

Predicting Central European Summer Heatwaves with Machine Learning

E. WEIRICH BENET,[1] M. PYRINA,[1] B. JIMENEZ ESTEVE,[1] E. FRAENKEL,[2] J. COHEN,[3,4] and D. I. V. DOMEISEN [1,5]

1. Institute for atmospheric and climate science, ETH Zürich, Switzerland
2. Department of Biological Engineering, MIT, USA
3. Atmospheric and Environmental Research (AER), USA
4. Department of Civil and Environmental Engineering, MIT, USA
5. Institute of Earth Surface Dynamics, Université de Lausanne, Switzerland

SCREEN CAPTURE WELCOME

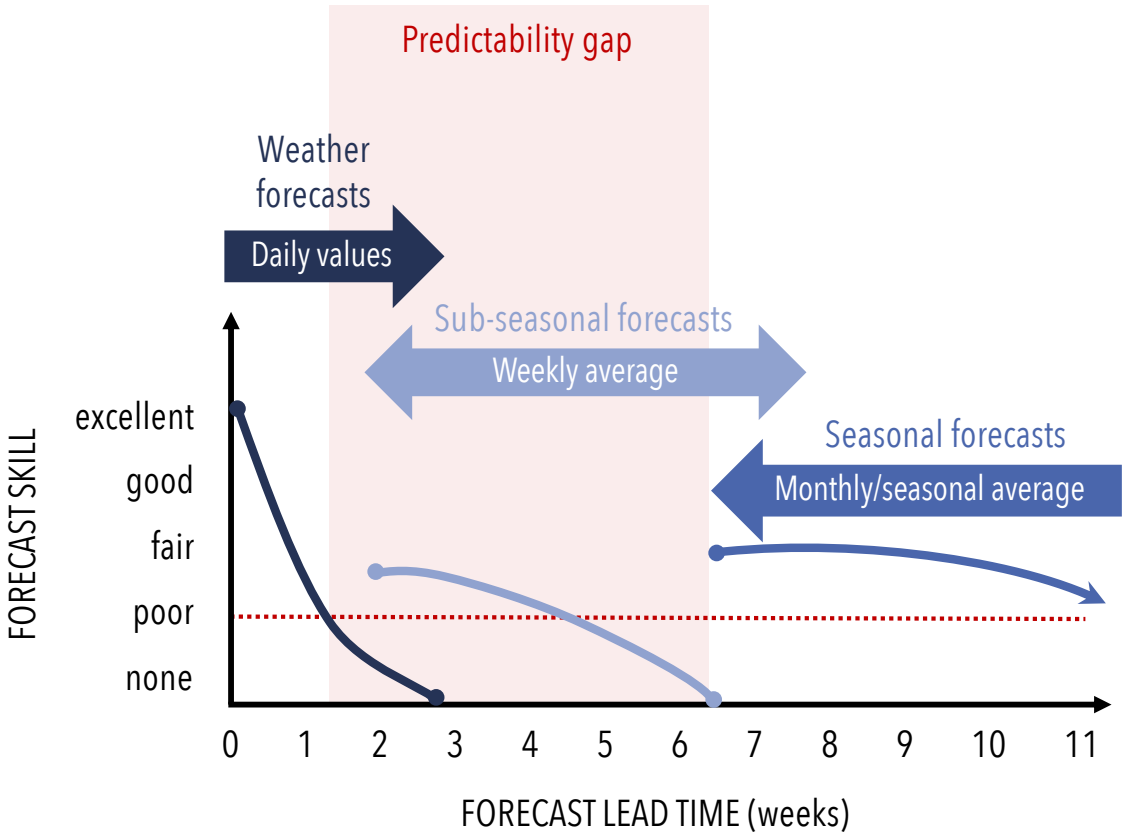


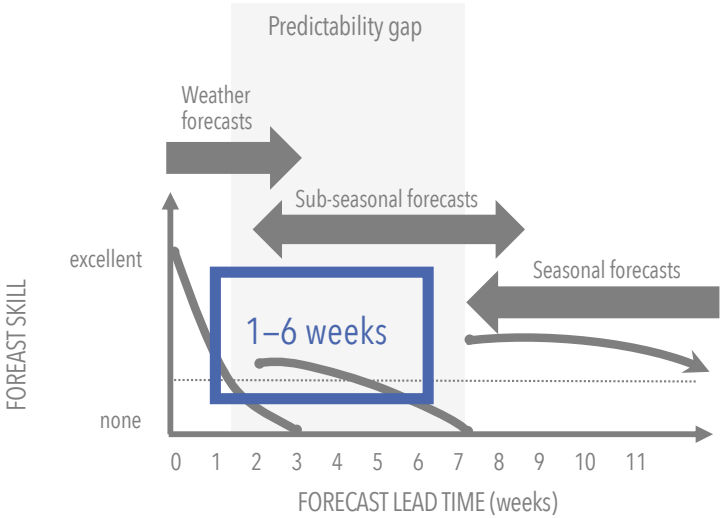
Figure adapted by E. Gawthrop from an original figure by T. Barnston (Columbia Climate School)

