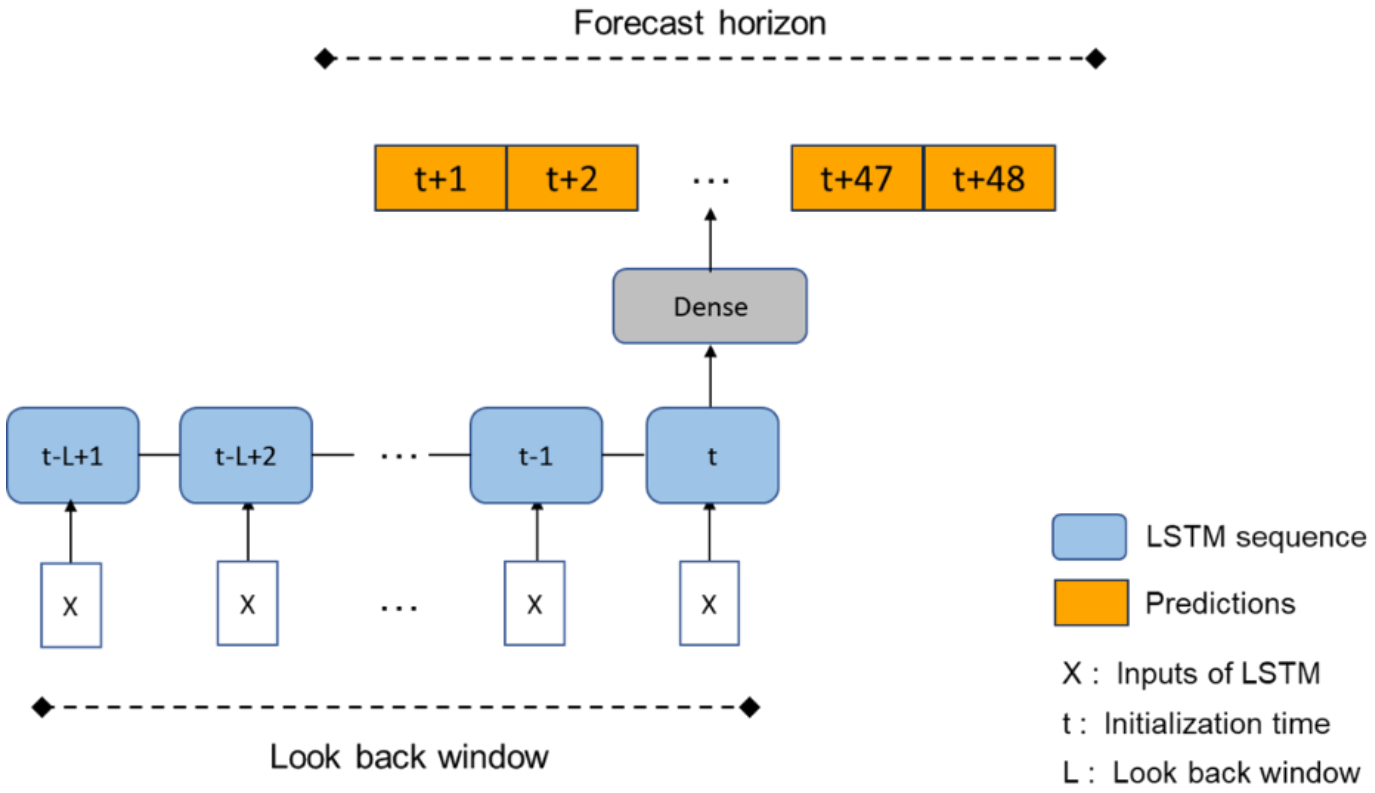
LobithNN: Deep learning model for short-term flow forecasts

- **Goal**
  - Improve the 48-hour discharge forecast at Lobith on an hourly basis
- **Input**
  - Historical discharge at Lobith
  - Water levels from 6 reliable upstream stations (Wesel, Düsseldorf, Köln, Andernach, Koblenz, Kaub)
- **Model**
  - Single-layer LSTM network, outputting 48 simultaneous forecasts
- **Data**
  - Trained on historical BfG data (1980-2021)
  - Implemented with operational Matroos data
- **Model tuning**
  - Variation in measurement stations and/or look-back windows (= time sequence length)
  - Use of time difference (ΔQ) instead of Q as target

