AI-Based Climate Data Processing: Cloud Removal and Weather Prediction for Agriculture

Pavel Kordík
DataLab, Faculty of Information Technology
Department of Applied Mathematics
Czech Technical University in Prague

October 22, 2025

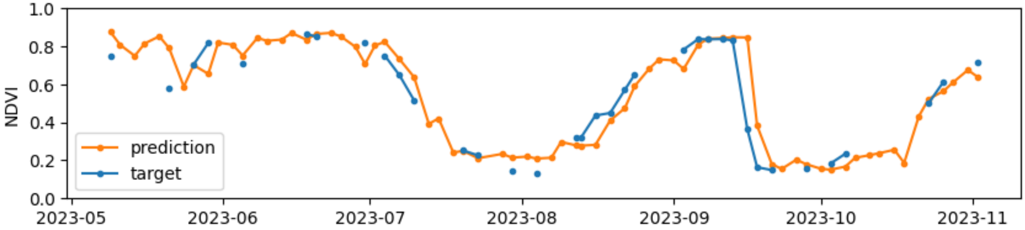
About DataLab @ FIT CTU

- Applied machine learning for Earth observation, meteorology, and fundamental AI research.
- Focus: cloud removal for optical satellites, NDVI (Normalized Difference Vegetation Index) time-series for crop health monitoring, local weather downscaling, and novel AI architectures.
- Notable achievements: Won NeurIPS 2022 challenges in weather prediction and exoplanet gas detection from spectral images.
- Contributions span research papers, open tooling, and operational prototypes with industry/academia partners.

Use Case — Cloud Removal: Motivation

- **Why**: Optical satellite monitoring is hindered by frequent cloud cover, breaking time-series needed for farm decisions.
- **Impact**: Missing observations delay detection of stress, pests, irrigation/fertilization effects, and yield risk.
- **Objective**: Restore gap-free reflectance/NDVI time series and provide confidence to support risk-aware analytics.

Use Case — Cloud Removal: Motivation (Example)

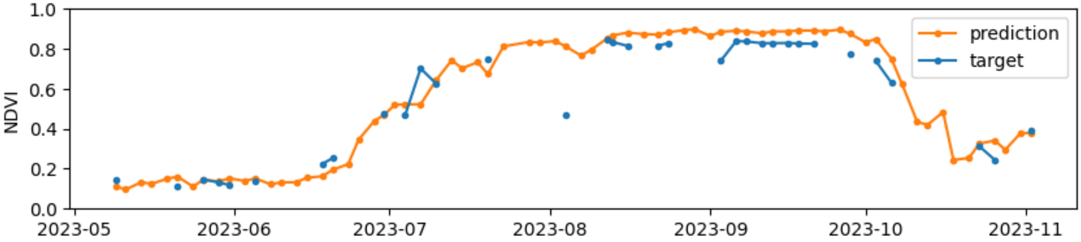


NDVI visualization for a representative area; cloudy periods disrupt continuity.

Use Case — Cloud Removal: Data

- **Optical**: Sentinel-2 L2A, 12 bands @ 10 m; SCL used for cloud/shadow masks.
- **Radar**: Sentinel-1 GRD (VV, VH) co-registered to Sentinel-2.
- **Context**: Optional VUMOP soil types; EUMETSAT LSA point products.
- **Sampling**: Czech Republic ROIs; 256×256 patches; 5-frame recent context; 2022 train, 2023 validation.
- **Prep**: Resample to 10 m, mask clouds for targets, normalize inputs.