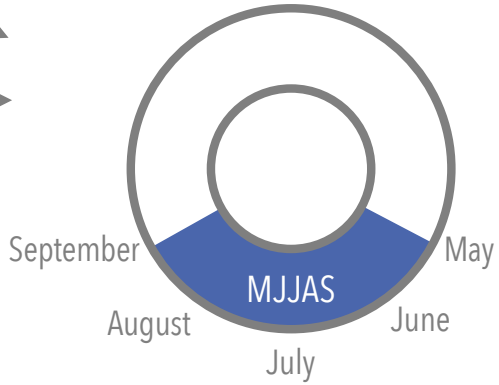
Sub-seasonal predictability gap

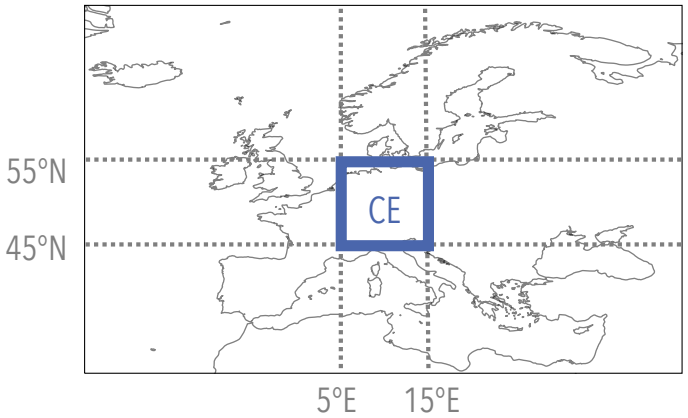
- Weather forecasts' accuracy decays fast with lead time
- Seasonal forecasts available
- Predictability gap
- Sub-seasonal forecasts are not accurate