Water Level Stations (_h) along the River Rhine



Architecture of the LobithNN LSTM-model

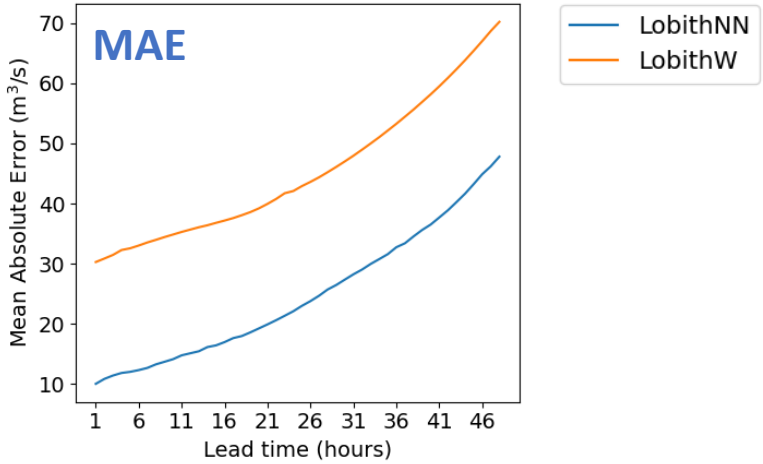


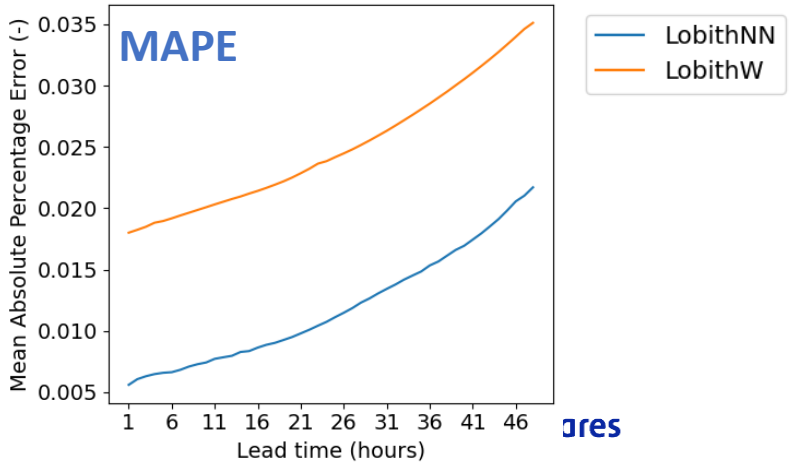
*The **LSTM model** uses an input sequence of **upstream water levels** over a 'look back window' to simultaneously **forecast 48 hours ahead**. The model processes time series and produces simultaneous forecasts through a **dense layer**.*

LobithNN Outperforms Current Model

- **Compared to LobithW** (= current model)
  - Lower MAE (= Mean Absolute Error) and MAPE (= Mean Absolute Percentage Error)
  - Across all lead times (2019-2021)
- **Split by discharge range** compared to LobithW
  - Consistently better at discharges <4000 m³/s
  - Comparable around 4000-6000 m³/s
  - LobithW better at >6000 m³/s (but limited data)
- **LobithNN better than previous version** of LobithNN
  - Especially in the first forecast step (MAE ~11 m³/s vs 22 m³/s)
- **Improvement in operational reliability**
  - By using fewer but more reliable measurement stations

