Combining machine learning and data assimilation

to improve hydrodynamic forecasting in a tidal estuary

Clemens Cremer, Jesper Sandvig Mariegaard, Henrik Andersson, Henrik Madsen
DHI A/S, Denmark

30 May 2022

Background and Motivation

Area of interest and challenge

Background and motivation: **Port of Hamburg**

- To ensure
  - navigability,
  - nautical safety,
  - efficiency in port operations

- knowledge of hydraulic conditions is crucial
  - Nautical navigation
    - Bridge passage heights
    - Under keel clearance
  - Accident investigation
  - Reduction of maintenance (dredging)

IHMA

DENMARK
COPENHAGEN
North sea
Baltic sea
UNITED KINGDOM
Elbe estuary
Port of Hamburg
HAMBURG
AMSTERDAM
THE NETHERLANDS
GERMANY

Background and motivation: **Operational model**

- numerical model of Elbe Estuary
  - Two dimensional
  - DHI Software 2D MIKE FM
- Online (**real-time**)
  - In operation for ~10 years with improvements in between
  - Hindcast 1.5 h
  - Forecast 8.5 h
  - Model run / Updated hourly

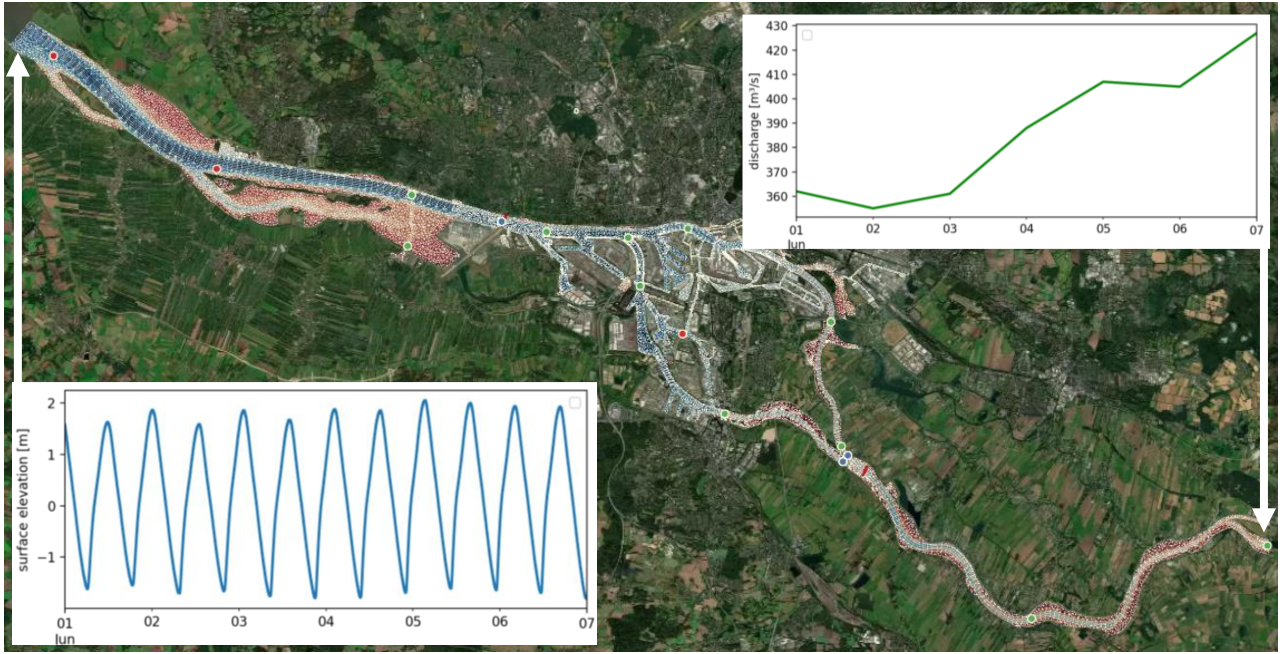
Background and motivation: **Characteristics** of current model

Boundary conditions

- Water level (downstream)
  - Water level
    external prediction
- Discharge (upstream)
  - naïve forecasting (last value)

Further model parametrization

- Bed resistance
- Consideration dynamic wetting and drying (tidal flats)