Use Case — Cloud Removal: Data (Illustration)

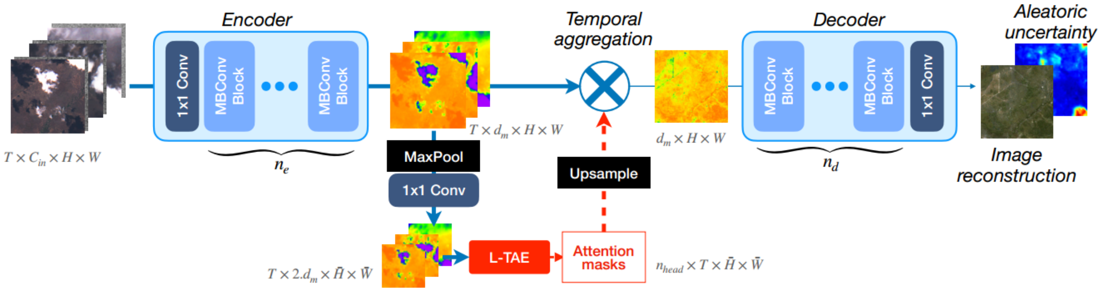


Another NDVI example highlighting spatial variability across fields.

Use Case — Cloud Removal: Methods

- **Models**: UnCRtainTS (mono-temporal) with uncertainty head; U-Net baseline.
- **Inputs**: Concatenate recent S2, S1 channels; optional contextual layers.
- **Training**: Adam (lr=1e−3), batch=6; MAE/MSE objectives; NDVI-loss variant for NDVI accuracy.
- **Inference**: Predict gap-free reflectance/NDVI and uncertainty for low-confidence regions.

Use Case — Cloud Removal: Architecture

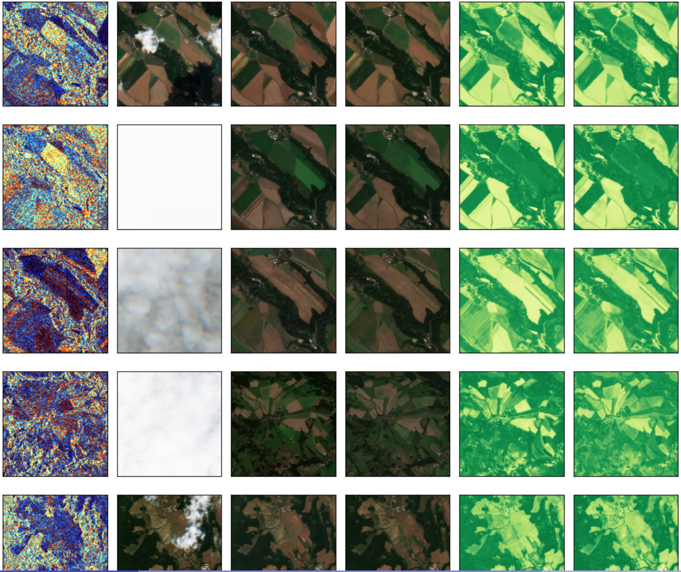


UnCRtainTS architecture with uncertainty estimation head.

Use Case — Cloud Removal: Results

- **Accuracy**: UnCRtainTS MAE 0.0255 vs. U-Net 0.0278, mosaicking 0.0386, recent-cloudless 0.0371.
- **Inputs**: S1+S2 > S2-only > S1-only; both are crucial.
- **NDVI**: Optimizing NDVI directly improves NDVI MAE (e.g., 0.0631 vs. 0.0721).
- **Ablations**: Soil/LSA context gave marginal gains in this setup.