More accurate, more stable - and still room to grow





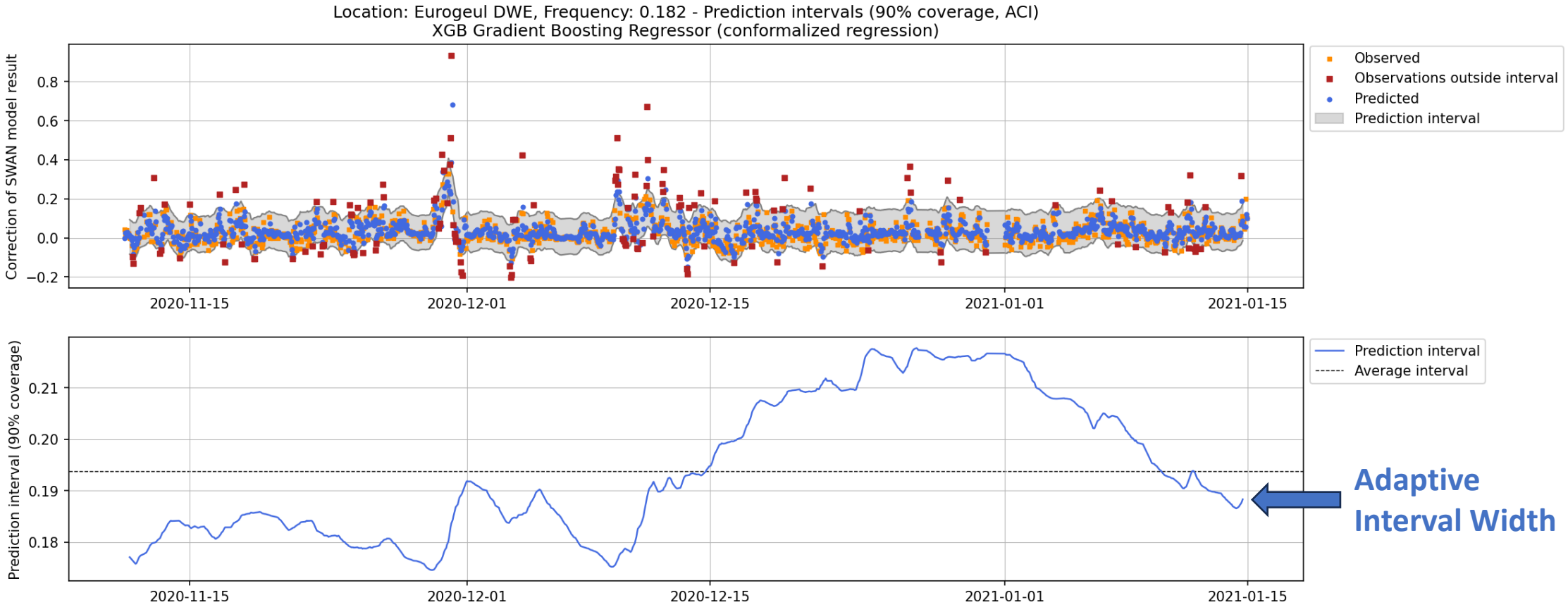
Robustness: Handling missing data

- **Detecting missing data** and anomalies
- **Filling gaps** in time series
  - Imputed values (based on statistics / correlations)
  - Label indicating imputed values
- **Re-training** of LobithNN
- **Validating** robustness

Reliability: Weather ensembles & model uncertainty combined

- **Current reliability**
  - Ensemble forecasts of river discharge
  - Only accounts for uncertainty in forcings
- **Future reliability**
  - Conformal prediction intervals for each ensemble member
  - Combined into a total ensemble forecast
  - Accounts for both model uncertainty and uncertainty in forcings
  - Conformal Prediction guarantees true value falling within the calibration interval (for example 70%)
  - Interval widths are expected to be larger than current
- **Next steps**
  - Calibrate ACI (Adaptive Conformal Intervals) on LobithNN forecasts
  - Combine intervals per ensemble member into one total forecast