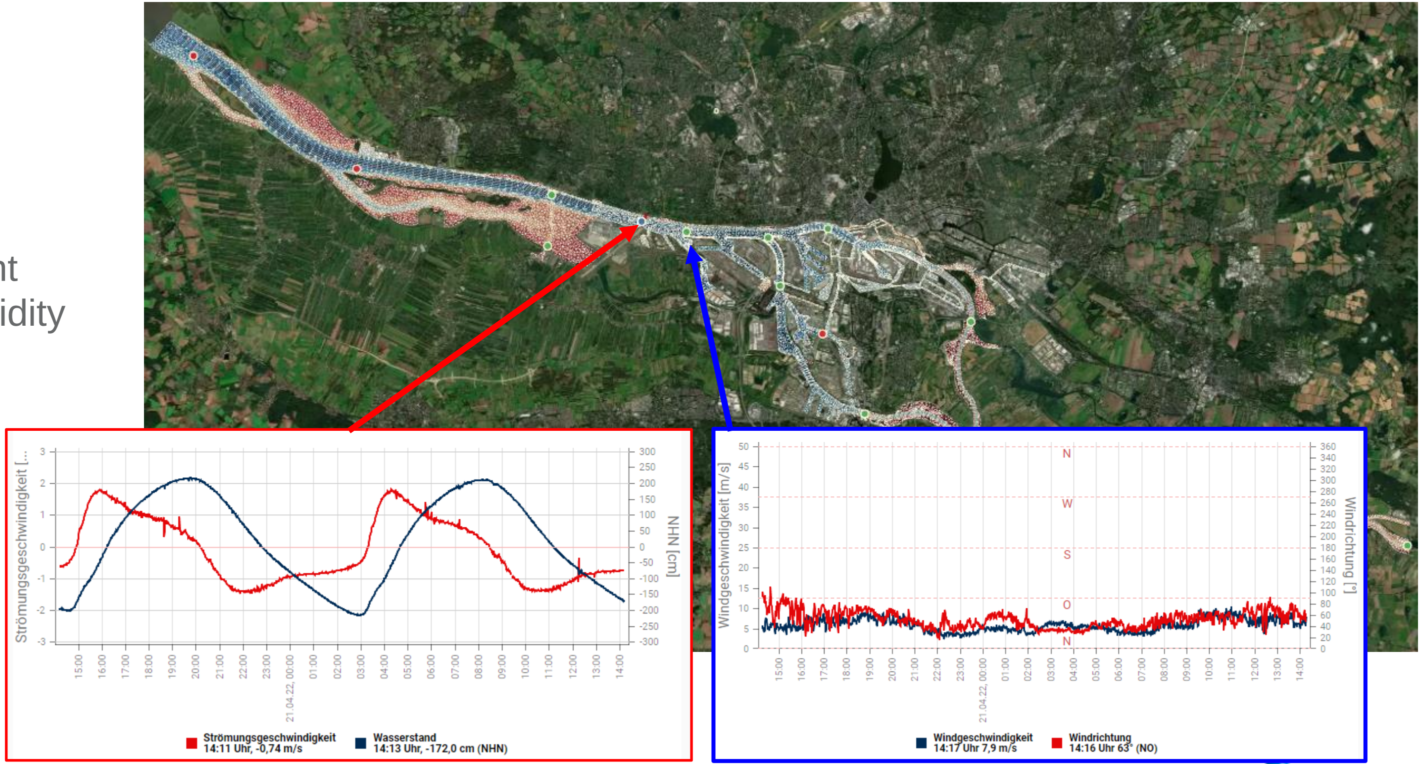
Problems:

- **Analysis shows strong uncertainty in provided water level boundary condition for forecast**
- **Further uncertainty in model parametrization (no meteorological forcings, bathymetry, roughness, geometry)**

Assessment of model skill

Model skill assessment: **Observations**

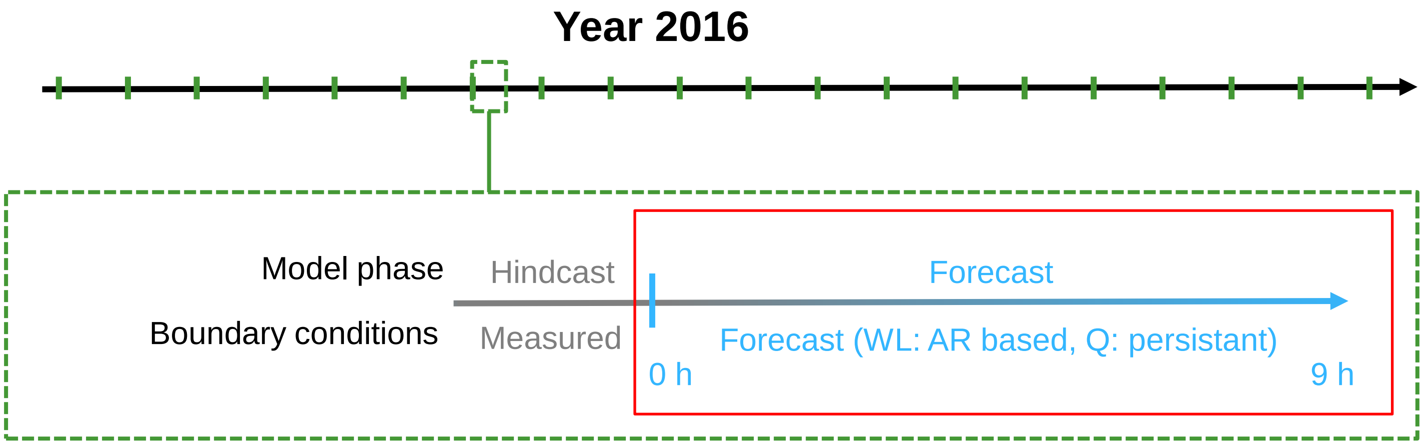
- **Water level gauges**
- Currents
- Discharge
- Wind
- Suspended Sediment Concentration / Turbidity



