Deutsches Geodätisches Forschungsinstitut (DGFI-TUM)
Technische Universität München

EGU22-5408

Machine Learning Approach for Forecasting Space Weather Effects in the Ionosphere with Uncertainty Quantification

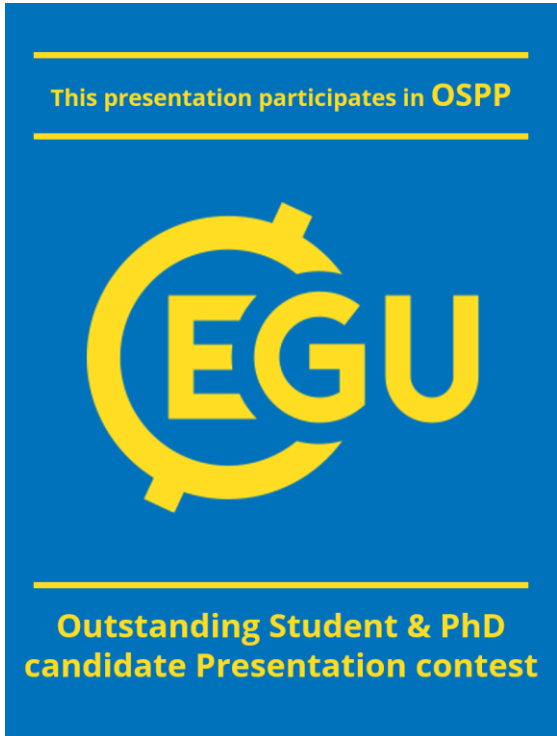
Randa Natras[1], Benedikt Soja[2], Michael Schmidt[1], Marie Dominique[3], Ayse Türkmen[1]

[1]Deutsches Geodätisches Forschungsinstitut der Technischen Universität München (DGFI-TUM), School of Engineering and Design, Technical University of Munich, Munich, Germany
[2]Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland
[3]Solar-Terrestrial Center of Excellence, SIDC, Royal Observatory of Belgium, Brussels, Belgium