Reliability Example: Correction of SWAN wave forecasts



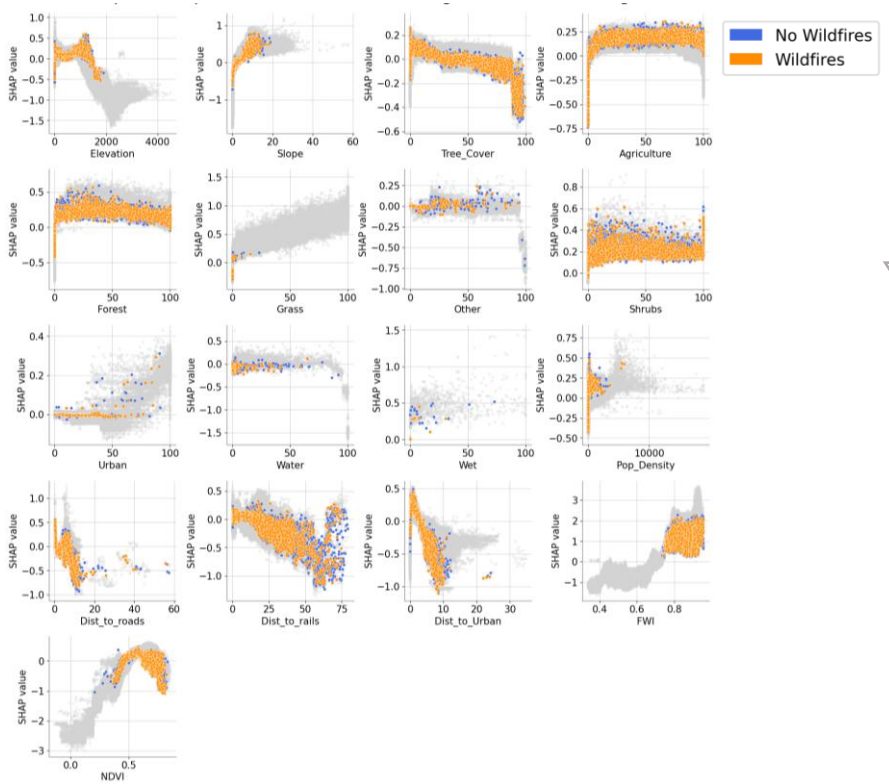
ACI (Adaptive Conformal Inference) *updates the width of prediction intervals dynamically, based on the distribution of past residuals (errors in the model's previous predictions). This approach* ***uses historical data to adjust its intervals over time in an adaptive way****, which is important in time series, as the underlying data distribution may change (non-stationarity).*

Explainability: Going beyond SHAP with causality

	Model-logica (XAI)	Data-logica (causal discovery)
What do you analyze?	Which variables contribute to the prediction?	Which variables actually influence the target?
What do you base this on?	SHAP contributions in relevant subsets	Causal relationships in the data within the same subsets
What do you want to know?	Is this variable important for the model?	Is this variable causally relevant in the data?
What do you test?	Overlap between explanatory contribution and causality	Consistency between model and real-world processes
What does deviation mean?	Model uses irrelevant or unexplainable input	Model ignores causal factors

Explainability Example: Wildfire susceptibility

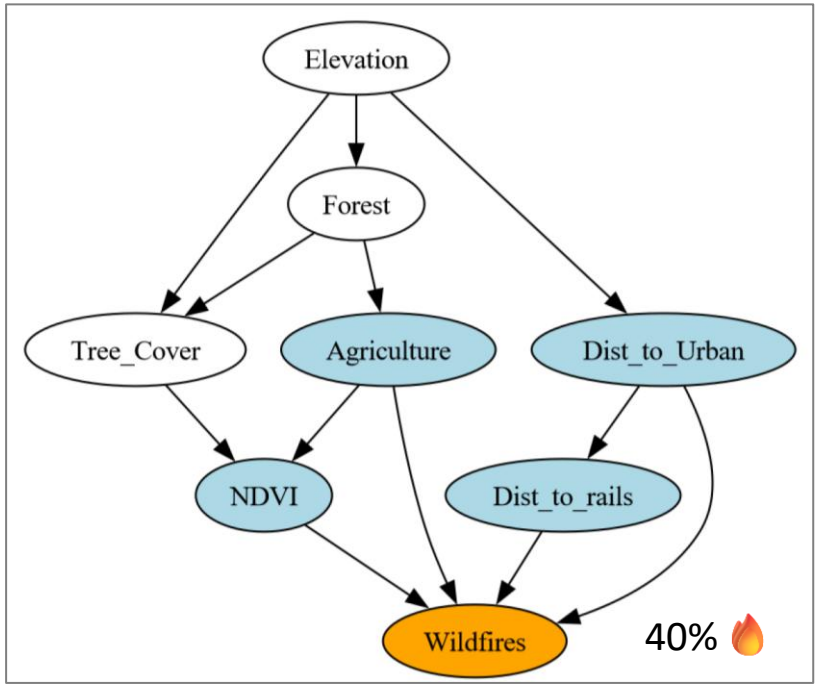
What the model has learned



SHAP contributions per point per feature

?

What the data tell us



Causal relations in the dataset

What the expert/stakeholder expects

Did the model learn the right mechanisms?

Accountable AI needs a human in the loop

But a loop isn't accountable unless it's understandable
And it won't be understandable without the right tools

Ask me anything?

hans.korving@deltares.nl