MAELSTROM

A machine learned weather forecast for Norway

Thomas Nipen (MET Norway)

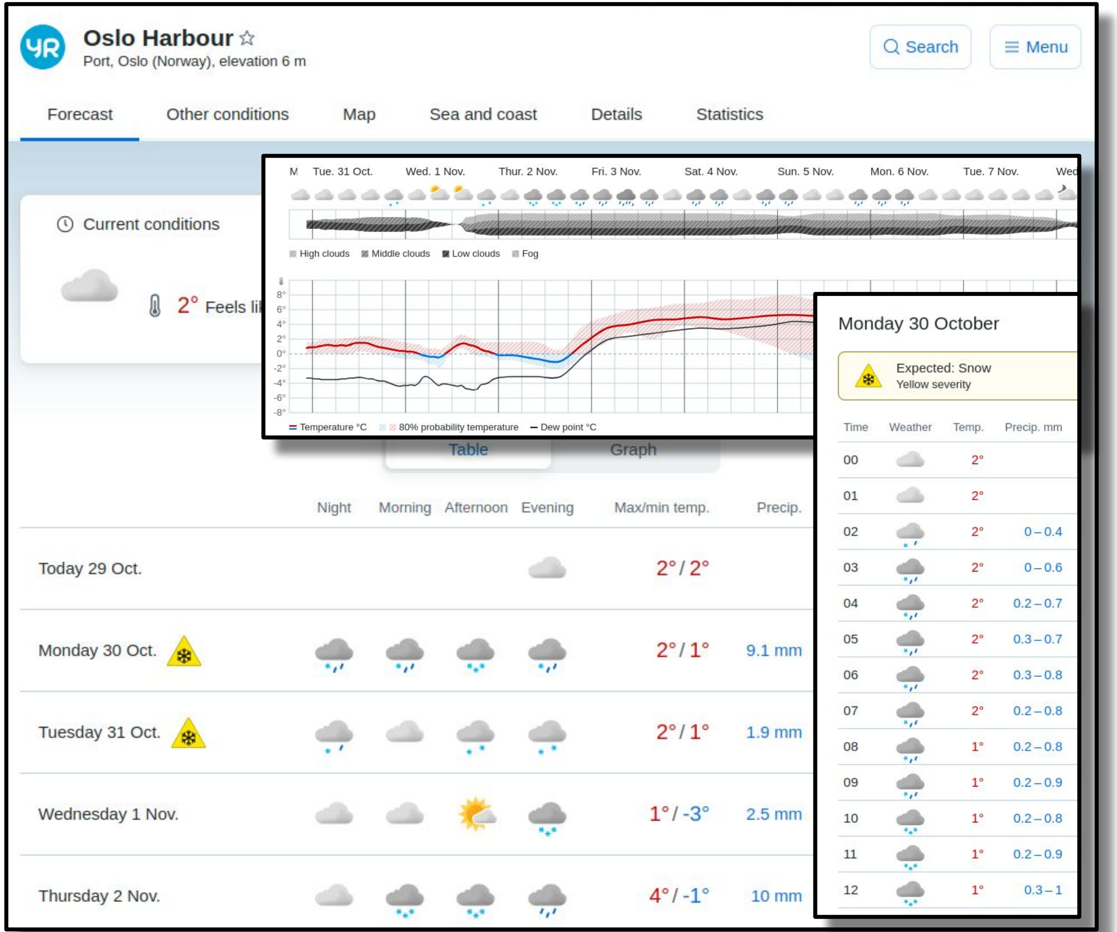
"The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway".