Use Case — Cloud Removal: Training/Validation



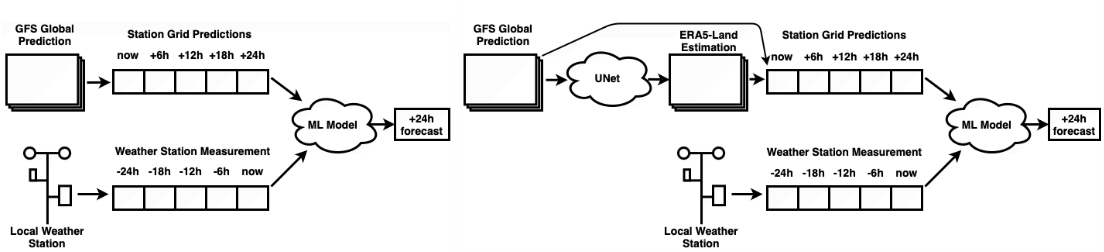
Use Case — Cloud Removal: Conclusion

- Gap-free NDVI series across cloudy seasons enable timely insights for agriculture.
- Combining S1+S2 is key; uncertainty maps support cautious decision-making.
- Next: multi-temporal UnCRtainTS, richer weather features, and field validation/ops deployment.

Use Case — Local Weather: Motivation

- **Why**: Global forecasts (e.g., GFS) are too coarse; microclimates drive site-specific conditions.
- **Impact**: Farmers need accurate 24h station-level forecasts for irrigation, spraying, and harvest planning.
- **Objective**: Improve local 24h forecasts by combining station history, GFS, and ERA5-Land-derived features.

Use Case — Local Weather: Pipeline



Data flow from GFS and stations to local predictions; ERA5-Land features estimated via U-Net.

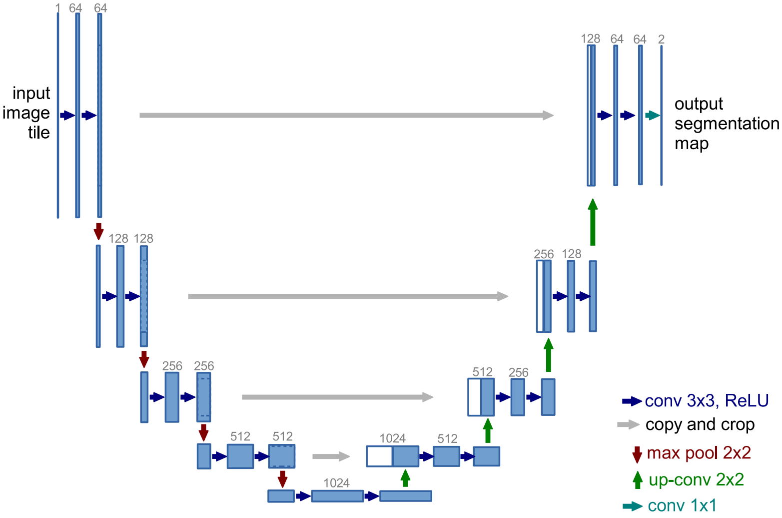
Use Case — Local Weather: Data

- **Stations**: HadISD subset — 27 stations in/near Czech Republic; hourly variables (T, dew point, wind, precip, cloud, SLP).
- **Global**: GFS 0.25° grid; rich atmospheric variables and lead times.
- **Reanalysis**: ERA5-Land — used offline; U-Net trained to estimate ERA5-Land-like features from GFS for near-real-time.
- **Splits**: Train 2022; validate 2023; station time windows + matched GFS rectangles.