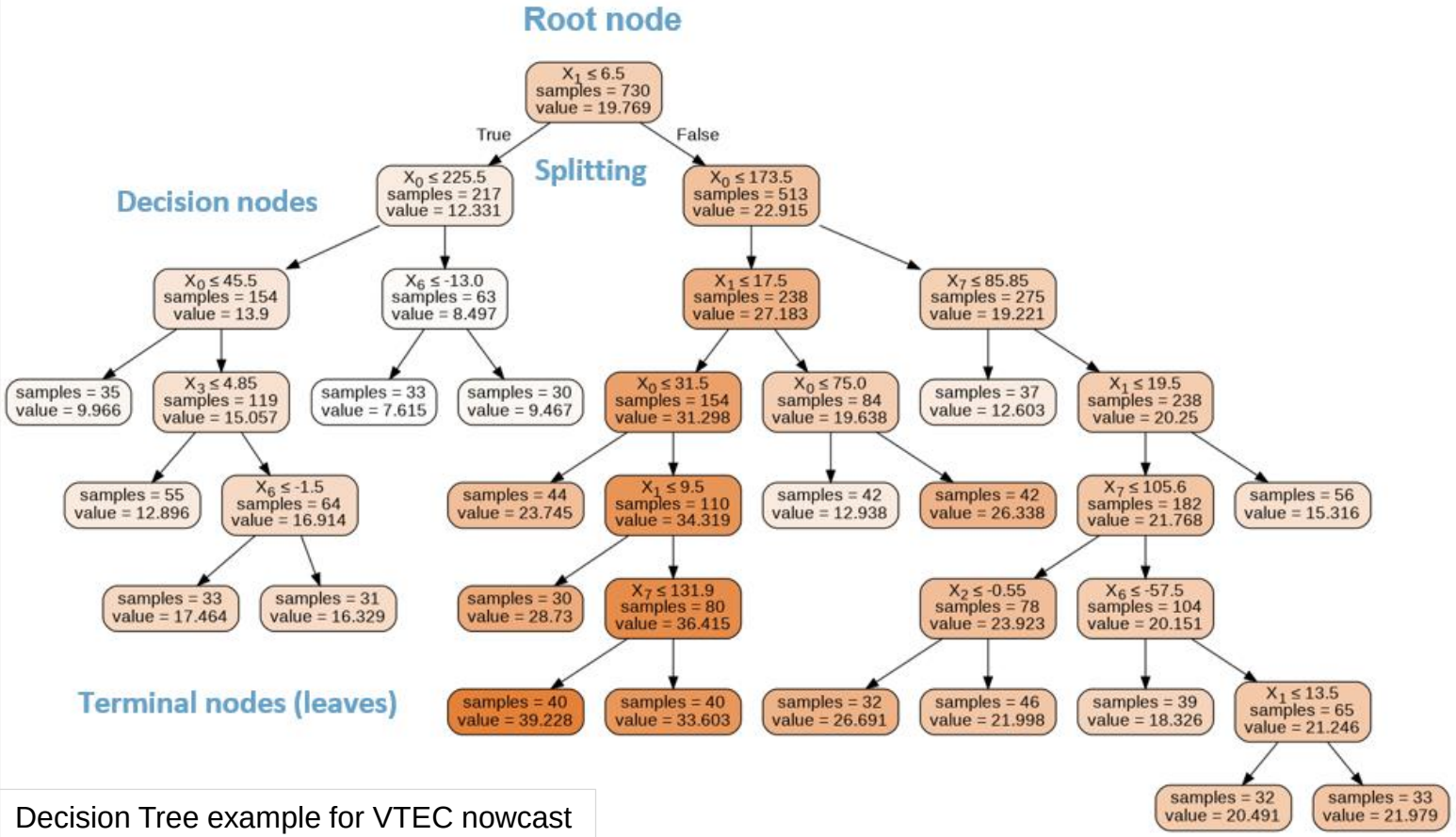
randa.natras@tum.de

European Geosciences Union General Assembly 2022 (EGU22)

Vienna, May 27, 2022



Learning algorithm: Decision Tree

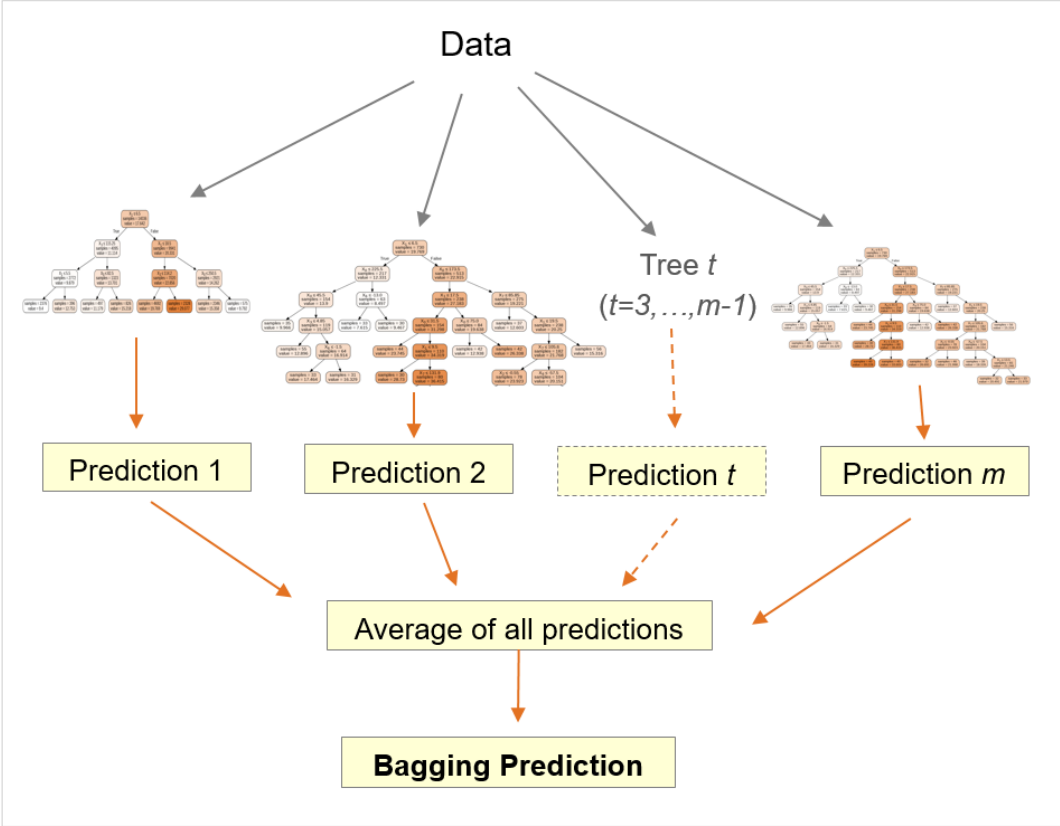


Decision Tree example for VTEC nowcast

- **Root node:** contains the entire dataset and further is divided into two subnodes
- **Splitting:** process of dividing a node into two subnodes by calculating reduction in variance
- **Decision node:** a subnode, which splits into fruther sub-nodes
- **Leaf / Terminal node:** nodes that do not split

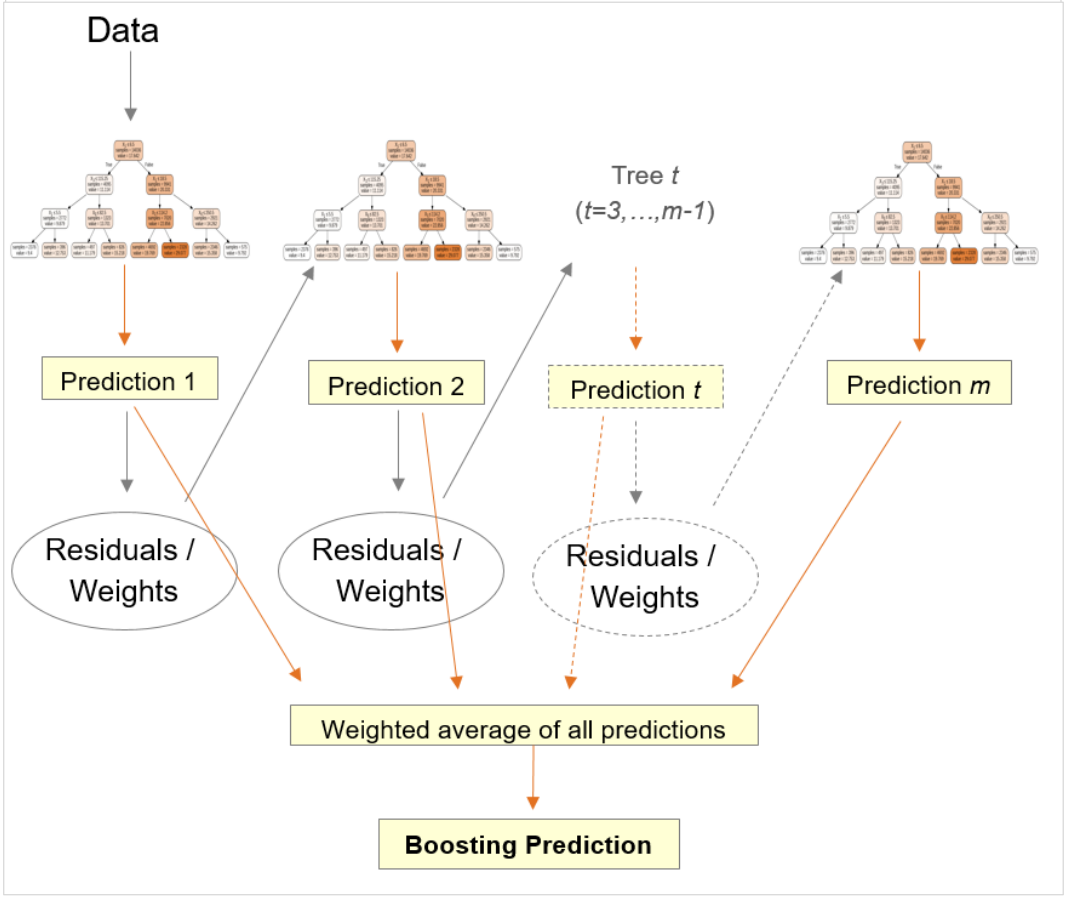
Final outcome: the average VTEC value in the particular leaf node.

Tree-based learning



- **Bagging** (Parallel Ensemble Learning):
  - Random Forest (multiple randomized trees)

- **Boosting** (Sequential Ensemble Learning)**:**
  - Adaptive Boosting - AdaBoost (training with weighted observations)
  - Gradient Boosting - GBoost (training with residuals)



1-day VTEC Forecasting, Data (time sampling 1h)

Input data:

Time: t

- Time: Hour of day and Day of year (DOY)
- Sunspot number R (daily)
- Solar radio flux F10.7 (daily)
- Solar wind plasma speed (hourly)
- Bz index (hourly)
- AE index (hourly)
- Dst index (hourly)
- Kp index (3-hour)
- VTEC from GIM CODE (hourly)
  - 10E 70N, 10E 40N, 10E 10N
- Exponential moving average of VTEC over previous 4 days and 30 days
- First time derivative of VTEC
- Second time derivative of VTEC

Output data:

- VTEC
  - 10E 70N,
  - 10E 40N,
  - 10E 10N

Time: t+24h

Data split:

- Training & Cross-validation: 2015 - 2016
  - Time series cross-validation
- Test: 2017

VTEC Nowcasting, Data (time sampling 1h)

Input data:

Time: t
- Time: Hour of day and Day of year (DOY)
- Sunspot number R (daily)
- PROBA2 Lyra (1 min)
- Solar wind plasma speed (hourly)
- Bz index (hourly)
- AE index (hourly)
- Dst index (hourly)
- Kp index (3-hour)

Time: t-1h
- VTEC from GIM CODE (hourly)
  - 10E 70N, 10E 40N, 10E 10N
- Exponential moving average of VTEC over previous 4 days and 30 days

Time: t-1h
- First time derivative of VTEC
- Second time derivative of VTEC

Output data:

- VTEC
  - 10E 70N,
  - 10E 40N,
  - 10E 10N

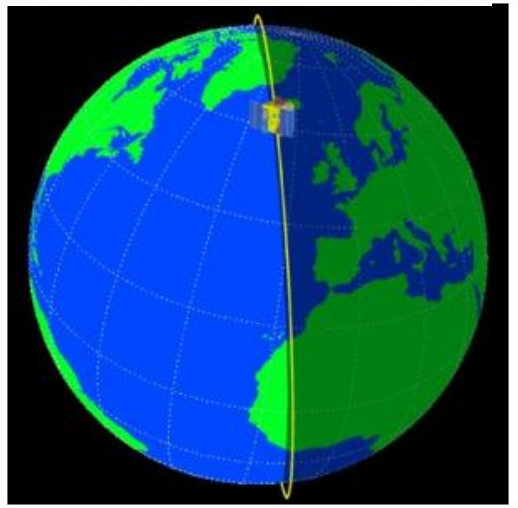
Time: t

Data split:

- Training & Cross-validation: 2015 - 2016
  - Time series cross-validation
- Test: 2017

PROBA2 LYRA Data

- **PROBA2 (PRoject for Onboard Autonomy) satellite:**
  - Solar instruments SWAP and LYRA
  - Orbit height: 720 km
  - Sun-synchronous orbit
- **LYRA data:**
  - Channel 4 data of nominal unit (Zirconium channel: 0.1-20 nm), soft Xray + EUV
  - Level 3 data: calibrated, 1-min sampling
  - For ML models: level 3 data averaged over one hour
- **Data cleaning:**
  - LAR (Large Angle Rotations),
  - SAA (South Atlantic Anomaly),
  - UV occultation (during winter, spacecraft in the Earth‘s shadow),
  - Visible and UV LED on (instrument calibration),
  - Moon in LYRA,
  - Offpoint (satellite manoeuver),
  - Spacecraft anomaly (data missing: May 1-4, 2015, October 29-30, 2016).

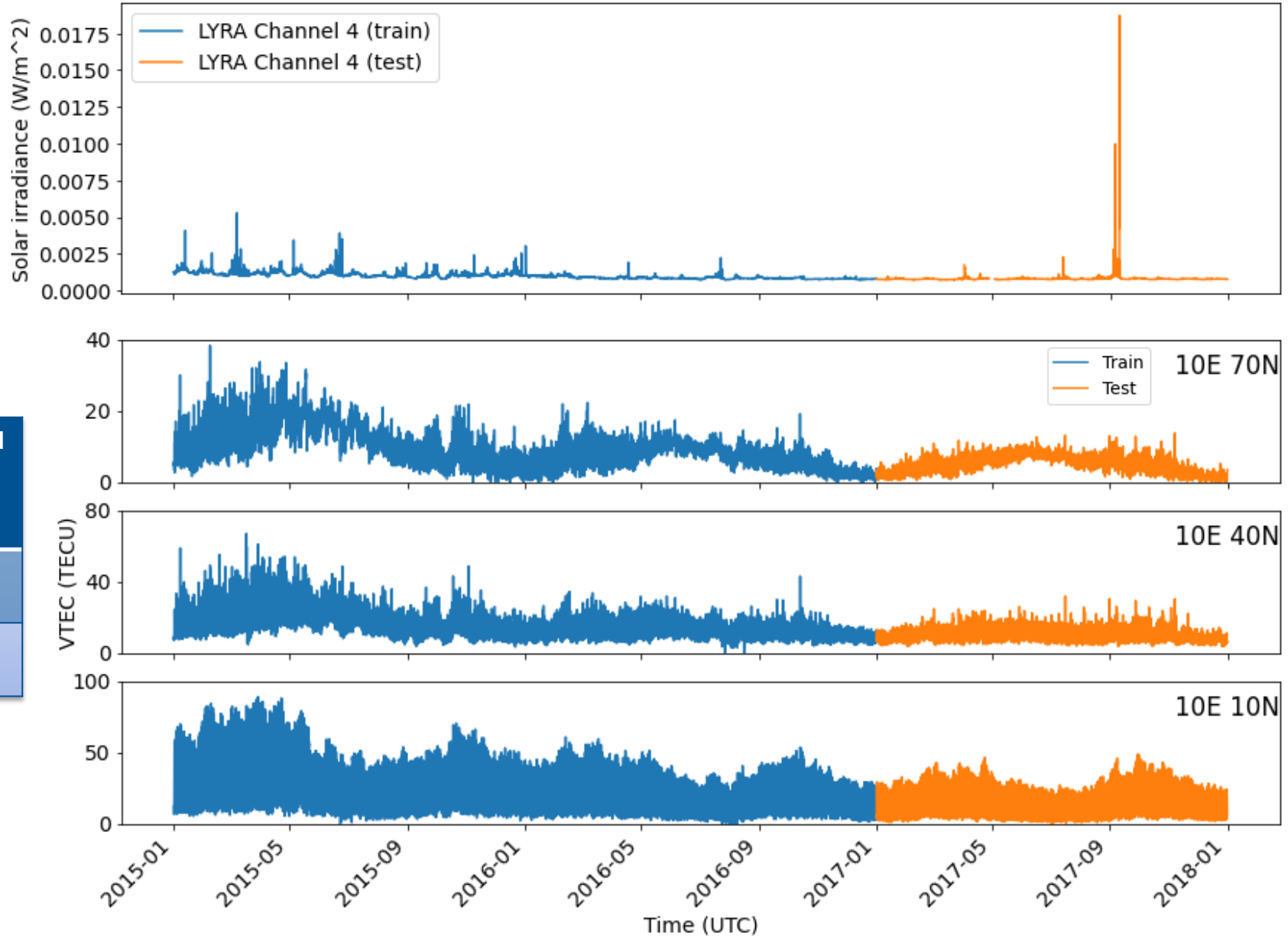


Cartoon of PROBA2's Orbit

Source: https://proba2.sidc.be/about/launch

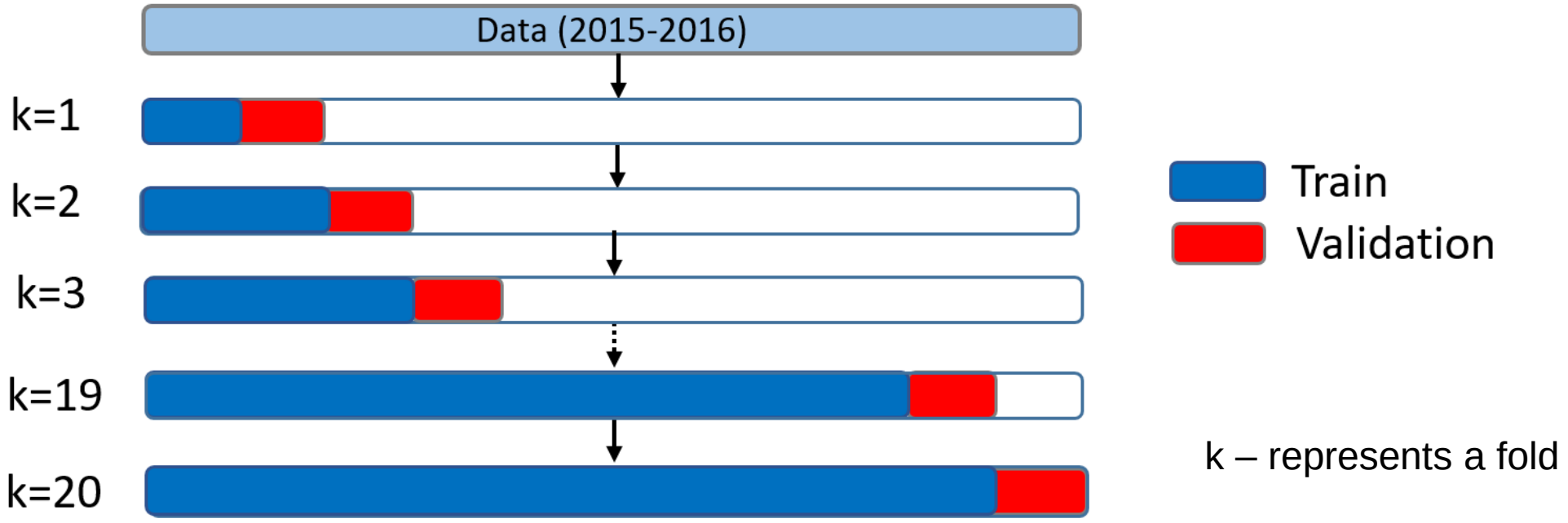
LYRA and VTEC data overview

Data (1h)	Amount of data (2015-2016)	Removed while cleaning (%)
LYRA original	17373	
LYRA cleaned	16239	1134 (6.5%)



Model training and evaluation

- Temporal structure of time series -> cross-validation on a rolling basis
- Evaluate model performance in a robust way
- **Final metrics:** average of RMS from every cross-validation iteration (k-fold).

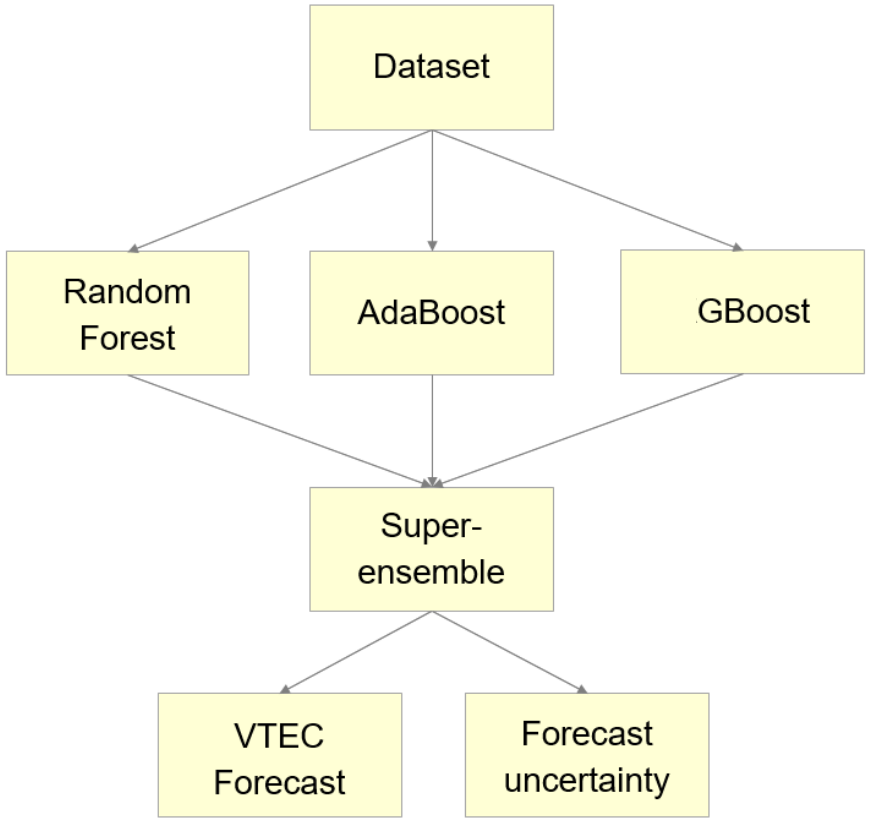


Uncertainty quantification (UQ)

- **Goals:**
  - define the accuracy and precision of VTEC prediction,
  - quantify the level of trust in VTEC prediction,
  - assess the reliability of VTEC prediction.

I. Multi-model and multi-data ensemble method

- VTEC forecast -> ensemble mean
- Forecast uncertainty -> ensemble spread (2σ)
- 3 datasets*:
  1. Original data as input and output
  2. Daily differences as input and output
  3. Original data + daily differences as input, original data as output



II. Confidence interval

- Quantile objective loss function
- Applied for GBoost and 3rd dataset
- Selected quantiles: $\alpha = 0.95$ and $\alpha = 0.05$

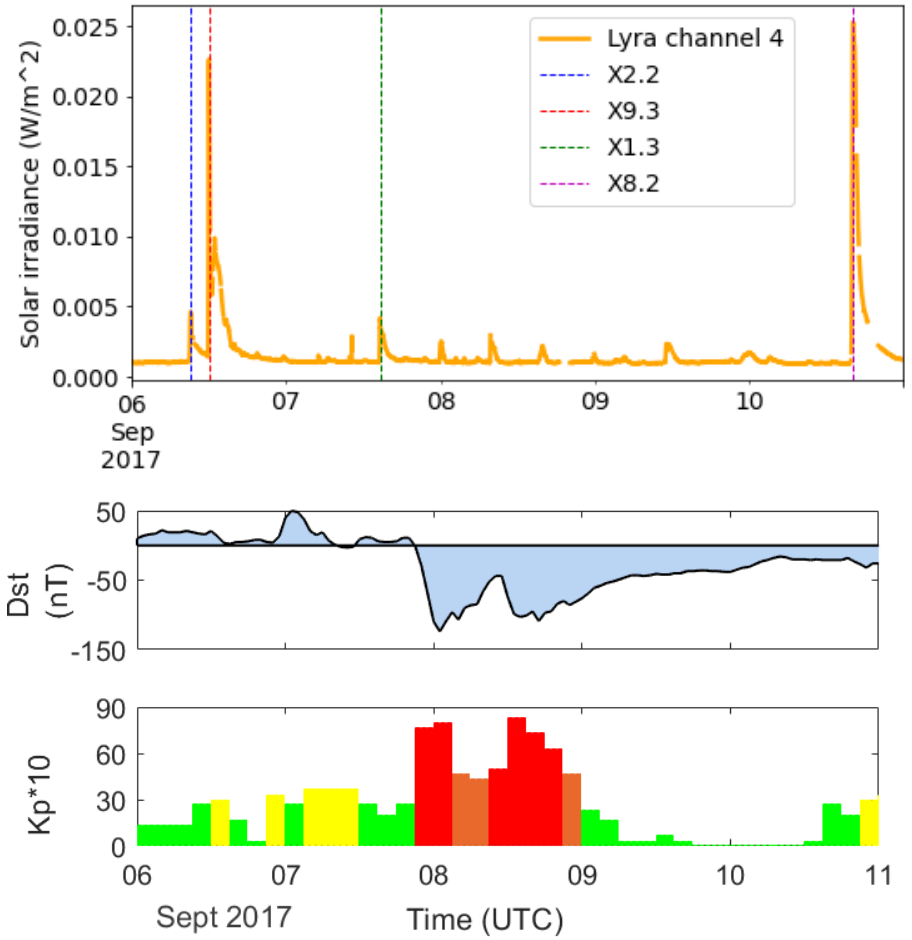
$$\mathcal{L}(e_i|\alpha) = \begin{cases} \alpha \cdot e_i & \text{if } e_i \geq 0, \\ (\alpha - 1) \cdot e_i & \text{if } e_i < 0 \end{cases}$$

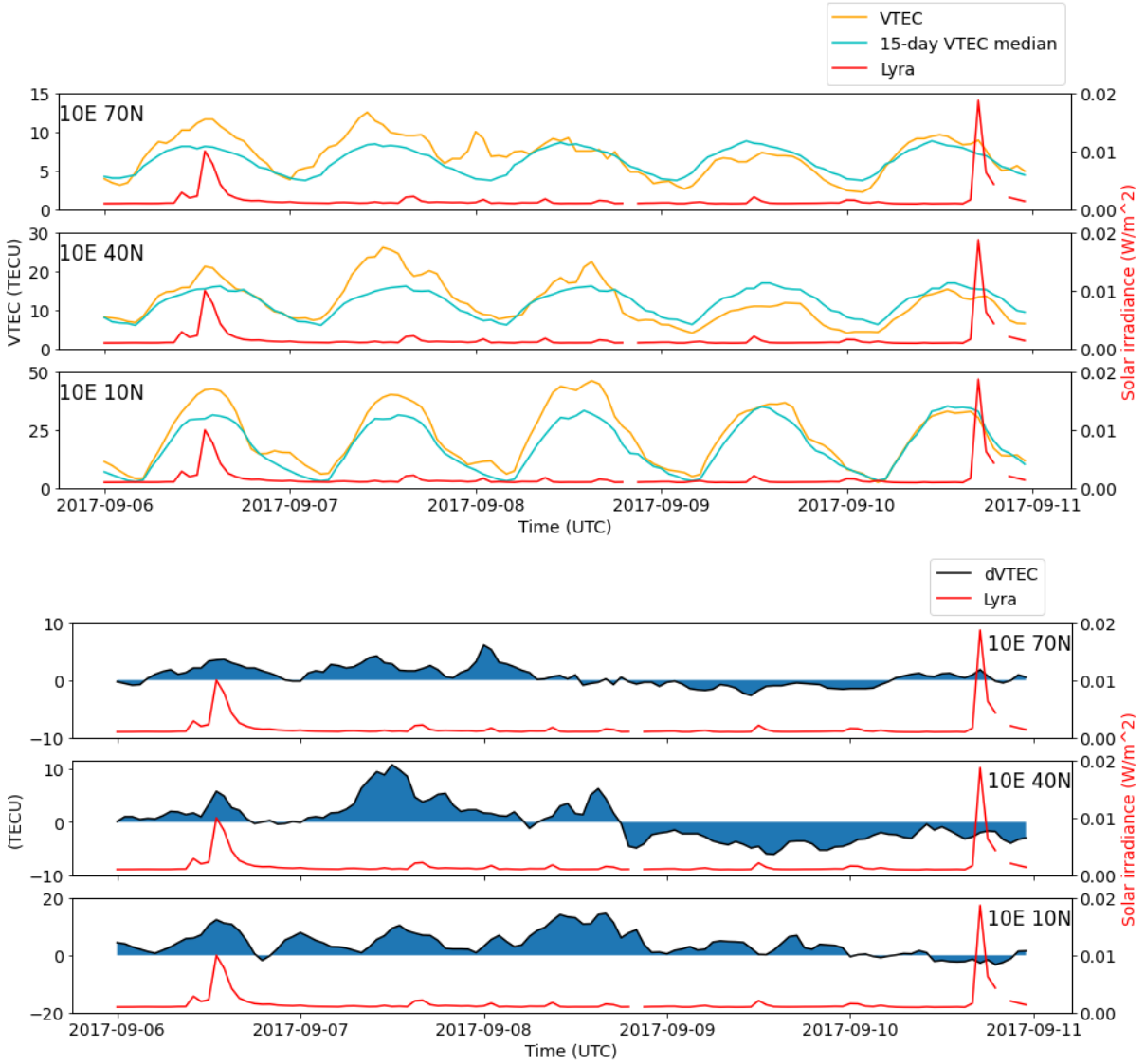
$$e_i = y_i - f(\mathbf{x}_i)$$

$$\mathcal{L}(\mathbf{e}|\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(e_i|\alpha)$$

*Observations were preprocessed / cleaned before training.

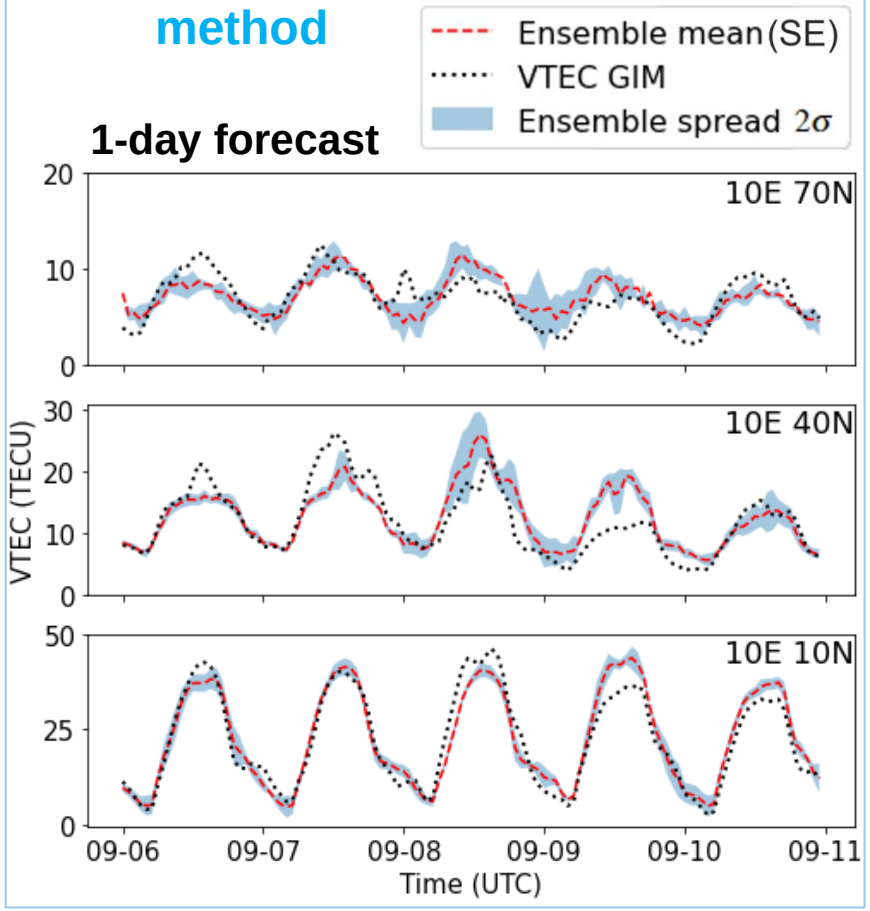
Space weather events: September 6-10, 2017



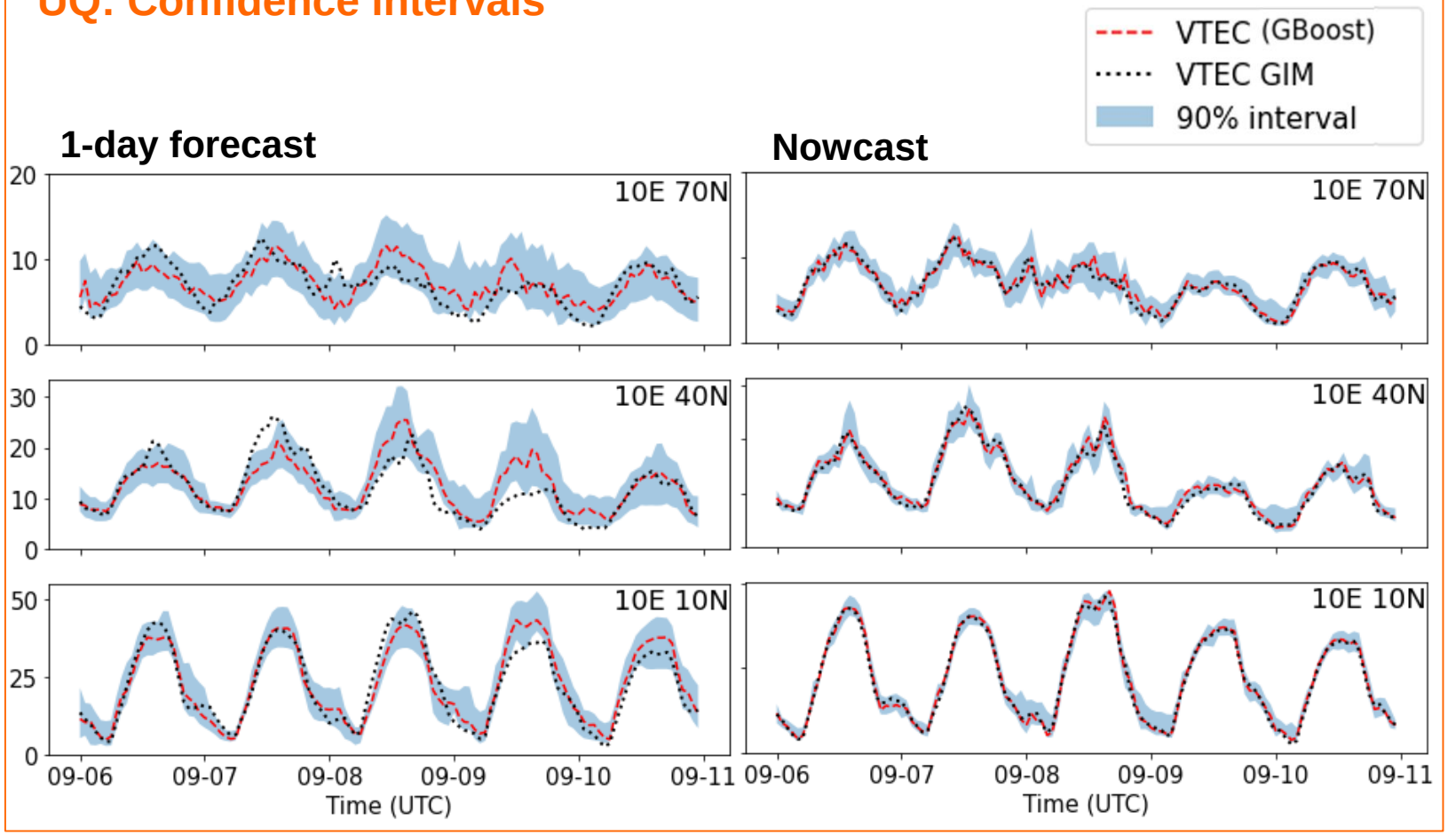


Results for September 2017 space weather events

UQ: Ensemble method



UQ: Confidence intervals



- The uncertainty information defines the reliability and precision of VTEC predictions.
- UQ allows to assess the trustworthiness of predictions.
- Ground-truth VTEC mostly within predicted confidence intervals for space weather events

Thank you for your attention!

randa.natras@tum.de

Scan for the abstract and display material

This presentation participates in OSPP

Outstanding Student & PhD candidate Presentation contest