EuroHPC Joint Undertaking

Co-ordinated by ECMWF

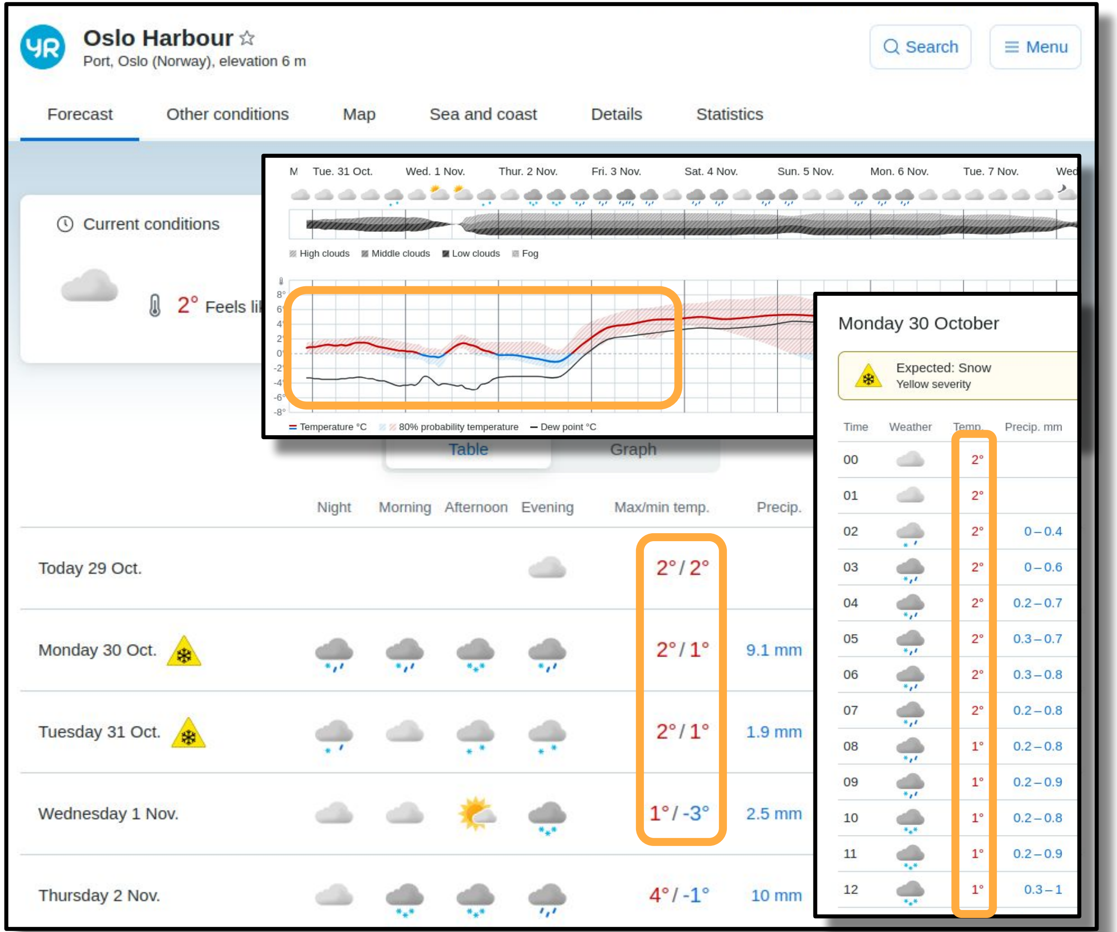
www.yr.no

10 million weekly users



www.yr.no

10 million weekly users

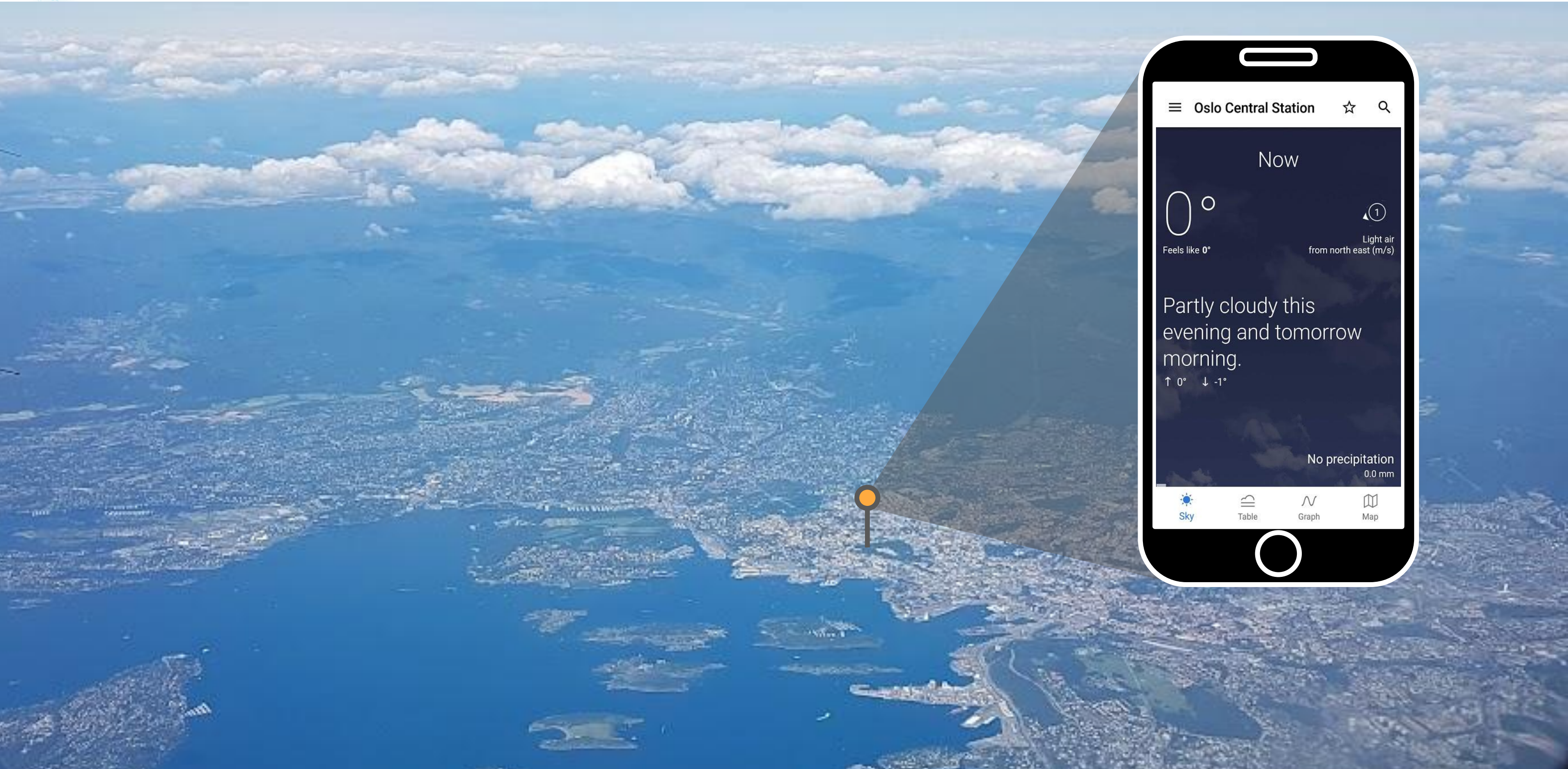


Part 1: Application

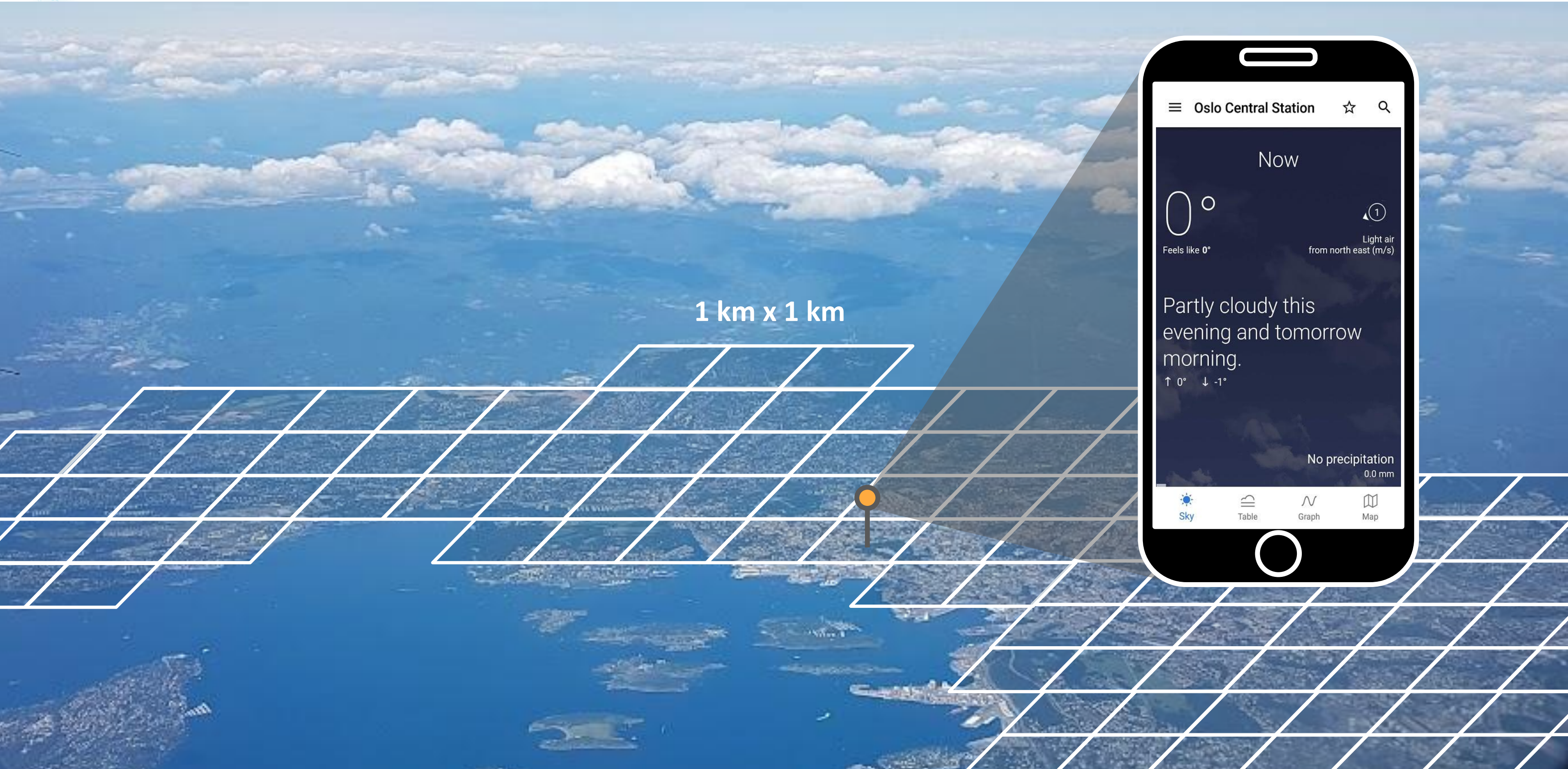
Part 2: ML solution

Part 3: Evaluation

5 Users expect high resolution forecasts

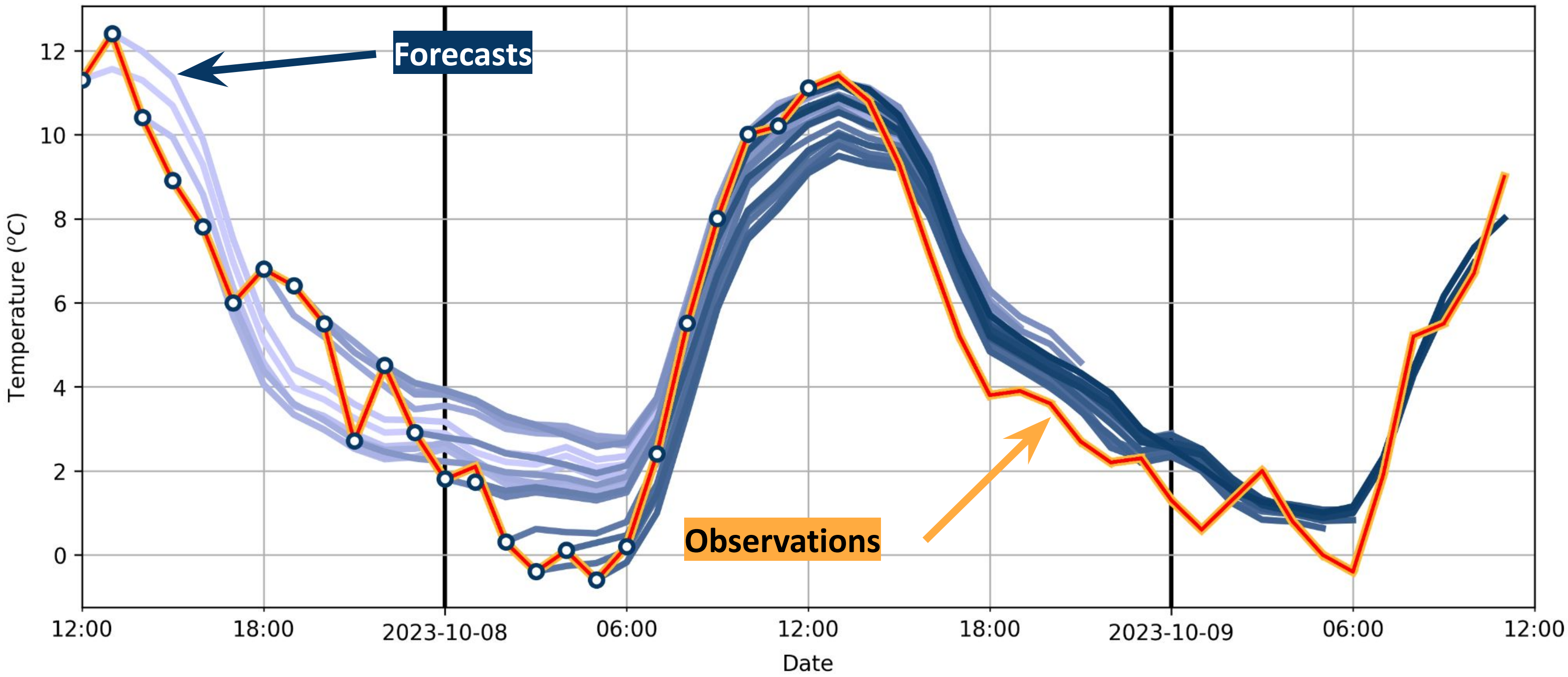


6 Users expect high resolution forecasts



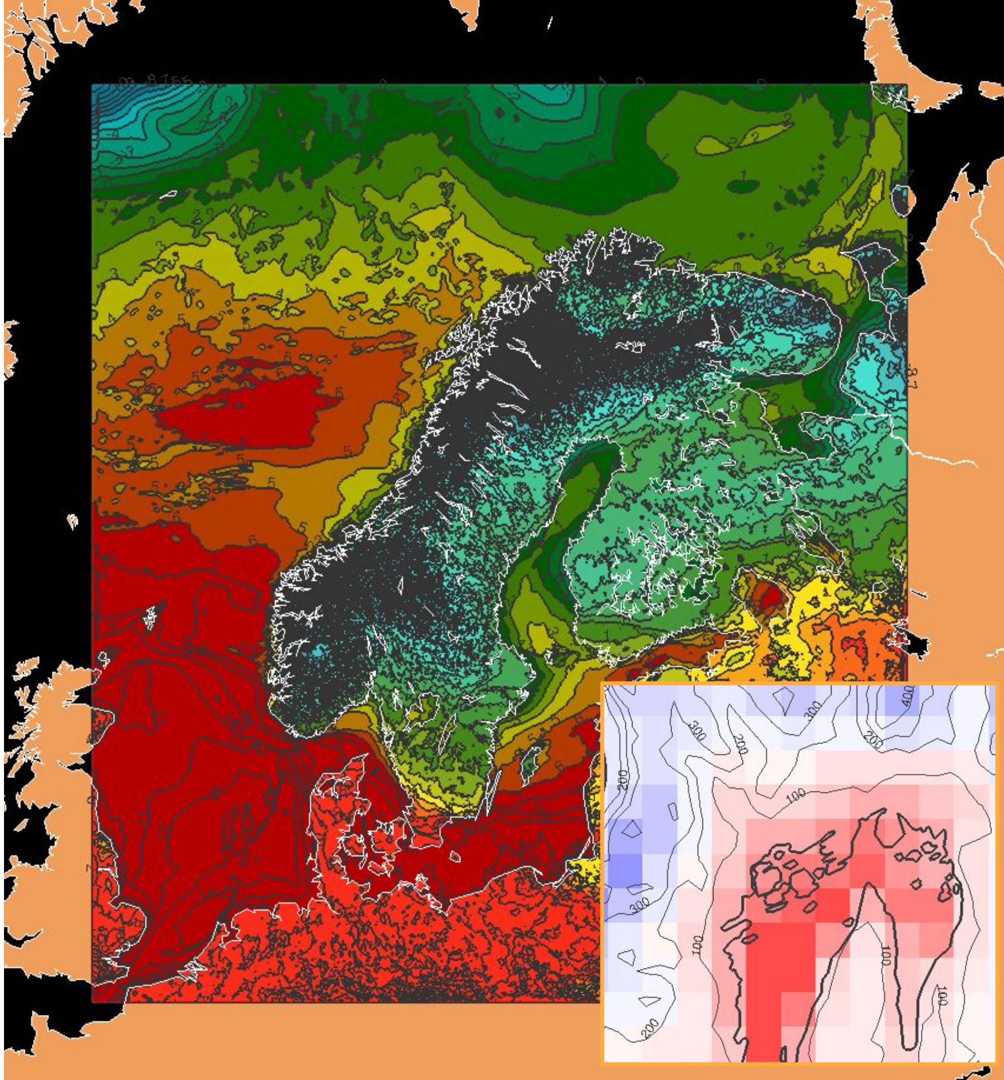
7 Users expect up-to-date forecasts

- New forecasts issued every hour as new observations become available