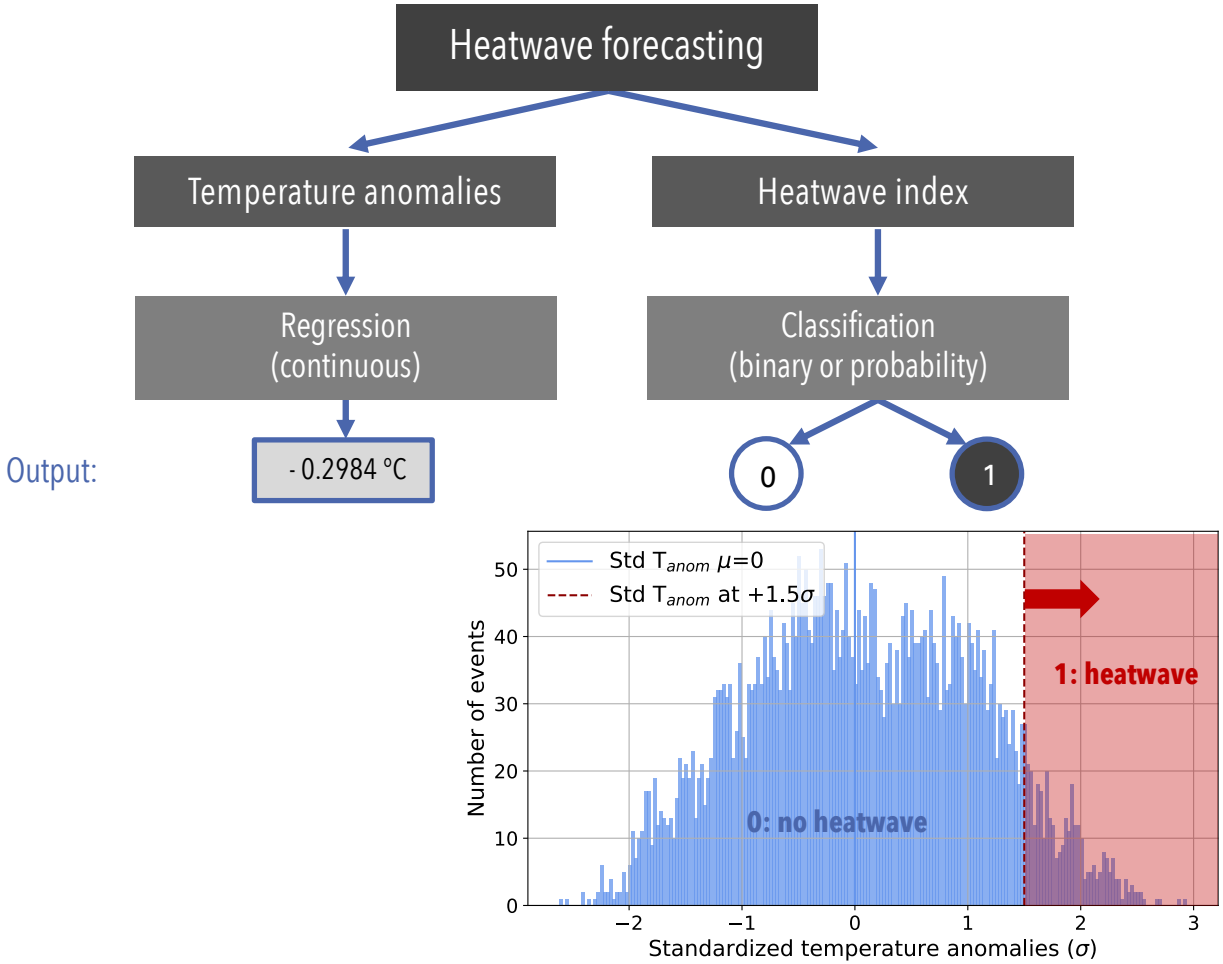
Research question

Can the sub-seasonal forecasting accuracy of summer heatwaves in Central Europe (CE) be improved by using Machine Learning methods?









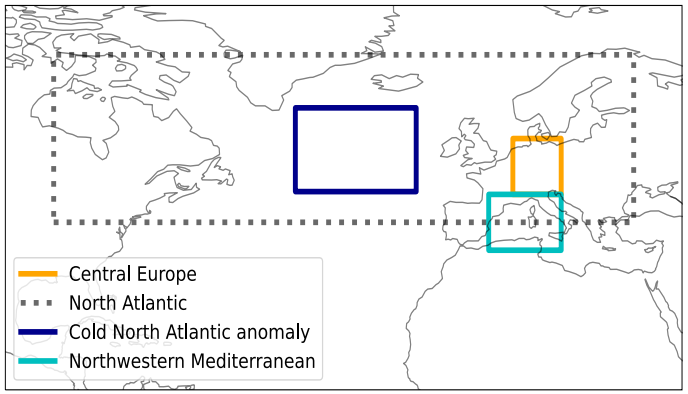
Target definition

1. Forecast temperature anomalies w.r.t. climatology
2. Forecast heatwave events → binary weekly heatwave index
   - 0: no heatwave
   - 1: heatwave