Current speed and water level

Wind speed and direction

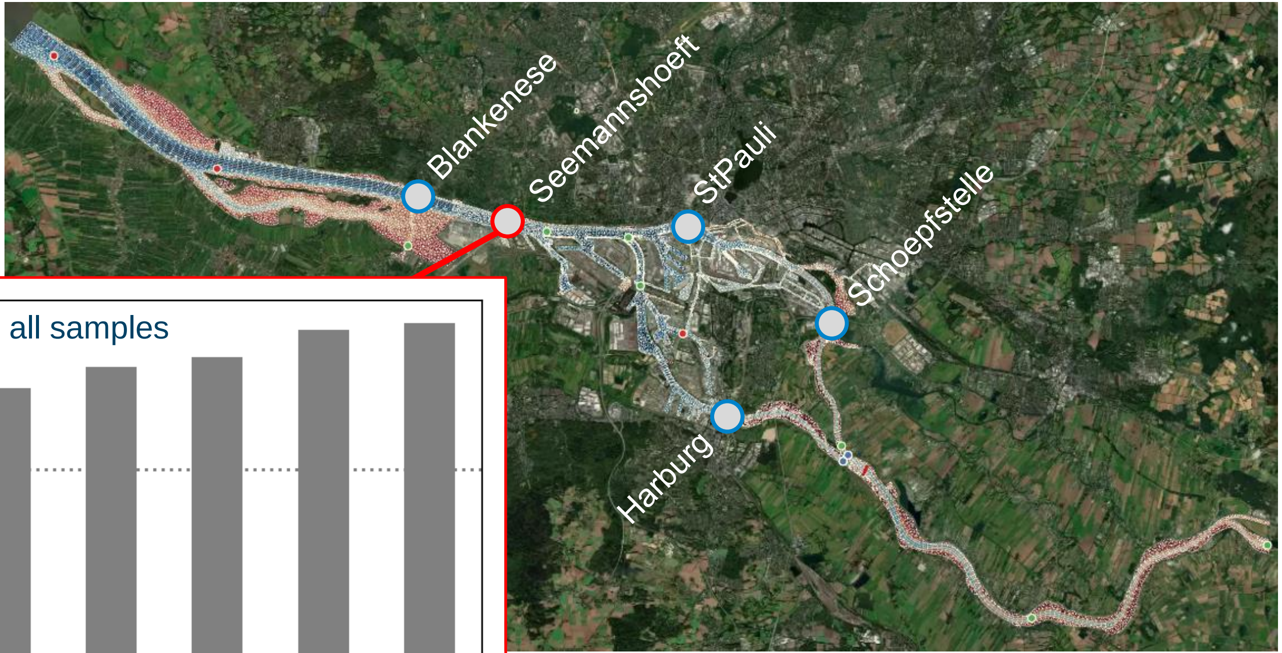
Model skill assessment: **Forecast sampling**