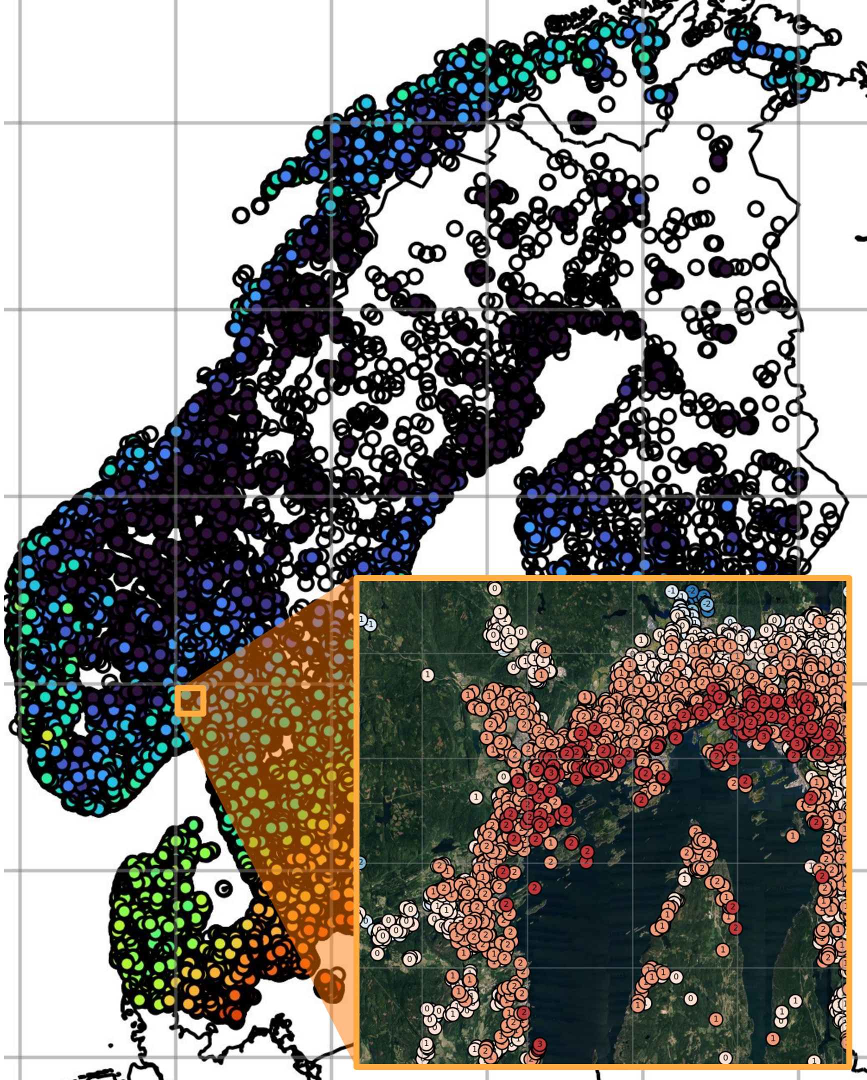
Predictors

- High resolution NWP ensemble (2.5 km)
- Hourly output for 59 hours
- Predictors:
  - 2m temperature (ensemble control)
  - 2m temperature (ensemble 10%)
  - 2 temperature (ensemble 90%)
  - 1h precipitation accumulation
  - Cloud cover
  - 10m wind (x-component)
  - 10m wind (y-component)
- Metadata variables:
  - Model altitude
  - Model land area fraction
  - “Real” altitude (1x1 km)
  - “Real” land area fraction (1x1 km)
  - Model x-coordinate
  - Model y-coordinate



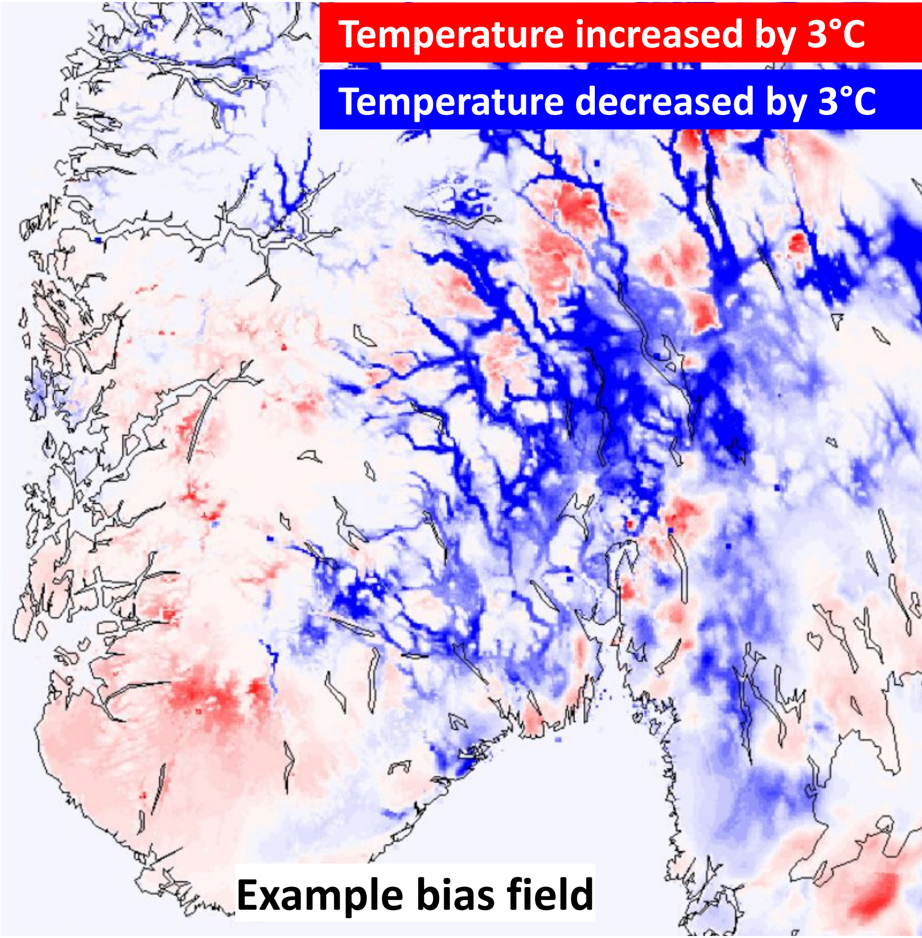
9 Target data

- Challenge to assemble an accurate target at high resolution
- Conventional observation networks are too sparse (at least in the Nordics)
- Citizen observations are an emerging data source (50-100x increase compared to SYNOP network)
- Target field based on:
  - Citizen observations
  - Early lead times (3-9h) from NWP
  - Combined using optimal interpolation (OI)



10 Gridded truths as input predictors

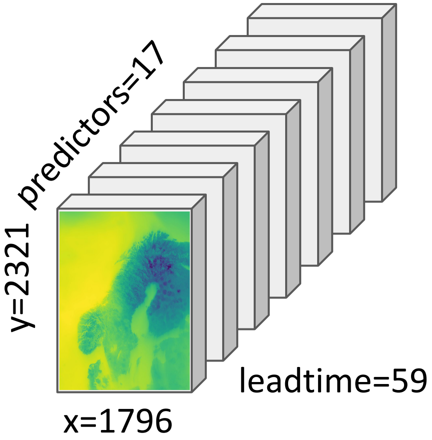
- Target fields for the 24h leading up to prediction also used as input predictors
- Allows us to keep forecasts up to date with recent observations
- NWP bias (target - NWP) used as predictor



Prediction problem

Input data (6 terabytes)

- 1x1 km downscaled NWP and recent biases
- 59 x 2321 x 1796 x 17
- 700 samples (2 years)

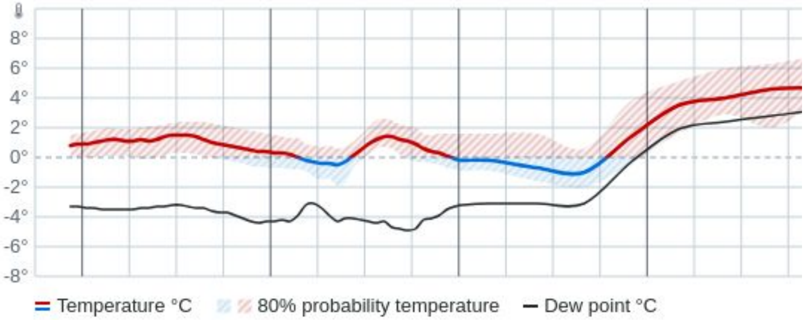


Output

- 1x1 km temperature forecasts
- 59 x 2321 x 1796 x 3 (10, 50, 90% quantile levels)

Target data

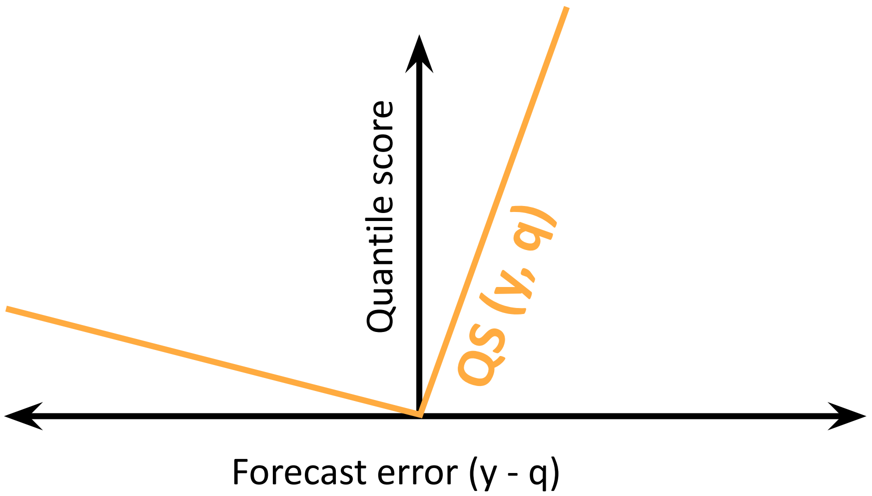
- 1x1 km gridded truth
- 59 x 2321 x 1796



12 Loss function

- Quantile scoring function is used to evaluate quantile forecasts (10, 50, 90%)

Example for quantile level = 0.9



Quantile score

$$QS_\tau(y,q) = \begin{cases} (y-q)(\tau-1)) & y \leq q \\ (y-q)\tau & y > q \end{cases}$$

τ : quantile level
y : observation
q : quantile

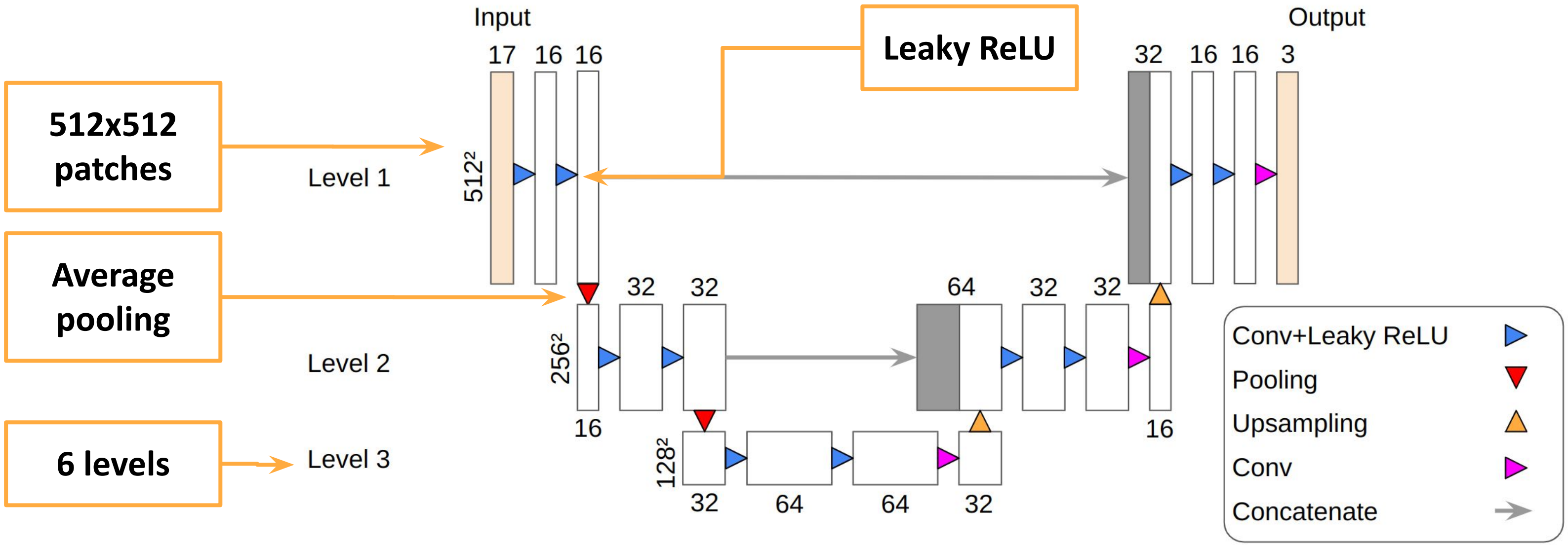
Part 1: Application

$\boxed{\text{Part 2: ML solution}}$

Part 3: Evaluation

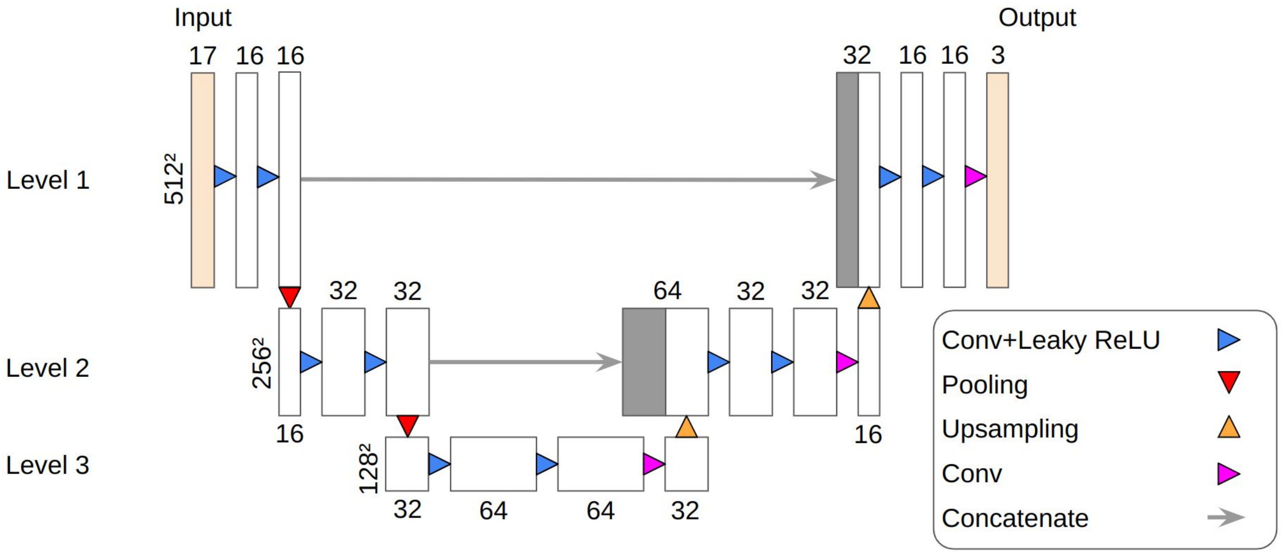
U-Net

- 2D U-Net, all leadtimes trained together (leadtime added as a predictor)
- 1,314,019 trainable parameters



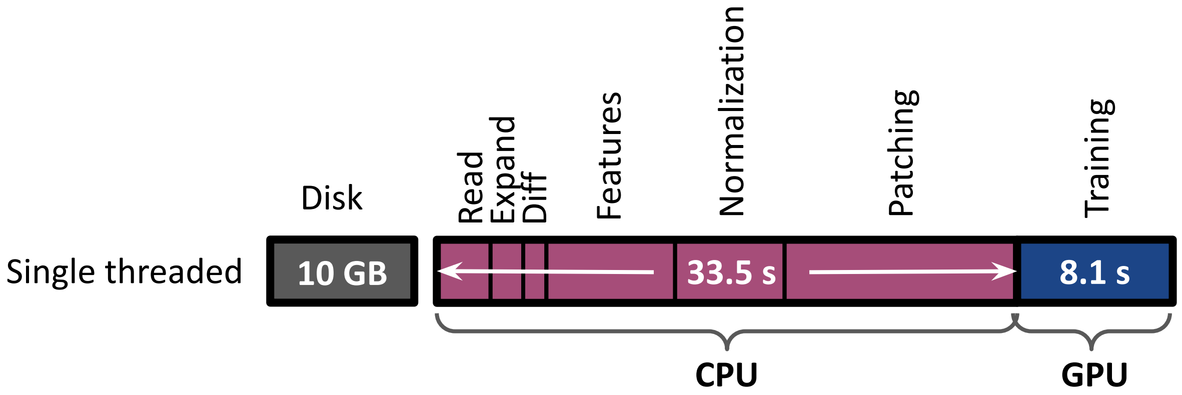
U-Net

- 2D U-Net, all leadtimes trained together (leadtime added as a predictor)
- 1,314,019 trainable parameters
- Trained on 4 NVIDIA A-100 GPUs, 2x24 cores AMD EPYC 7402, 512GB RAM
- Extensive optimization of processing performance and memory footprint



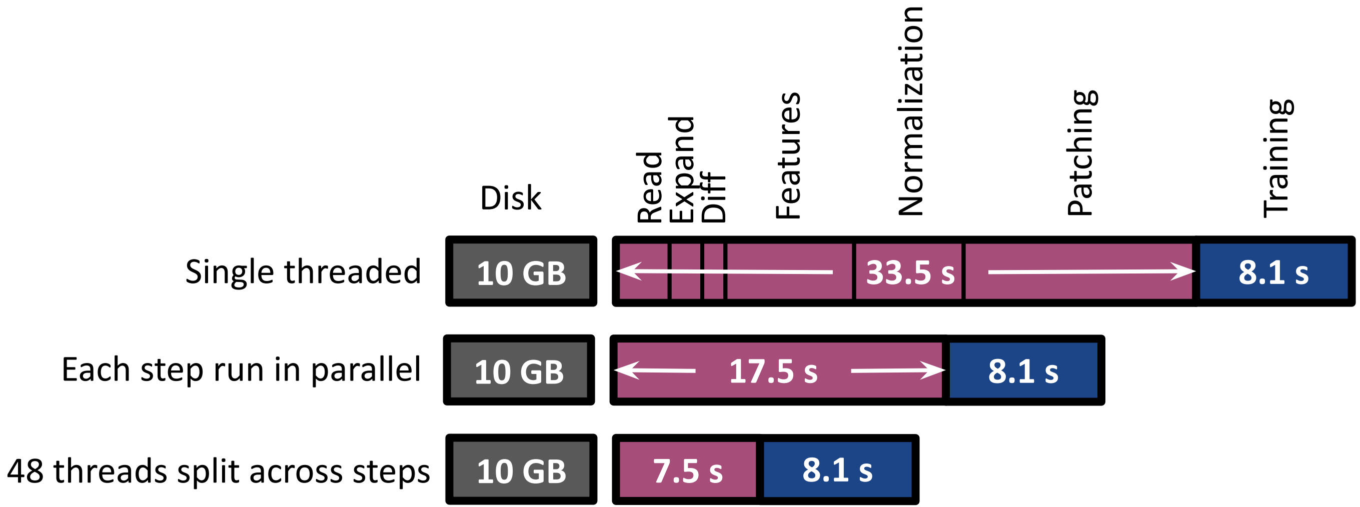
16 Optimizing the data loader

- We needed a data loader that:
  - Streames data from disk (6TB too large for memory)
  - Doesn't cause an I/O bottleneck
  - Can read data as we have them stored on our systems (i.e. reusable in other applications)
  - Allows loading options to be easily changed



17 Optimizing the data loader

- We needed a data loader that:
  - Streames data from disk (6TB too large for memory)
  - Doesn't cause an I/O bottleneck
  - Can read data as we have them stored on our systems (i.e. reusable in other applications)
  - Allows loading options to be easily changed



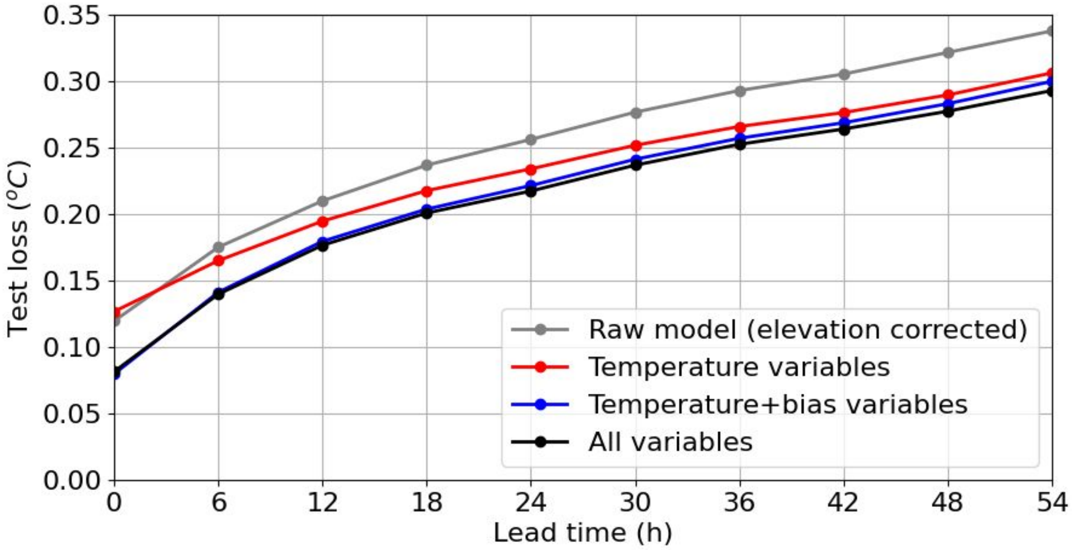
Part 1: Application

Part 2: ML solution

Part 3: Evaluation

Verification

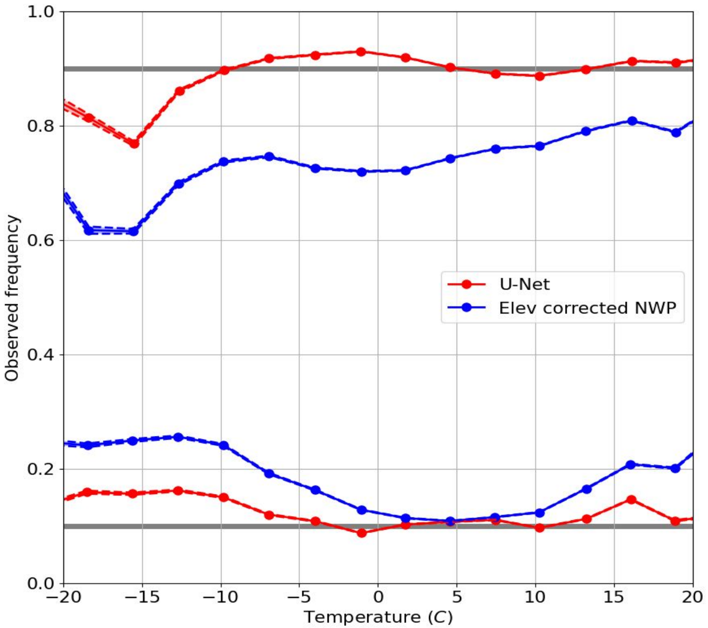
- 1 year training period, 1 year testing period
- Bias variables are important contributors to overall skill of forecast
- Precip/winds/clouds also have a (small) positive effect



Air temperature (ensemble control)
Air temperature (ensemble 10%)
Air temperature (ensemble 90%)
Altitude (model and "real")
Bias yesterday
Bias right now
Land area fraction (model and "real")
Precipitation
Winds
Clouds

20 Verification

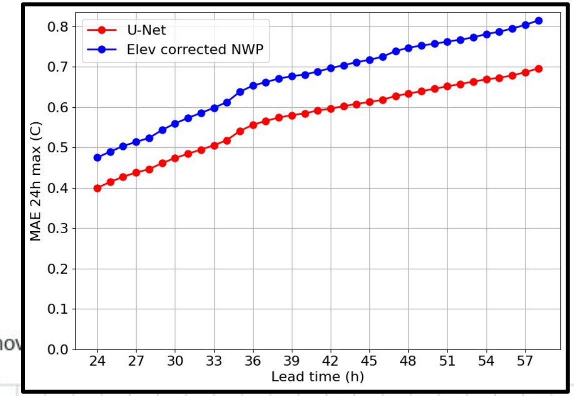
- 1 year training period, 1 year testing period
- Bias variables are important contributors to overall skill of forecast
- Precip/winds/clouds also have a (small) positive effect
- 10 and 90% quantiles are much more reliable

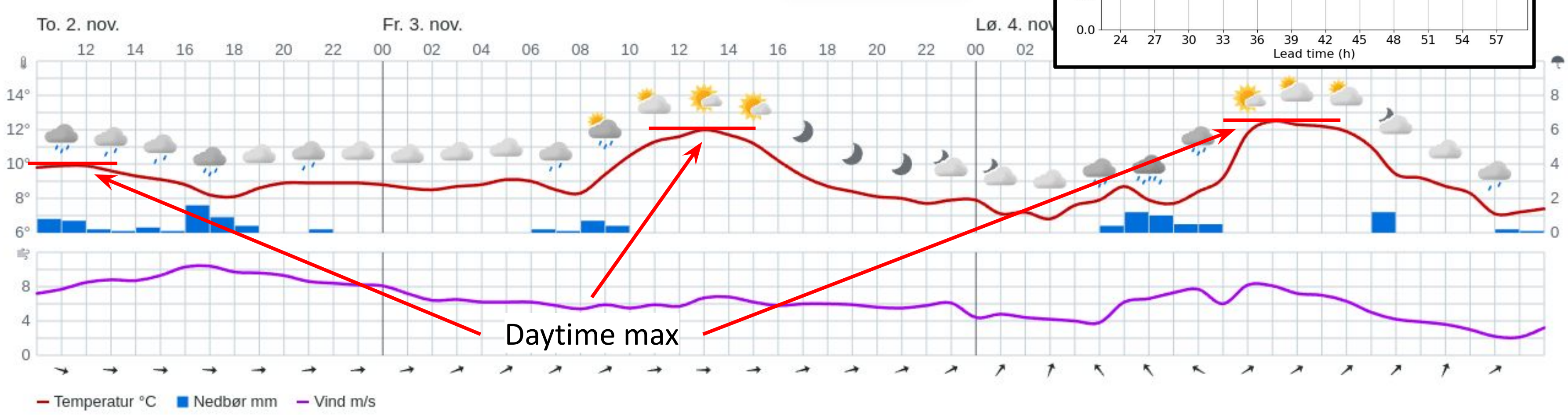


21

Verification

- Other properties not captured by the loss function:
  - Daily min/max
  - Sharp temporal changes
  - Spatial consistency (users comparing different locations)
- So far, these metrics also look promising





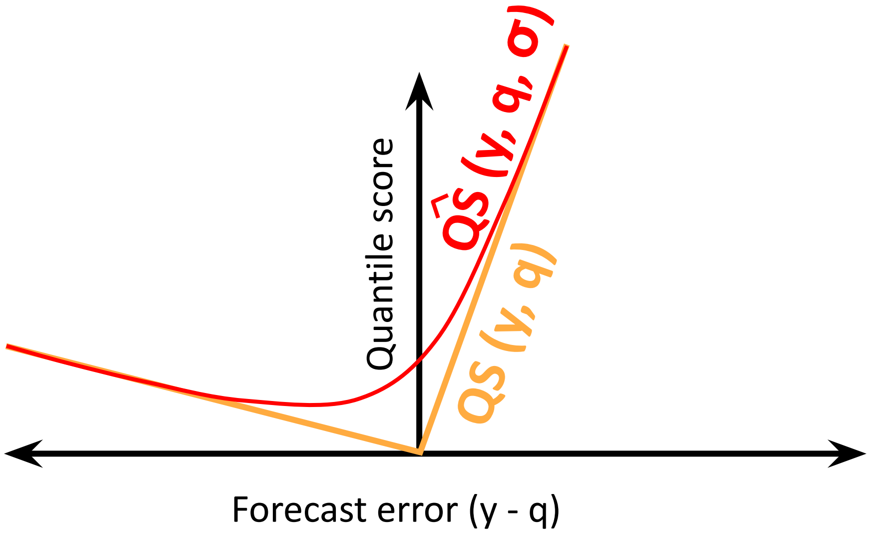
Summary

- The MAELSTROM project has contributed to:
  - Development of an ML solution for forecasting temperature suitable for the general public
  - Optimization of the training pipeline by exploiting the available hardware
  - Development of a high-resolution benchmarking dataset for testing new ML methods
- Links:
  - Forecast site: www.yr.no
  - Data access via climetlab: https://github.com/metno/maelstrom-yr
  - Jupyter notebooks: https://gitlab.jsc.fz-juelich.de/esde/training/maelstrom_bootcamp (AP1)
  - Contact: Thomas Nipen (thomasn@met.no)

Loss function

- Quantile scoring function is used to evaluate quantile forecasts (10, 50, 90%)
- Convolve the forecast error with the target uncertainty
- Developed a computationally efficient approximation

Example for quantile level = 0.9



Quantile score

$$QS_\tau(y,q) = \begin{cases} (y-q)(\tau-1)) & y \leq q \\ (y-q)\tau & y > q \end{cases}$$

τ : quantile level
y : observation
q : quantile

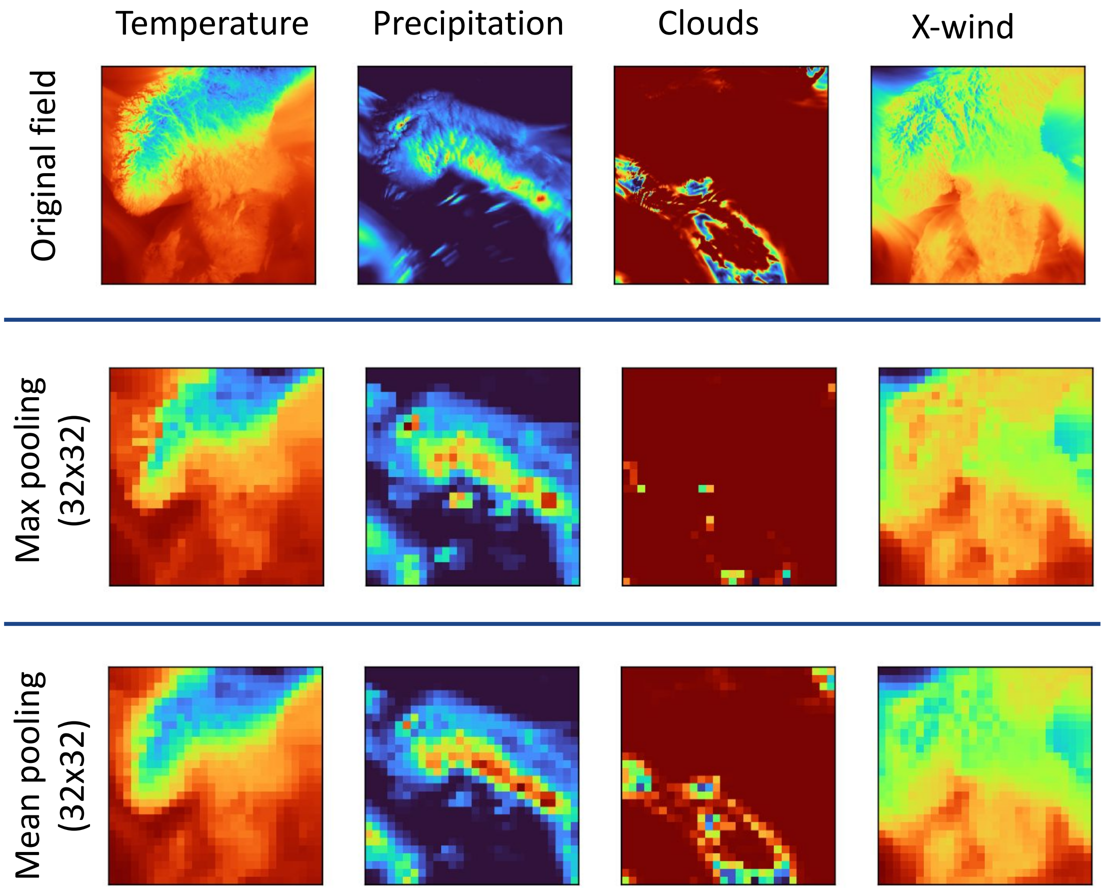
Quantile score (uncertain target)

$$QS_\tau(y,q,\sigma) = \int_{-\infty}^{+\infty} QS_\tau(y-x,q)\phi\left(\frac{x}{\sigma}\right)dx$$

$$\hat{QS}_\tau(y,q,\sigma) = QS_\tau(y,q) + 0.4\sigma e^{\frac{1.4|y-q|}{\sigma}}$$

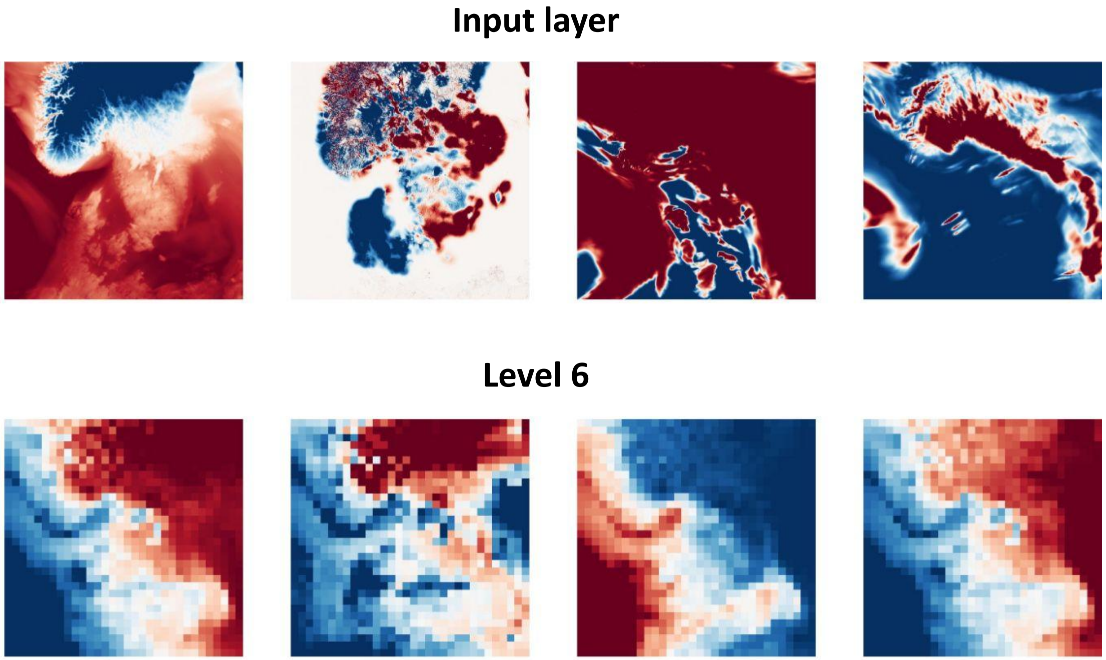
Average pooling

- Improvements also found in U-Net in MAELSTROM application 5



Leaky ReLU

- Standard ReLU activation disabled the layers in the U-Net due to dead ReLU nodes.
- Discovered through visualizing the tensors as they pass through the network




Verification

- 10 and 90% quantiles are much more reliable
- Still, reliability has regional patterns suggesting room for improvements