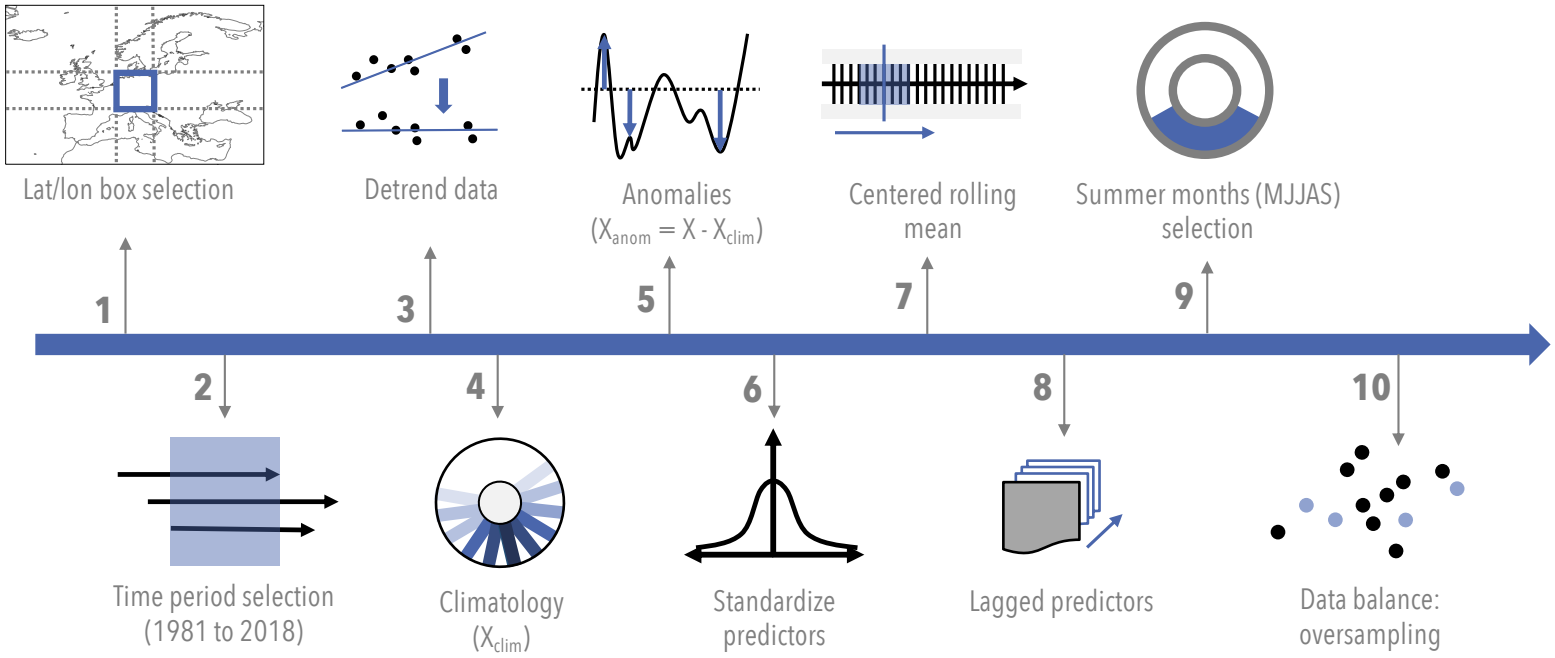
σ: standard deviation from the μ: mean

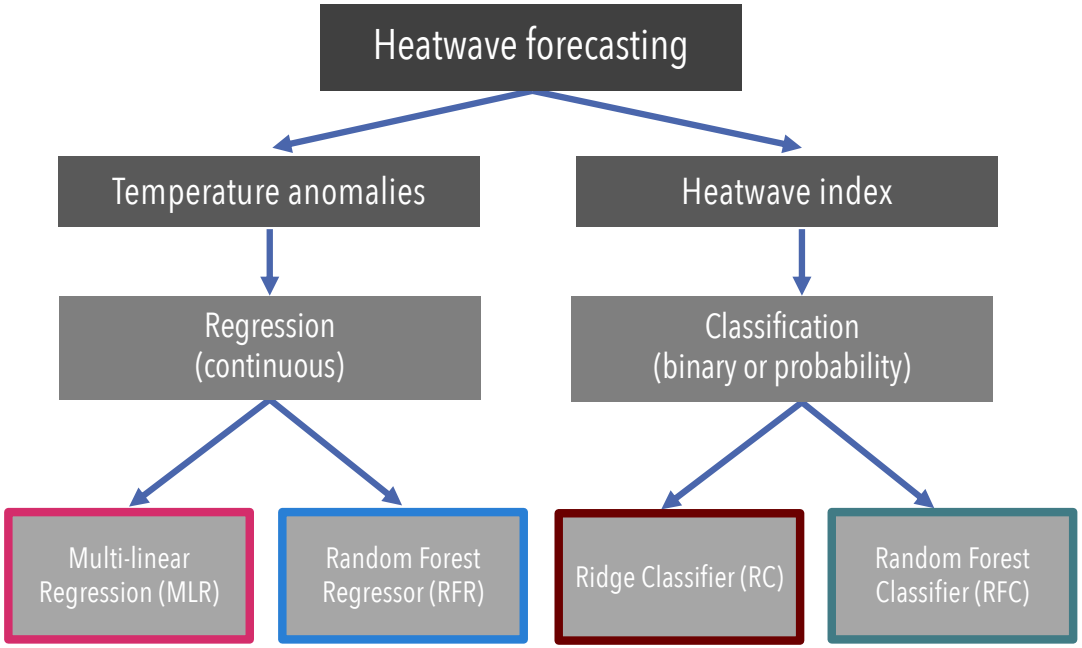
Temperature predictors

Predictor	Level	Box	Latitude	Longitude	Method
Temperature	2 m air	Central Europe	45°N—55°N	5°E—15°E	average
Geopotential	500 hPa	Central Europe	45°N—55°N	5°E—15°E	average
Precipitation	surface	Central Europe	45°N—55°N	5°E—15°E	average