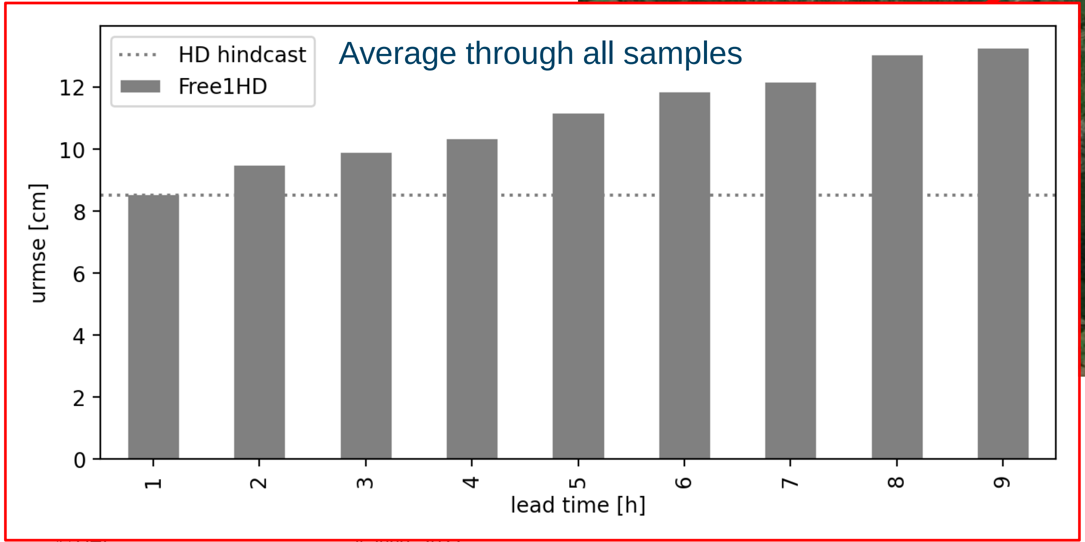


- Extracting 100 samples for individual forecast
- Throughout one year (2016)
- Forecasts initialized with water levels and current speeds in domain (from Hindcast)
- Boundary conditions
  - Waterlevel forecast: artificial forecast, mimicking externally provided (AR process)
  - Discharge forecast: with persistence (last observed value)

Model skill assessment: **Observation gauges**

- Evaluating unbiased rmse
- All samples average
- urmse increasing with lead time





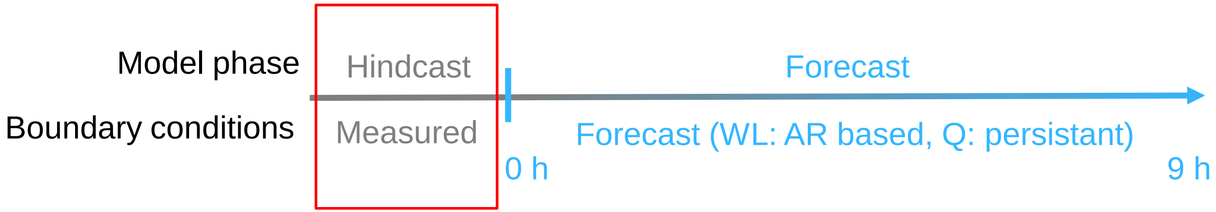
Improving the operational model

Data assimilation

Solution? The promise of data assimilation

- Data assimilation
  - **Continuous update with observations** / model skill assessment
  - Spatial and **multivariate** propagation / interpolation of point information
  - **Physically consistent** representation

→ Without future observations: Improve Hindcast and initial conditions

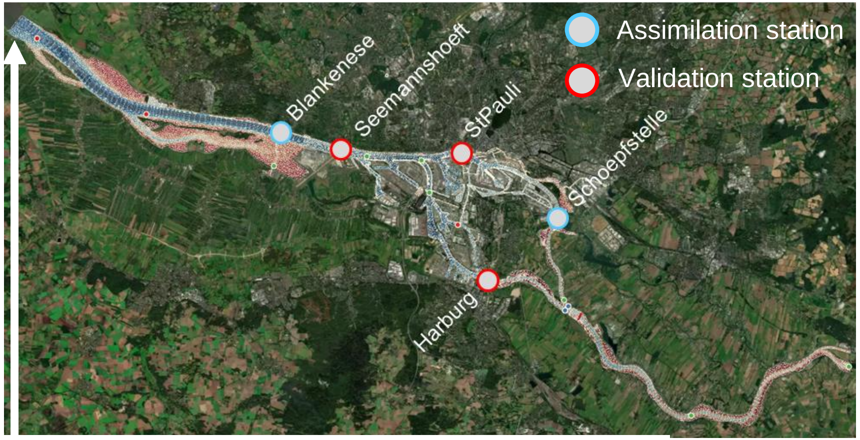


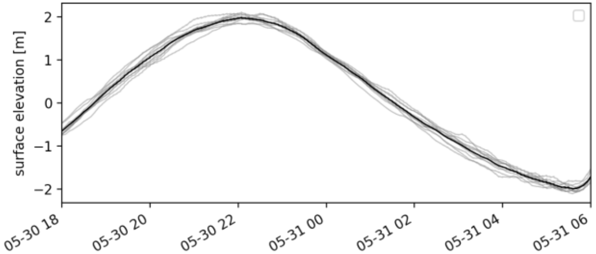
Solution? Assimilation setup

- Application of EnKF (serial, “Potter-scheme”)
  - Temporal smoothing
  - No localization

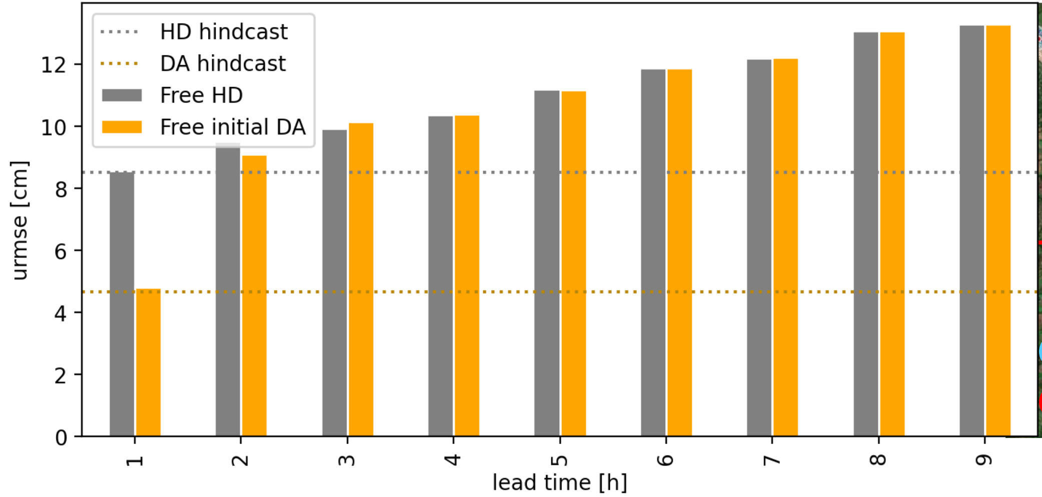
- **Ensemble** created
  - with 10 members
  - by perturbation of the water level boundary conditon
  - perturbation propagation via AR(1) process with a half-life of 3 hours
  - Sampled from gaussian with std dev of 0.2 m

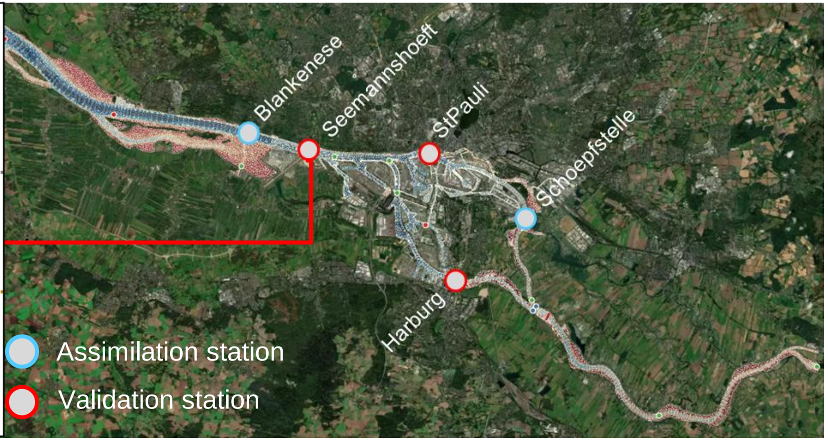
- Two stations utilized for **assimilation**





Solution? First results with data assimilation





- DA with observations in hindcast
- Initial conditions with DA improve next 2 hours of forecast
- Reversion to non-DA forecast results afterwards

→ Dynamical system **strongly driven by boundary conditions**

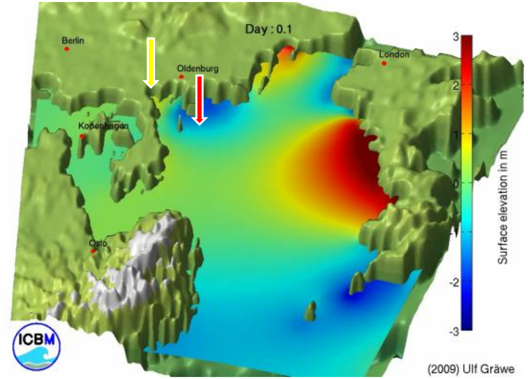
→ To improve long term forecast better future "observations" for assimilation required

Providing future „observations“ for assimilation

Via a machine learning approach

Solution Part 2? The promise of ML predictions

- **Fast timeseries forecasting**, suited for prediction
- **Easy to consider features** (e.g. wind, more distant gauges outside of model domain)

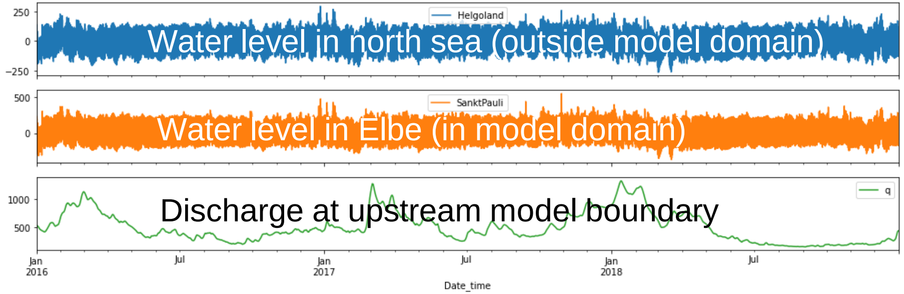


- **Effects not considered in numerical model** might be learned and considered by data-driven model (e.g. seasonal vegetation)
- Best case: easy setup and little calibration for decent results
  - Tweaking of parameters and hyperparameters potentially easier than obtaining and processing input data required for deterministic model

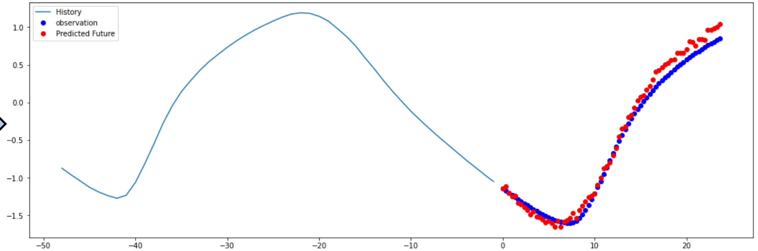
Solution Part 2? Forecasting water levels via LSTM

- Predict water levels at assimilation stations
- Water levels are a function of…
  - Wind speed and direction in north sea
  - Upstream Discharge
  - Water levels of neighboring stations
- **L**ong **S**hort **T**erm **M**emory model found suitable

Exemplary input features



LSTM water level forecast of next 6 hours



* coarse 15 min input data
ca. 1 min model training time (4 core CPU, 4 GB RAM)
TensorFlow

Solution Part 2? Setup and features of LSTM

Long **S**hort **T**erm **M**emory features

- **Water levels** in North Sea
- **Wind** speed and direction
- **Discharge** (upstream)
- **Time** (hour of day, day of year)
- Years 2017-2019 used for training and validation
- Split 80%-20%

Long **S**hort **T**erm **M**emory settings

- Input: 12 lagged hours of features
- Two LSTM layers (256-128 units)
- Output layer (Dense) → 12 lead hours

Hybrid model setups

Combining hydronumeric model with machine learning and data assimilation

Recap: Combining ML prediction with data assimilation

Data assimilation

- Continuous update with observations
- Spatial and multivariate propagation
- Physically consistent representation

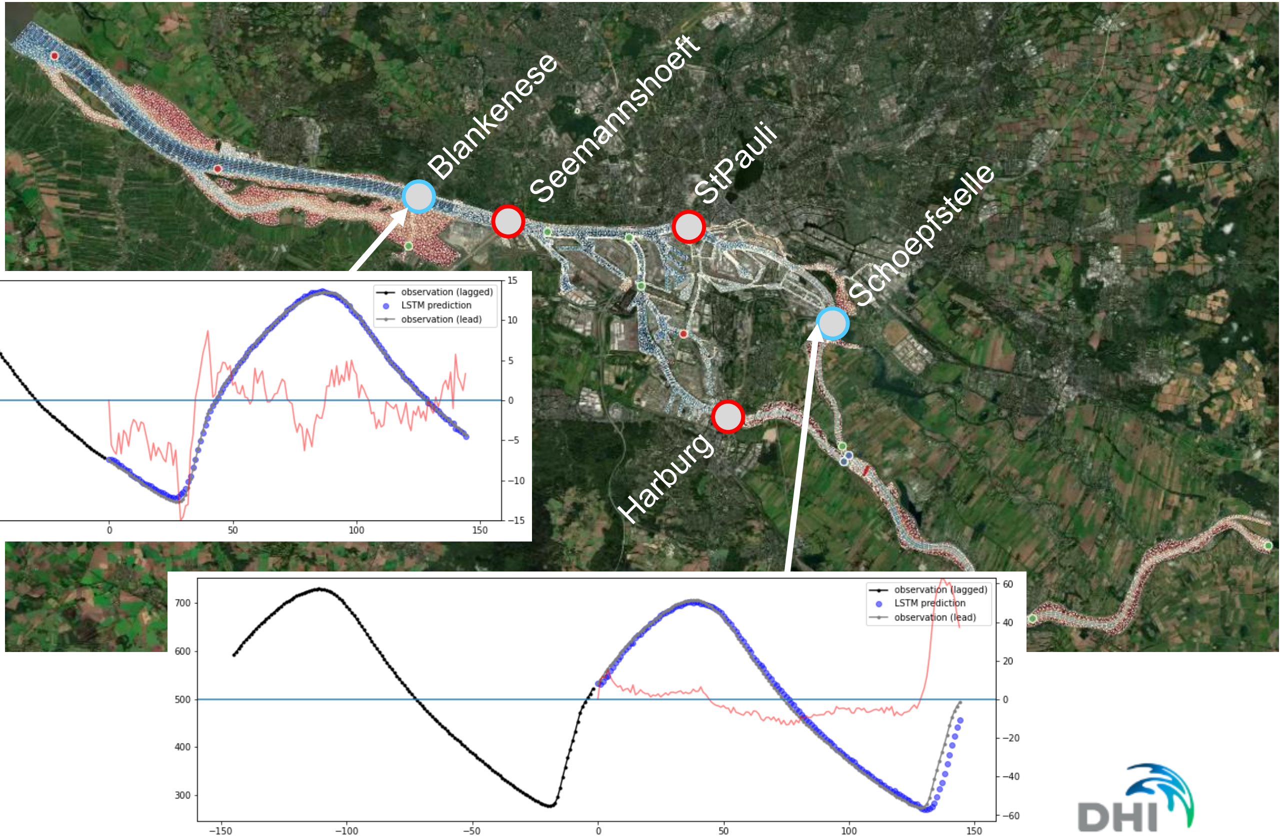
ML prediction

- Fast timeseries forecasting (“observations”)
  - Taking into account features outside of model domain

- Incorporation of ML predicted point “observations” in data assimilation
- Combining forecasts with different uncertainties into outcome with smaller overall uncertainty

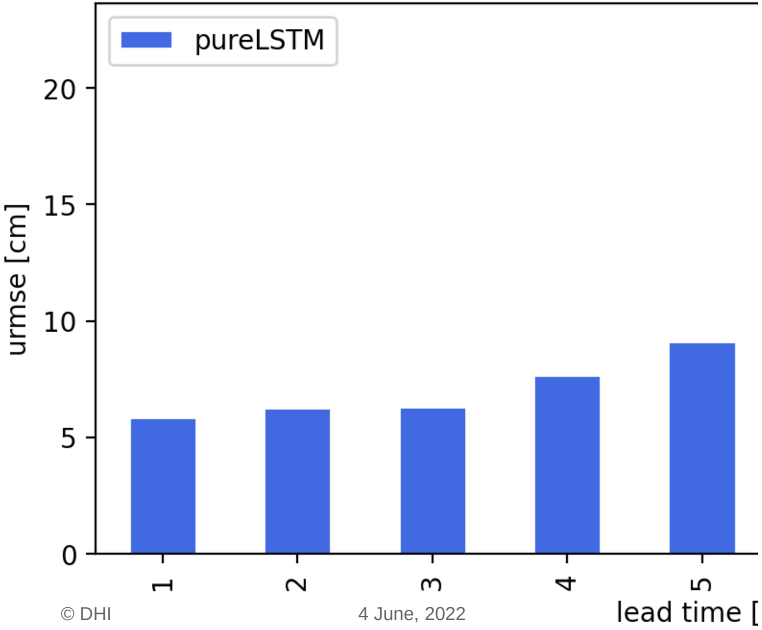
Combining ML prediction with data assimilation

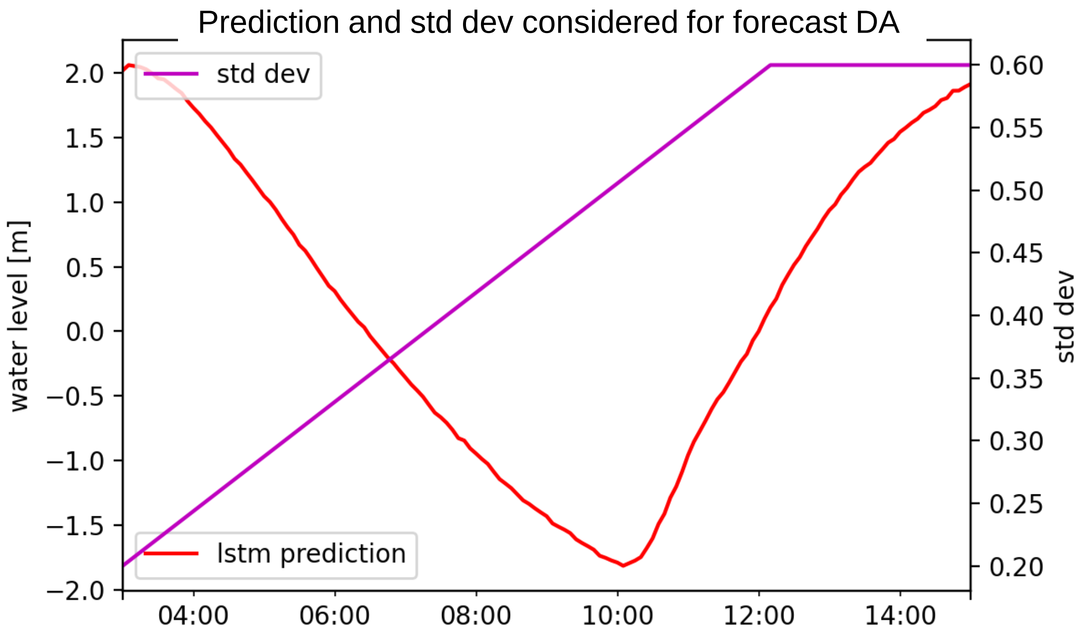
- **DA model utilizing ML predicted “observations”** at Blankenese and Schoepfstelle for assimilation



Combining ML prediction with data assimilation

- LSTM forecasts in general less reliable with increasing lead timesteps
- Considering decreasing confidence with time in LSTM predicted values DA

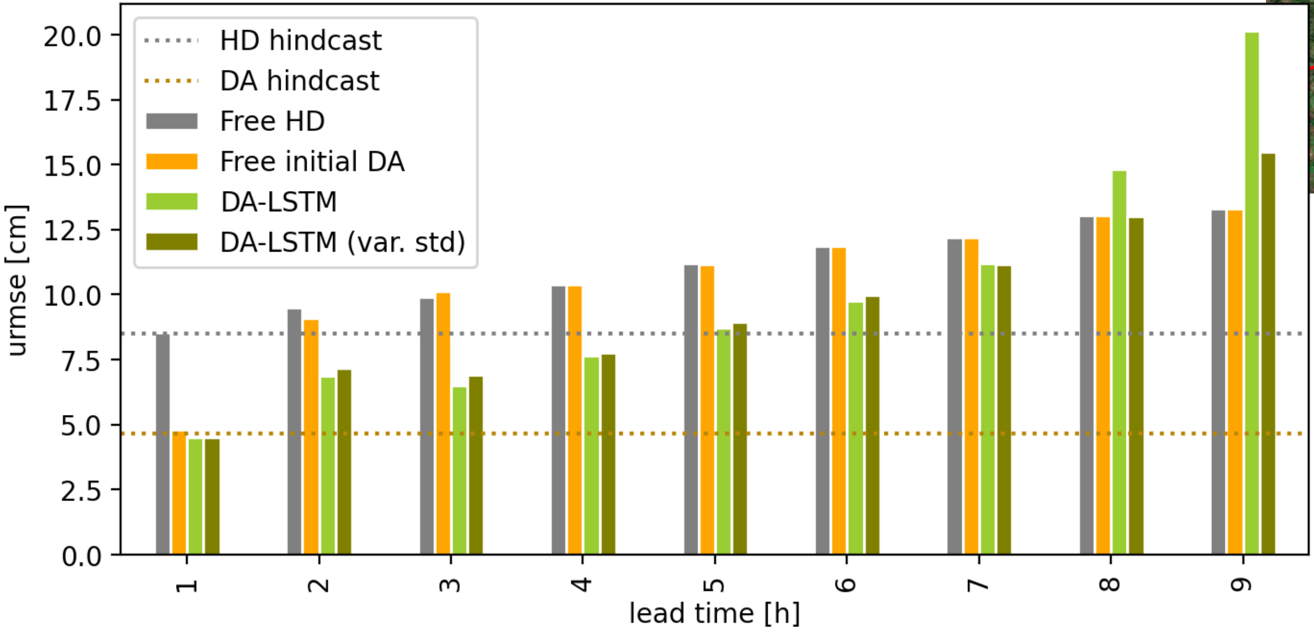




Results

Evaluating for different lead times

Comparing model performance



- Constant measurement error (std dev of 0.2) best performance up to ~7 hours lead time
- Considering stronger uncertainty in LSTM for larger lead times beneficial

Summary & Outlook

Summary & Outlook

- Combination of data-driven forecast and numerical model with data assimilation investigated
- Overall forecast quality improved
- it is worth to combine timeseries forecasting and data assimilation for a physically consistent, multivariate representation

- Future work will focus on
  - Evaluation of other variables (current speeds, discharge, sediments)
  - Improving LSTM timeseries predictions
  - Improve prediction of rare / extreme events

Thank you for your attention!