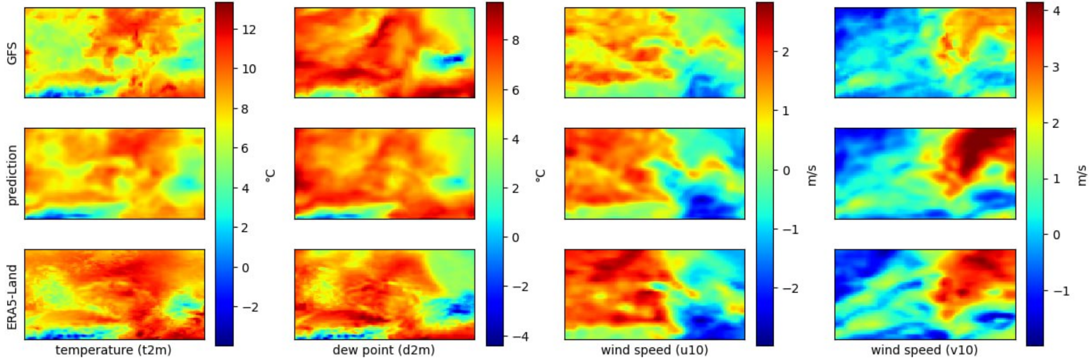
Use Case — Local Weather: Methods

- **Stage 1**: U-Net super-resolution maps GFS $\rightarrow$ ERA5-Land (T, dew point, wind u/v) to produce higher-quality features.
- **Stage 2**: Station-level predictors (CatBoost, MLP, LSTM seq2seq) using station history + GFS + generated ERA5-Land features.
- **Targets**: 24h-ahead T, dew point, wind speed at 2 m.

Use Case — Local Weather: U-Net Architecture



Use Case — Local Weather: U-Net Example

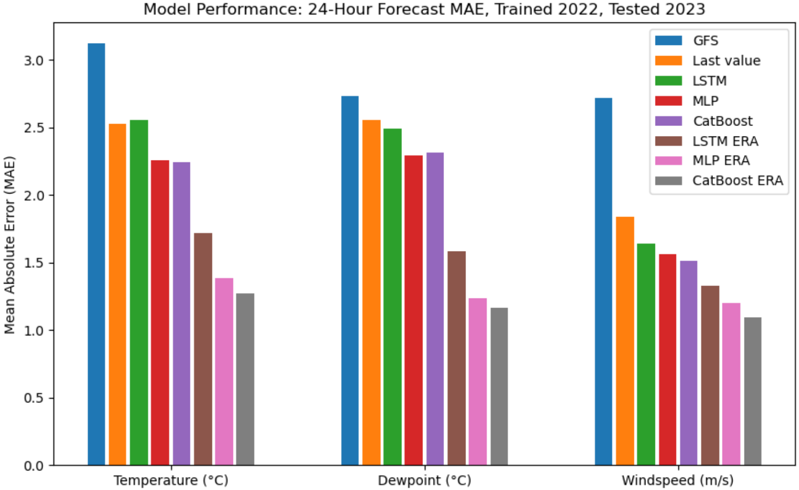


Example mapping: GFS (top) → U-Net output (middle) → ERA5-Land (bottom).

Use Case — Local Weather: Results

- **Accuracy**: CatBoost achieves $\approx$ 1.06–1.07 $^{\circ}C$ MAE (T) and $\approx$ 1.01 m/s (wind), improving over GFS and persistence.
- **ERA5-Land**: Generated features yield slight gains; best overall remains CatBoost.
- **Robustness**: Better diurnal cycles and extremes vs. raw global forecasts.

Use Case — Local Weather: Model Comparison



Comparison of baselines vs. our models across target variables.

Use Case — Local Weather: Conclusion

- Local 24h forecasts substantially improved using station+GFS (+ERA5-Land features when available).
- Deployable pipeline; next: reduce station history needs and scale to ∼200 stations.

Use Case — Climate Superresolution: Motivation

- **Why**: Global NWP (e.g., GFS) is too coarse for field-level agriculture; microclimates matter.
- **Impact**: Operations (frost protection, irrigation, spraying) need accurate 24h local forecasts.
- **Objective**: Superresolve coarse forecasts using station data and learned spatial patterns.

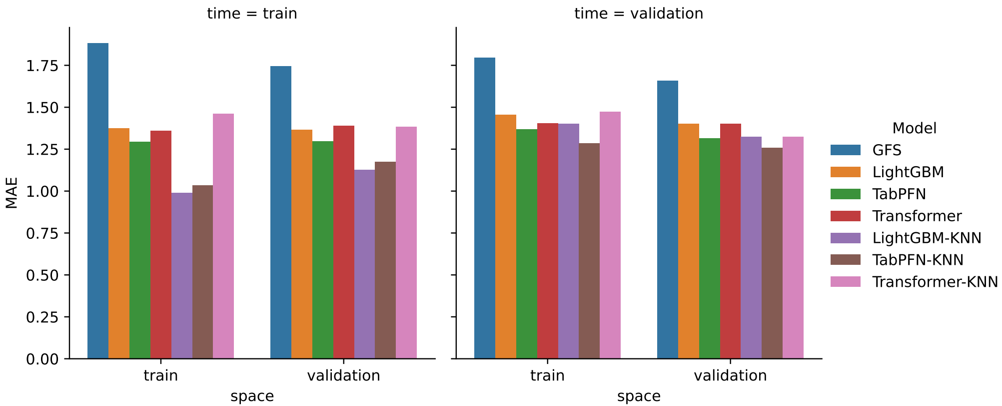
Use Case — Climate Superresolution: Data

- **Global**: GFS 0.25° predictions; multiple variables and lead times.
- **Stations**: High-frequency vineyard stations in S. Moravia (temperature at 15 min).
- **Geography**: Elevation and gradients to capture slope/aspect effects.
- **Splits**: Spatial+temporal splits to test generalization across stations and years.

Use Case — Climate Superresolution: Methods

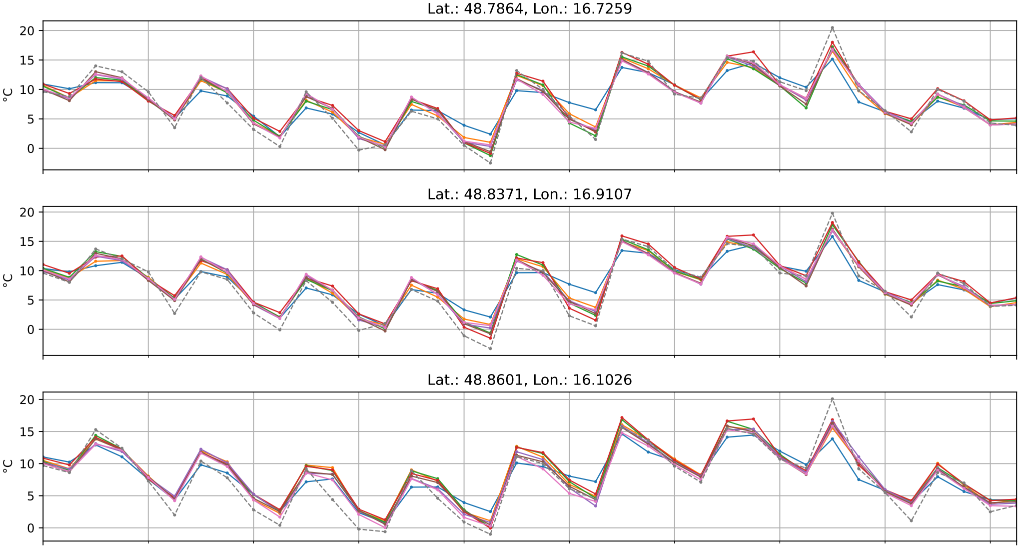
- **Models**: LightGBM, TabPFN, Transformer; KNN hybrid for spatial interpolation.
- **Targets**: 24h-ahead 2 m temperature; bias-correct GFS to local conditions.
- **Evaluation**: MAE across space/time splits; qualitative checks of spatial realism.

Climate Superresolution: MAE



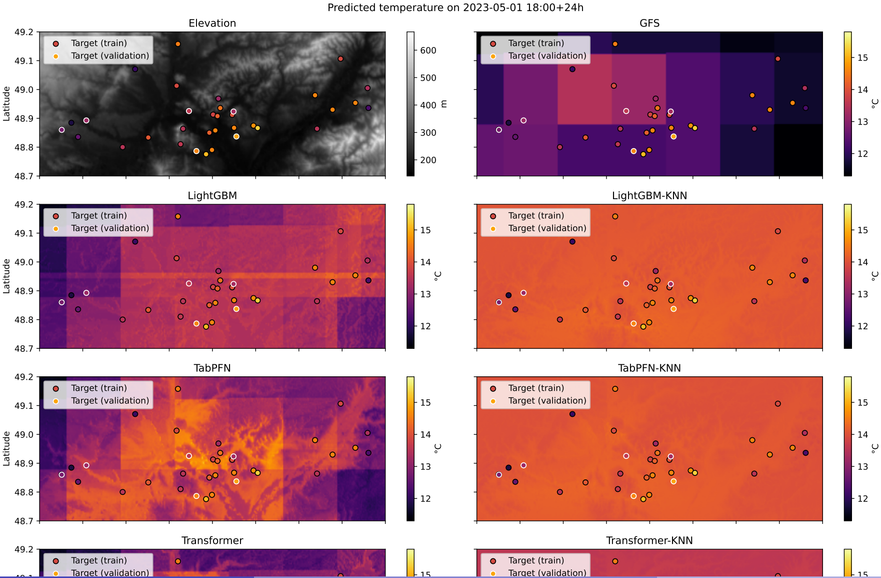
MAE of 24h temperature across space/time splits and models.

Climate Superresolution: Time Series

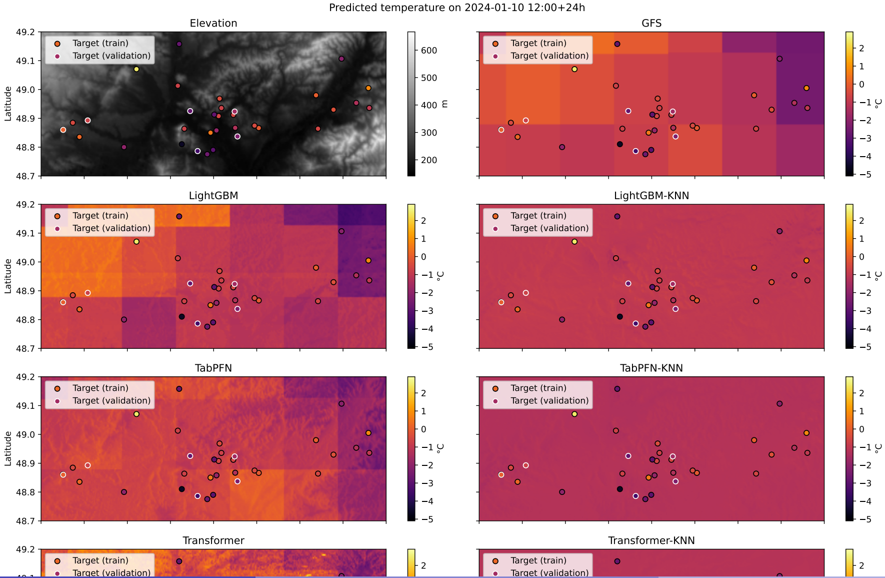


Lat.: 48.8928, Lon.: 16.2164

Climate Superresolution: Spring Example



Climate Superresolution: Winter Example



Use Case — Climate Superresolution: Results

- **Accuracy**: Up to ∼24% MAE reduction vs. GFS; TabPFN-KNN best on hardest split.
- **Realism**: Base models preserve topographic patterns better than KNN-smoothed fields.
- **Ops**: Low compute cost vs. physics downscaling; practical for deployment.

Use Case — Climate Superresolution: Conclusion

- Superresolution bridges global forecasts and field needs; improves local 24h temperature.
- Future: satellite LST for broader validation; refine trade-off between MAE and spatial realism.

Contact

- DataLab @ FIT CTU, Prague — collaborations welcome.
- Pavel Kordík: `https://kordikp.github.io` Email: `pavel.kordik@fit.cvut.cz`
- Web: `https://fit.cvut.cz` Email: `datalab@fit.cvut.cz`