Soil moisture	0—28 cm underground	Central Europe	45°N—55°N	5°E—15°E	average
Summer East Atlantic pattern	500 hPa	North Atlantic	40°N—70°N	90°W—30°E	PCA
Sea surface temperature (SST)	sea level	Cold North Atlantic anomaly	45°N—60°N	15°W—40°W	average
SST	sea level	Northwestern Mediterranean	35°N—45°N	0°—15°E	average



Data pre-processing pipeline





Machine Learning Models

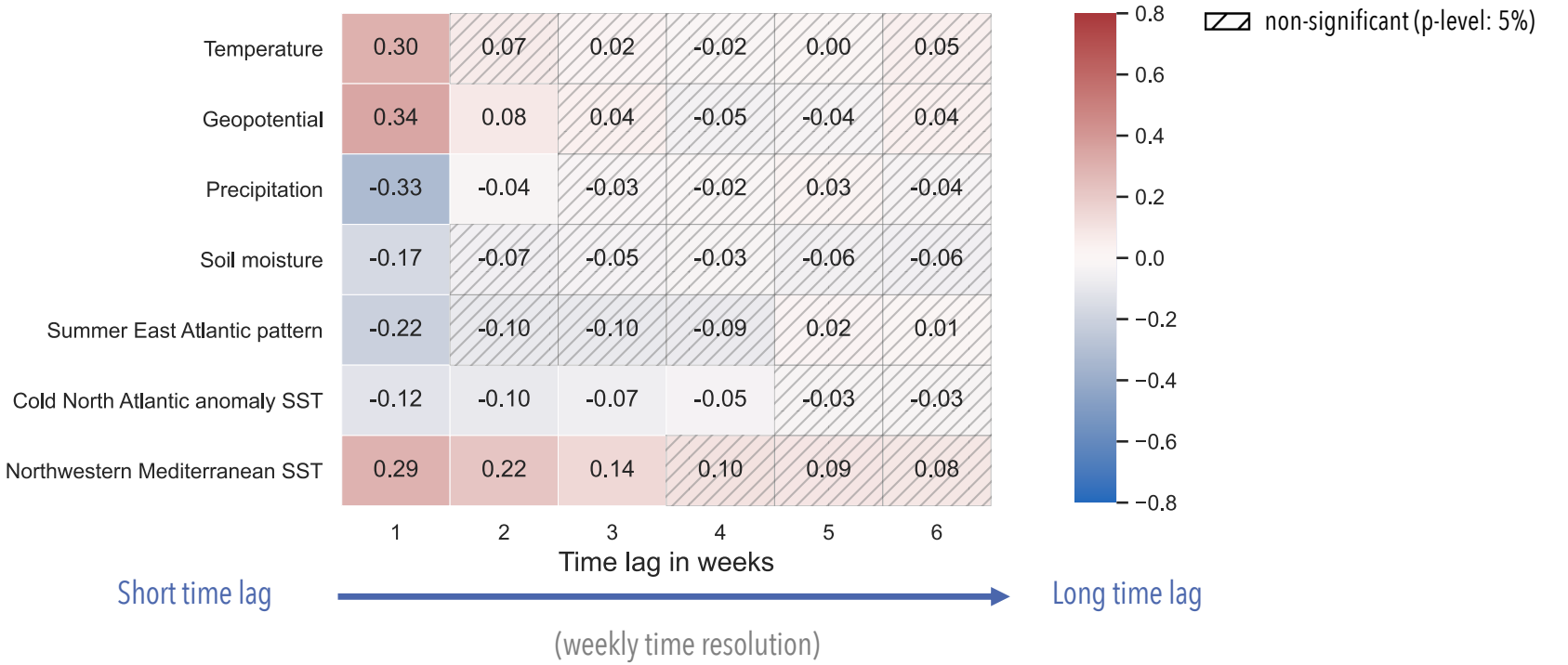
Two problems:

1. Regression
2. Classification

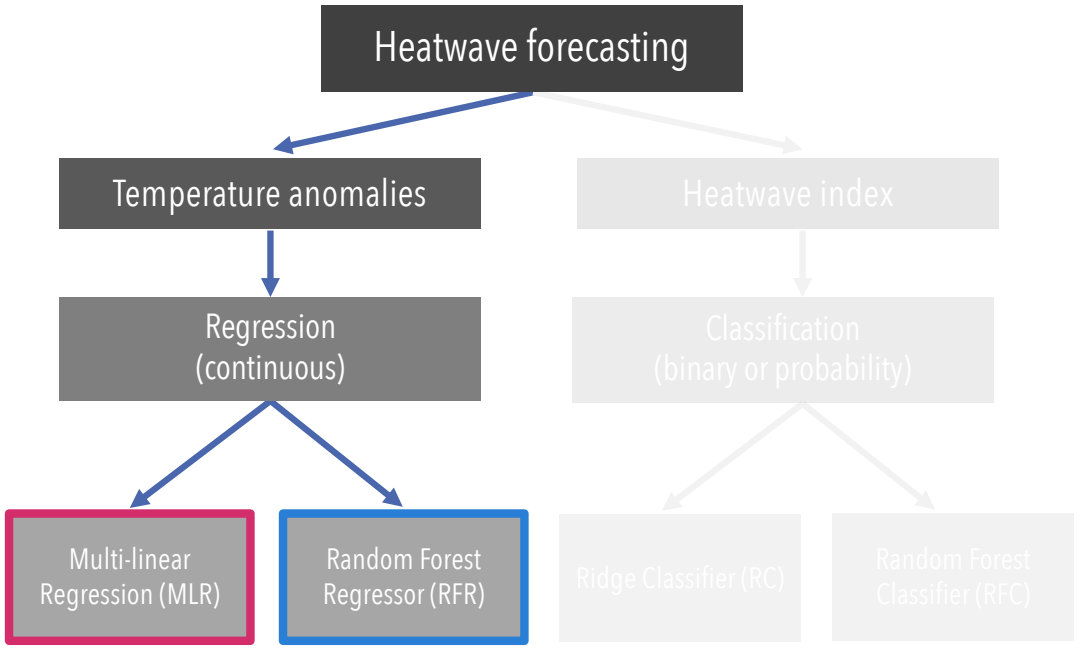
Four models:

- Two linear models
- Two random forests

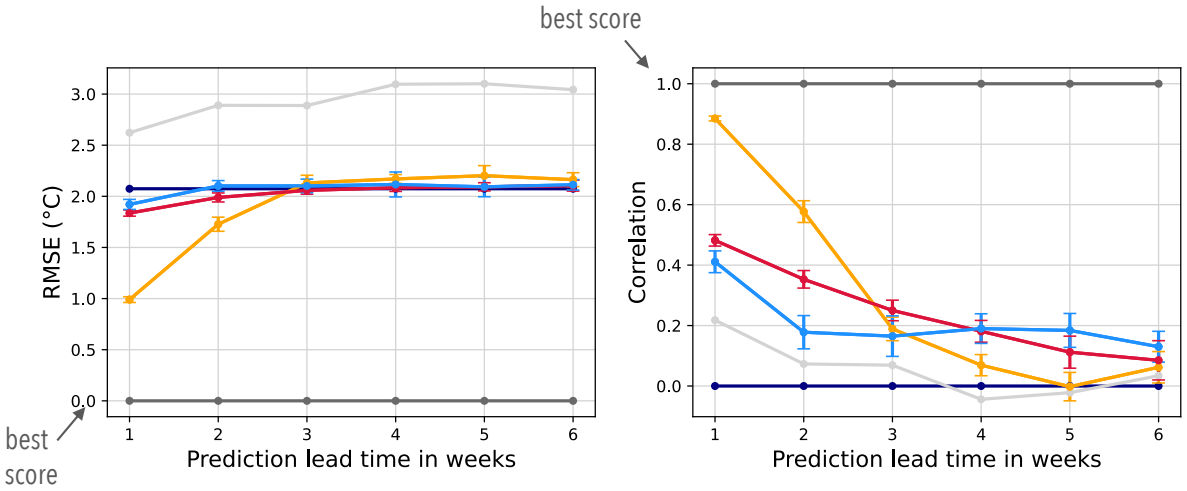
Linear Lagged Correlation Analysis



Regression forecasts

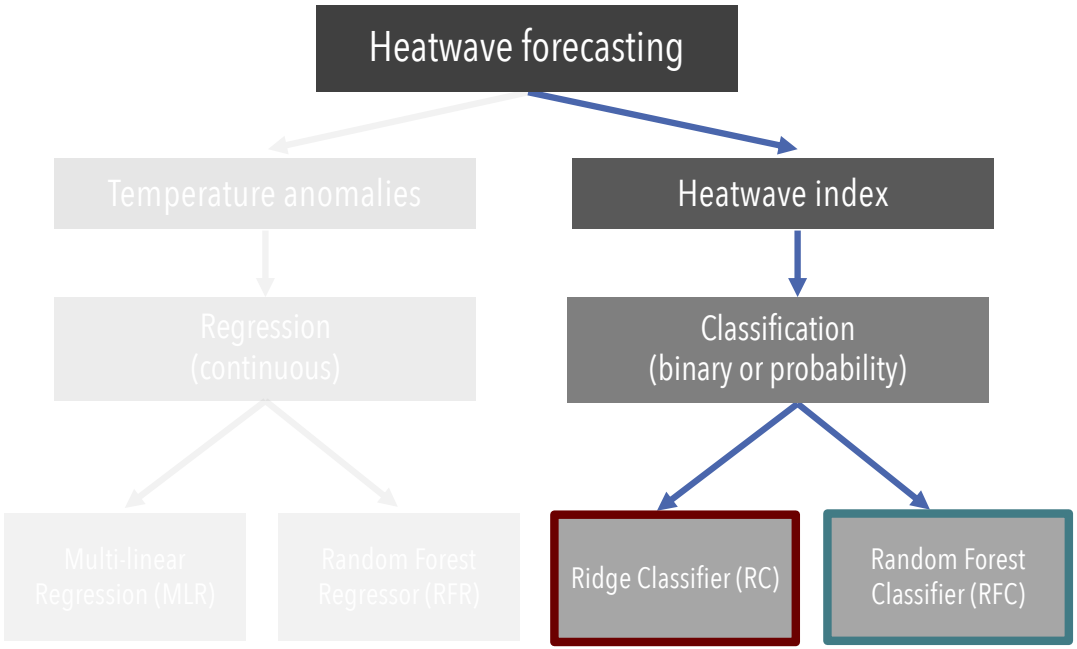


Regression forecasts' skill

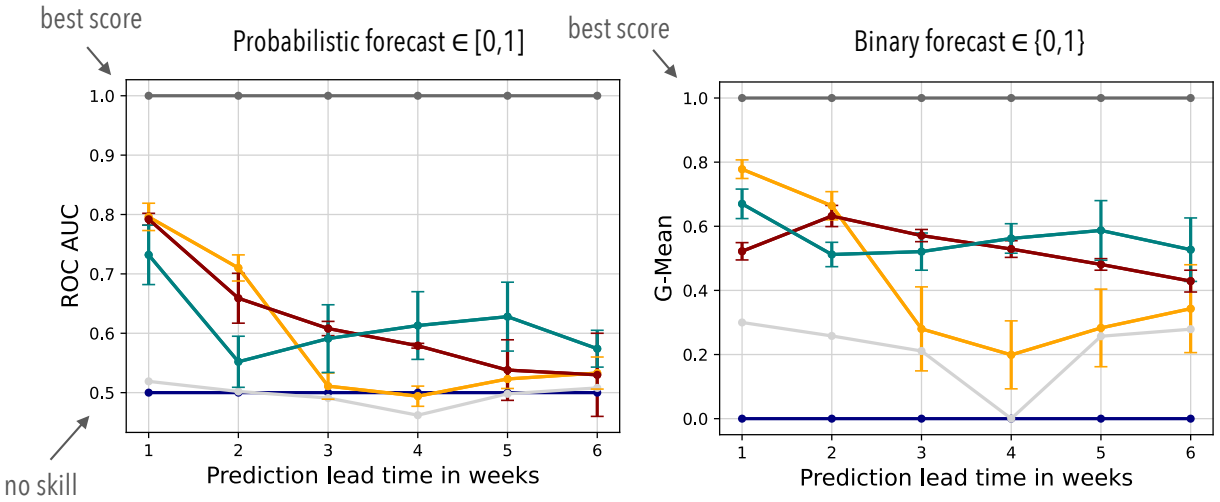


- Both ML models outperform persistence and climatology forecasts at all lead times
- ECMWF dominates at short lead times (up to two weeks)
- MLR dominates at three weeks lead time
- RFR dominates at long lead times (more than three weeks)

Classification forecasts



Classification forecasts' skill



RC: Ridge Classifier
RFC: Random Forest Classifier
} machine learning models

Ground Truth
Climatology
Persistence
ECMWF
} baselines for comparison

- All four ML models outperform persistence and climatology for all lead times
- ECMWF dominates for short lead times (up to two weeks)
- ML models outperform ECMWF for lead times higher than two weeks

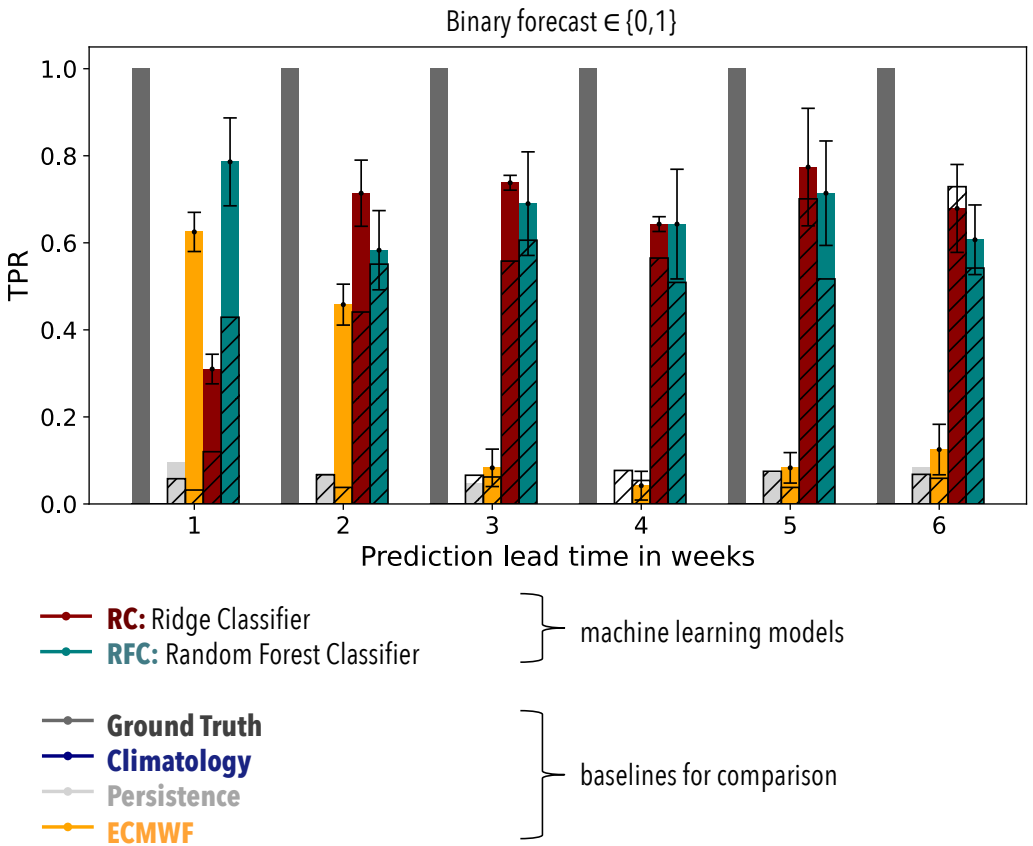
ROC AUC: Receiver Operating Characteristic Area Under Curve

Geometric Mean: $\text{G-Mean} = \sqrt{\text{TPR}\ (1\ -\text{FPR})}$

TPR: True Positive Rate = TP/(TP+FN)

FPR: False Positive Rate = FP/(FP+TN)

Classification forecasts' skill



- Useful ⟺ TPR > FPR
- For all lead times, at least one ML model provides a useful forecast
- ML models have both higher TPR and FPR than ECMWF

TPR: True Positive Rate = TP/(TP+FN)
FPR: False Positive Rate = FP/(FP+TN)

Conclusions

Can the sub-seasonal forecasting accuracy of summer heatwaves in Central Europe be improved by using Machine Learning (ML) methods?

Yes, but only at lead times higher than two weeks.

At short lead times (1-2 weeks):

- ECMWF more accurate than ML models
- Linear models more accurate than random forests

At longer lead times (3-6 weeks):

- ML models more accurate than ECMWF
- Random forests more accurate than linear models



Thank you for your attention!

Work under review by AIES (AMS):

Sub-seasonal Prediction of Central European Summer Heatwaves with Linear and Random Forest Machine Learning Models, *Weirich Benet et al.*

Earth arXiv pre-print:



LinkedIn:

