Deutsches Geodätisches Forschungsinstitut (DGFI-TUM)
Technische Universität München

Interpretable Machine Learning for Ionosphere Forecasting with Uncertainty Quantification

Randa Natras[1], Benedikt Soja[2], Michael Schmidt[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

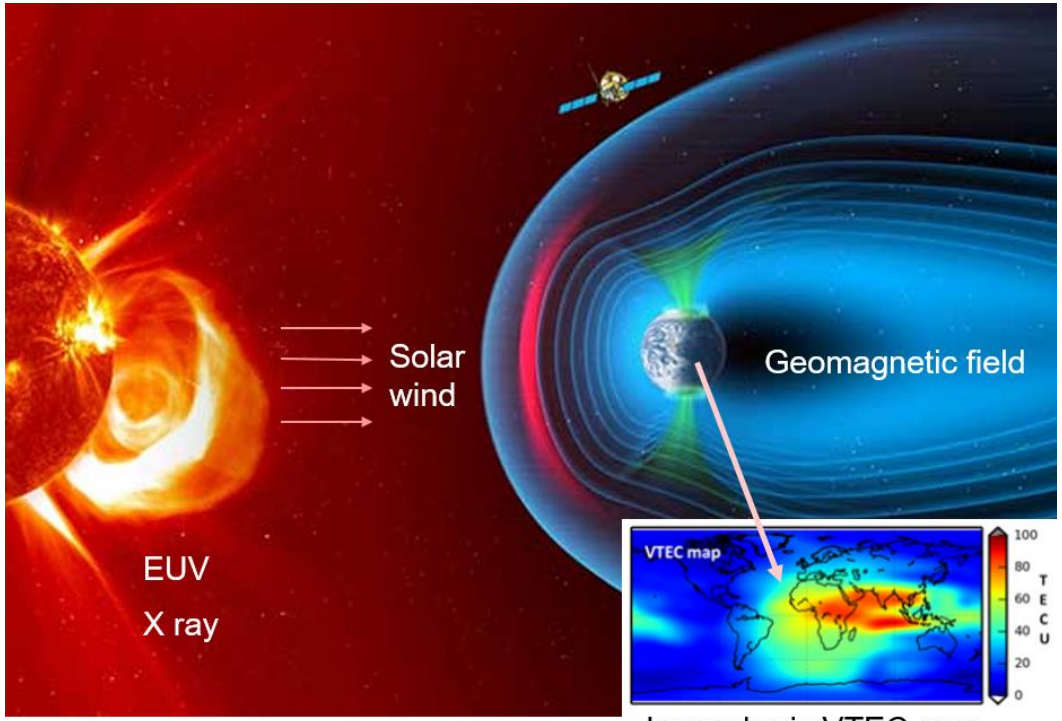
randa.natras@tum.de

1st Workshop on Data Science for GNSS Remote Sensing (D4G), GFZ Potsdam

June 13, 2022

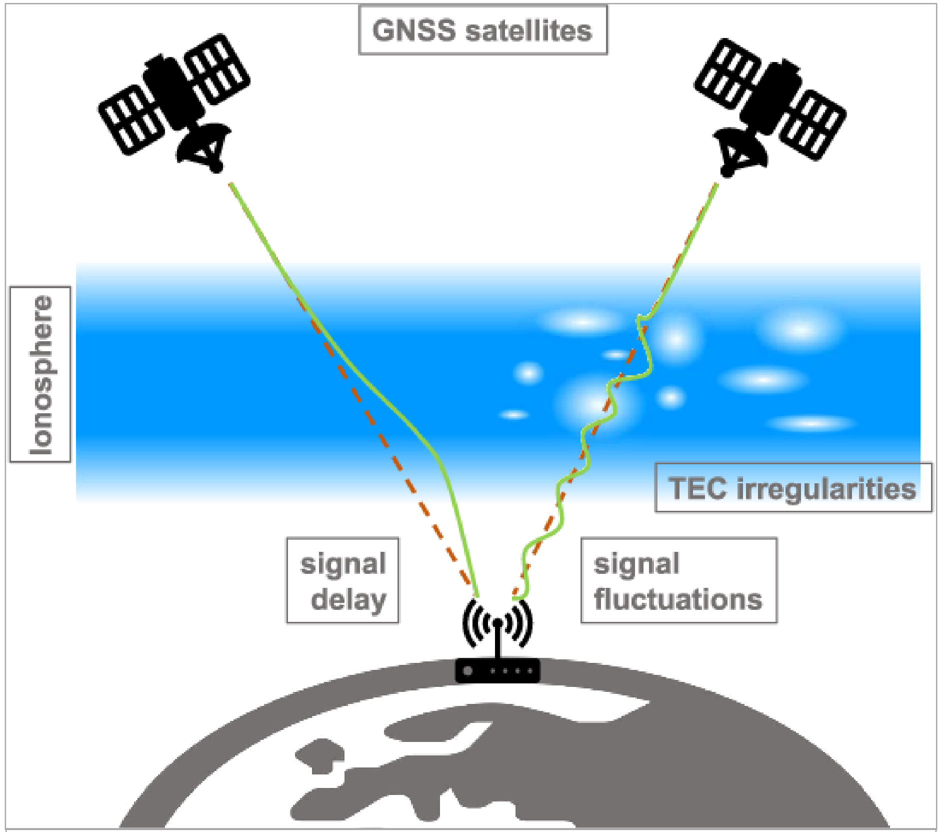
Research problem

- ➢ Ionospheric refraction of GNSS signals
- ➢ Dual-frequency obs. → integrated electron density
- ➢ Vertical Total Electron Content (VTEC)



Ionospheric VTEC map

Image source: ESA (background), DGFI-TUM (VTEC map).

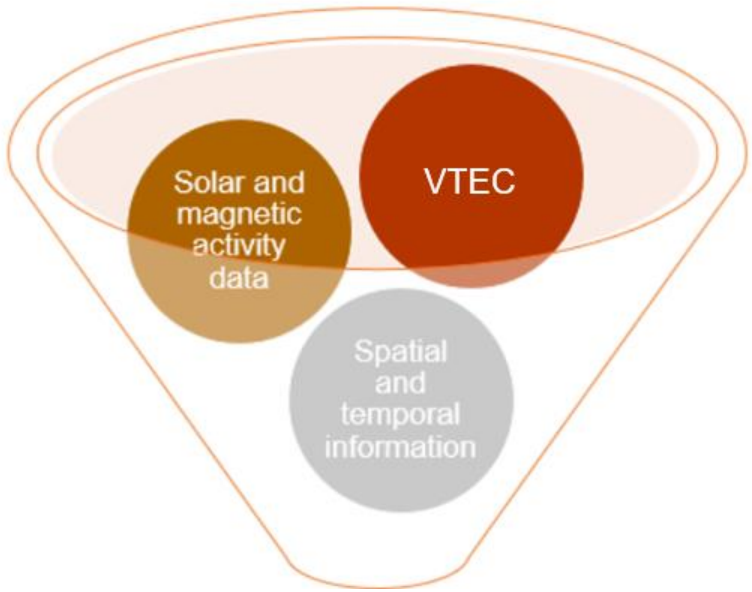


Source: https://www.semanticscholar.org/paper/Detection-of-GNSS-Ionospheric-Scintillations-Based-Linty-Farasin/3bc53da7342d4cdcd1a8bacfdc92651aeb62d5dc

Research problem

Objectives:

- ➢ Model / forecast VTEC accurately and precisely
  - ➢ Including solar-terrestrial processes (space weather)



- ○ Nonlinear processes
- ○ Complex, dynamical conditions
- ○ Limited understanding
- ○ Unkown function

➢ How to find a function?

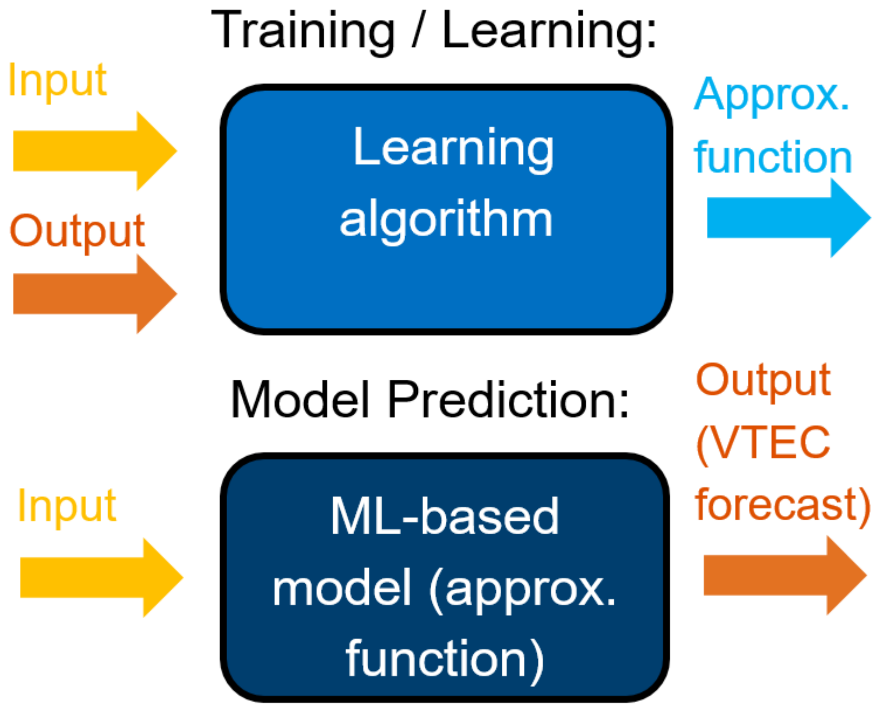
Research problem

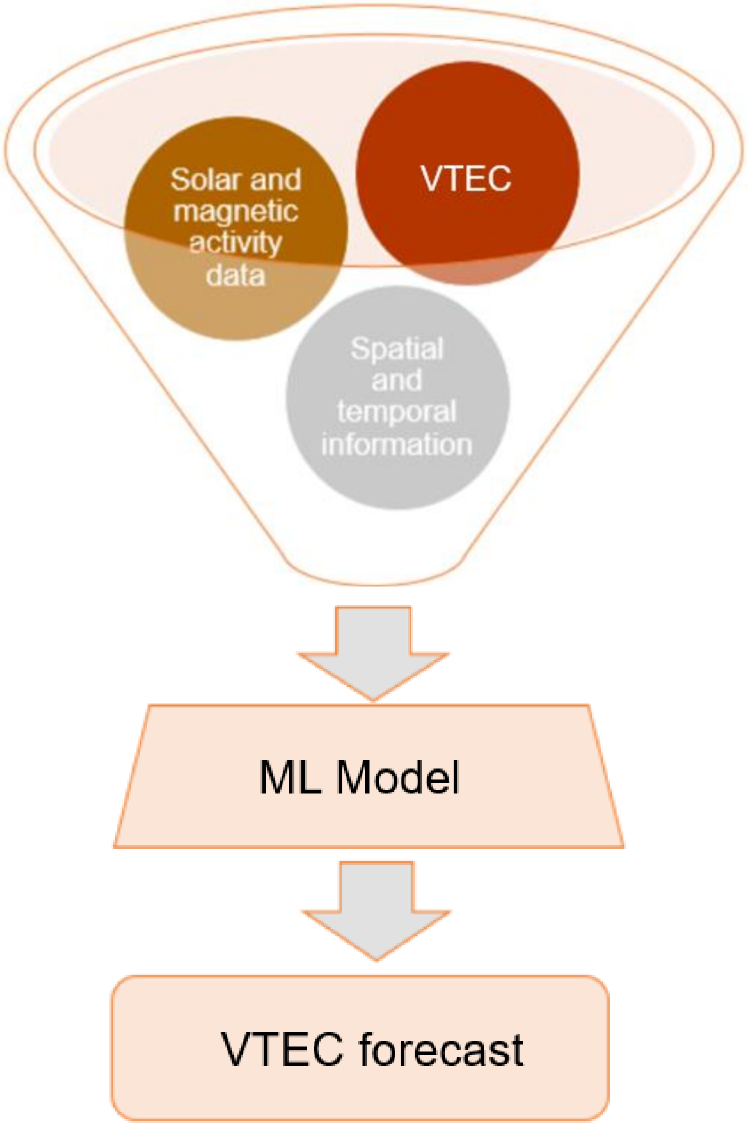
I. How to find a function?

- “Learn” from the (high-dimensional) data
- Approximate nonlinear functions

Machine Learning (ML)

Supervised learning

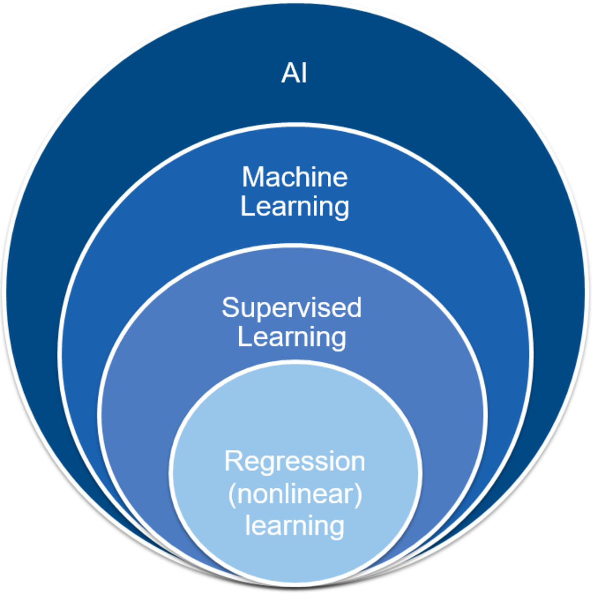




Learning algorithms

I. How to find a function?

- ➢ Linear models → (Deep) Neural networks
- ➢ Low complexity → High complexity
- ➢ Interpretability / performance trade-off



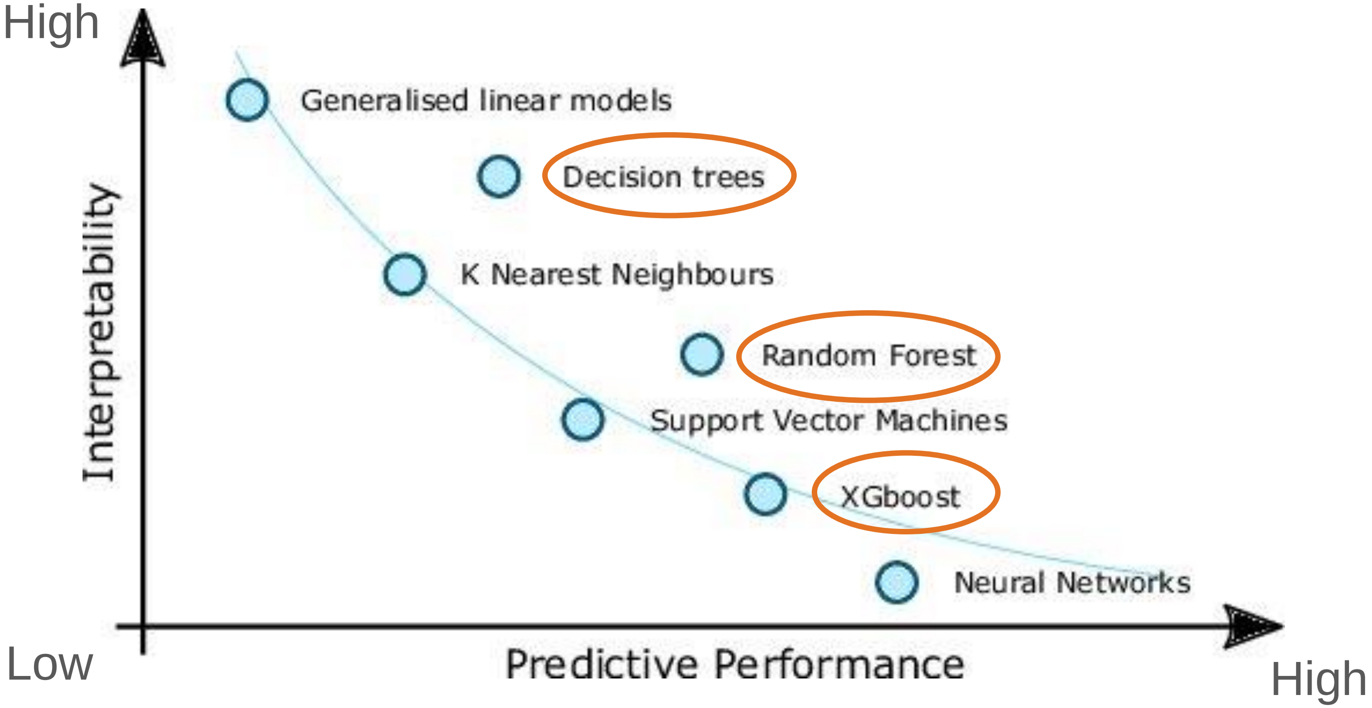
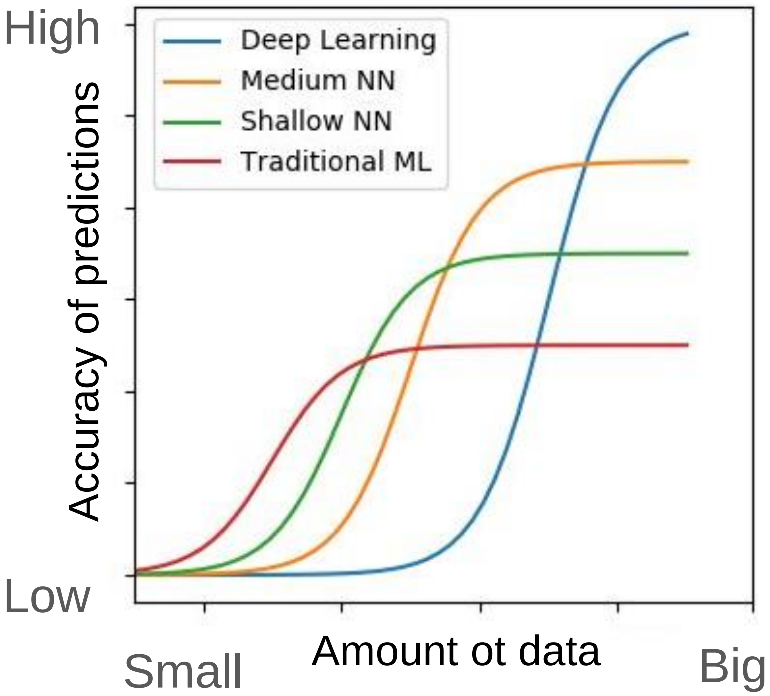
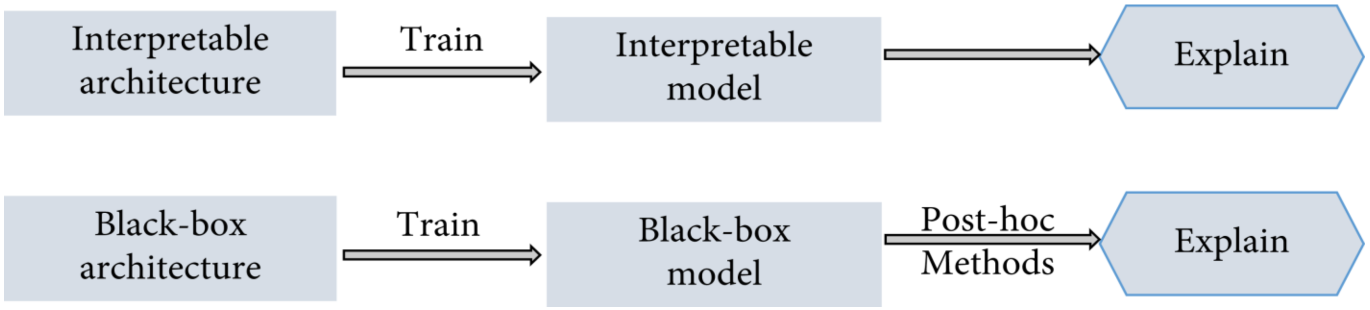


Image source: https://www.researchgate.net/publication/341509975_Agent-Based_Explanations_in_AI_Towards_an_Abstract_Framework



Adapted from: https://doi.org/10.1145/3332186.3332209

- Aim → Predictions:
  - Accurate and precise
  - Reliable
  - Trustworthy
  - Explainable / Interpretable



Source: https://doi.org/10.1155/2021/2939334

II. How to accomplish that?

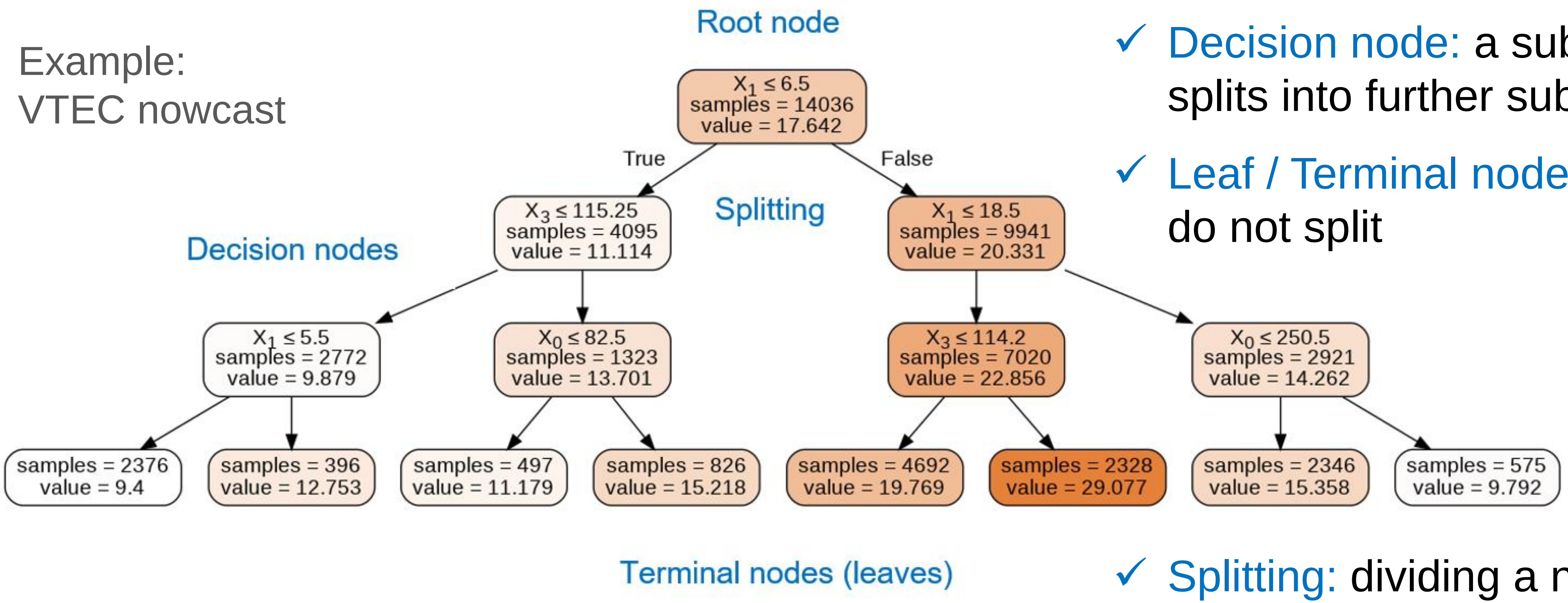
- ✓ (Permutation) feature importance
- ✓ Feature interaction
- ✓ Partial dependence
- ✓ LIME (local surrogate)
- ✓ SHAP (SHapley Additive exPlanations)

….

1. Using interpretable models

Decision tree learning

Example:
VTEC nowcast

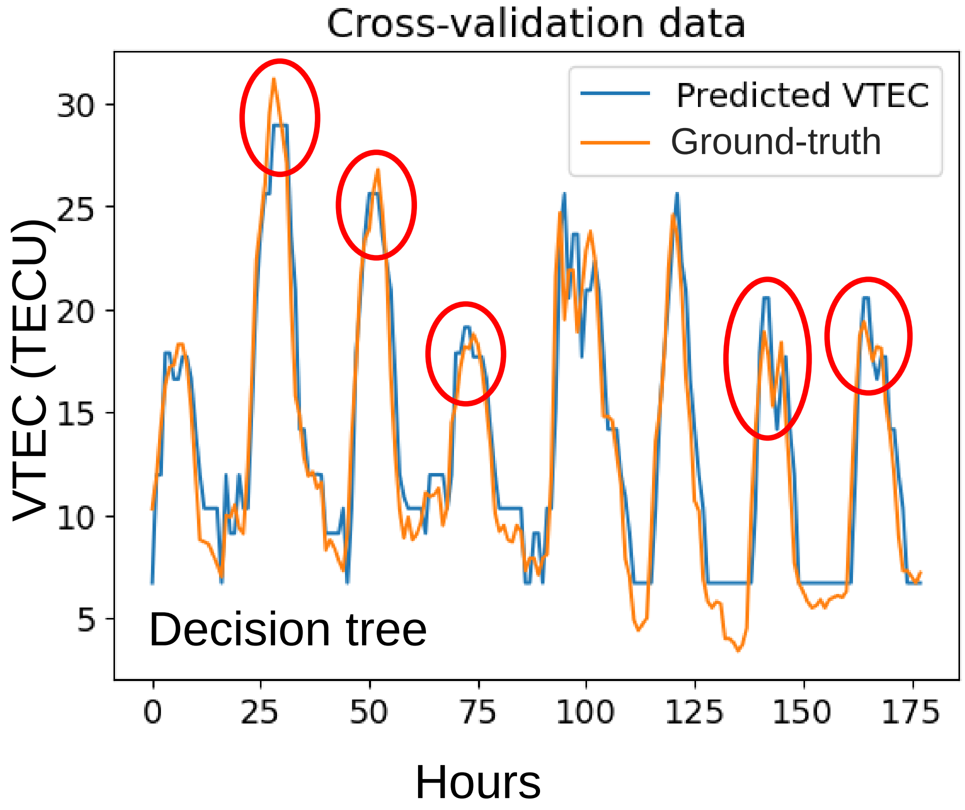


Final outcome: average VTEC in the particular leaf node.

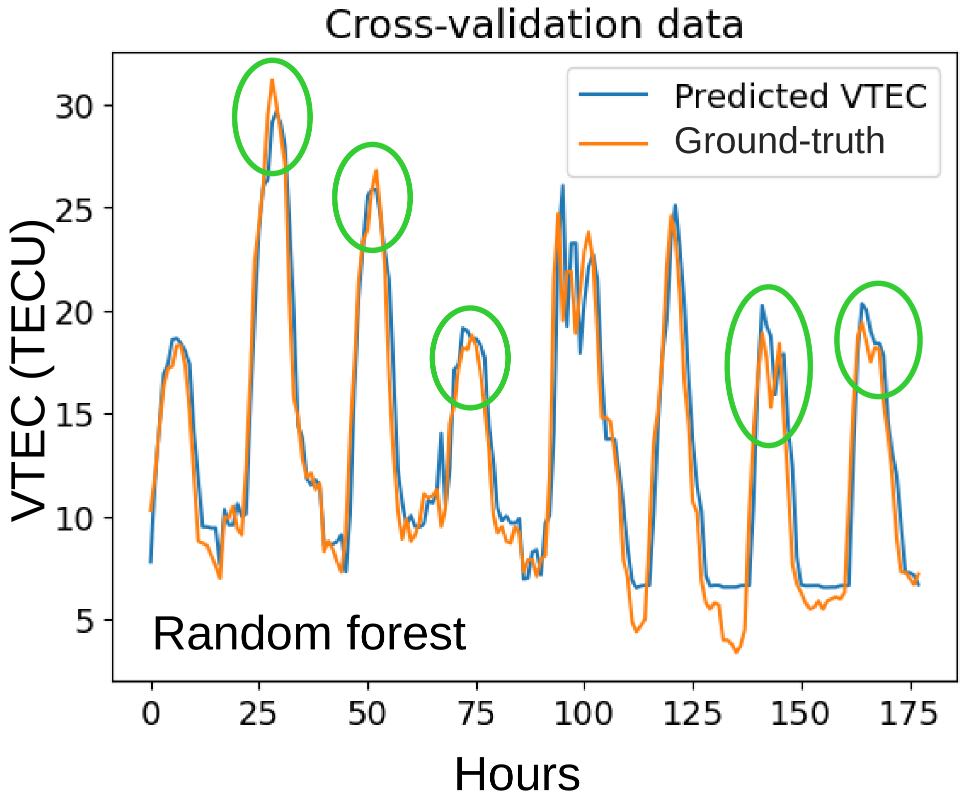
- ✓ Root node: entire dataset; further divided into 2 subnodes
- ✓ Decision node: a subnode that splits into further subnodes
- ✓ Leaf / Terminal node: nodes that do not split
- ✓ Splitting: dividing a node into 2 subnodes by calculating reduction in variance

Decision tree

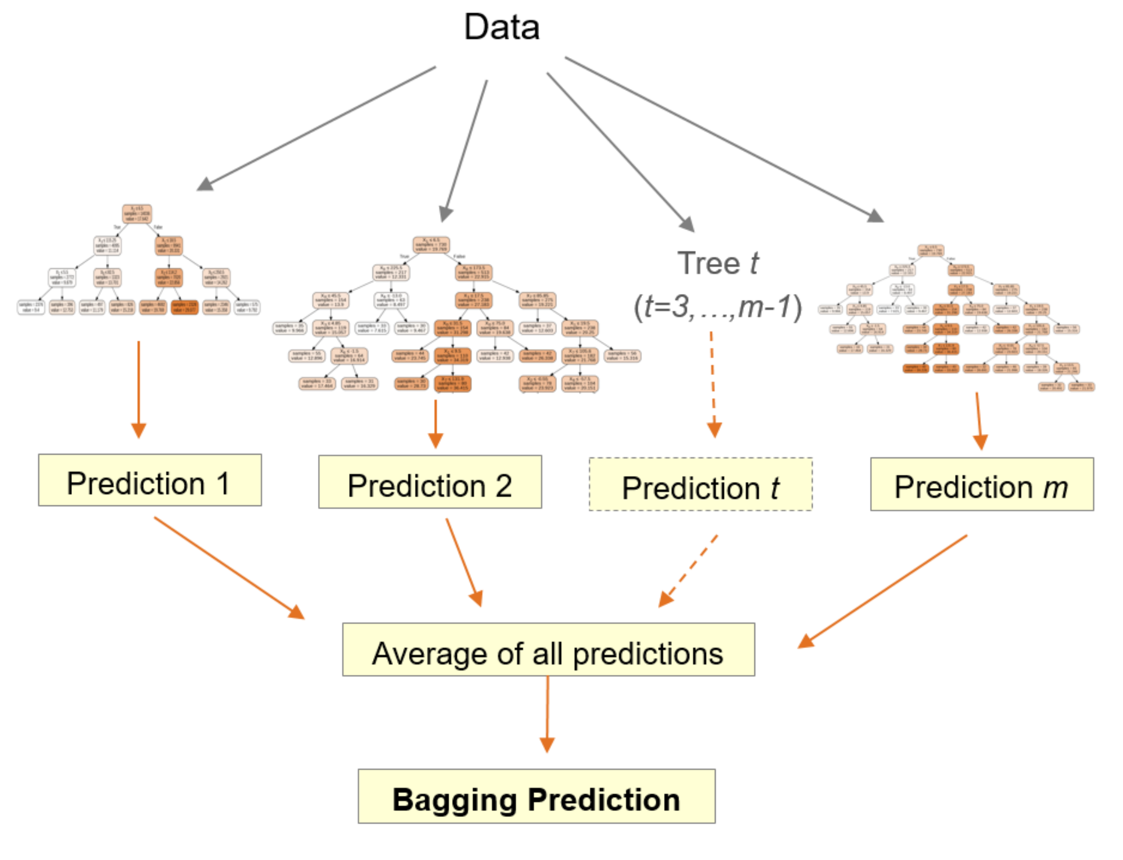
✓ Highly interpretable → follows a similar pattern to human thinking
✓ Lack of smoothness



Overcome issues → combining many trees
✓ Prediction smoother, more robust and more accurate
✓ Cost: reducing interpretability by increasing number of trees



Tree-based ensemble

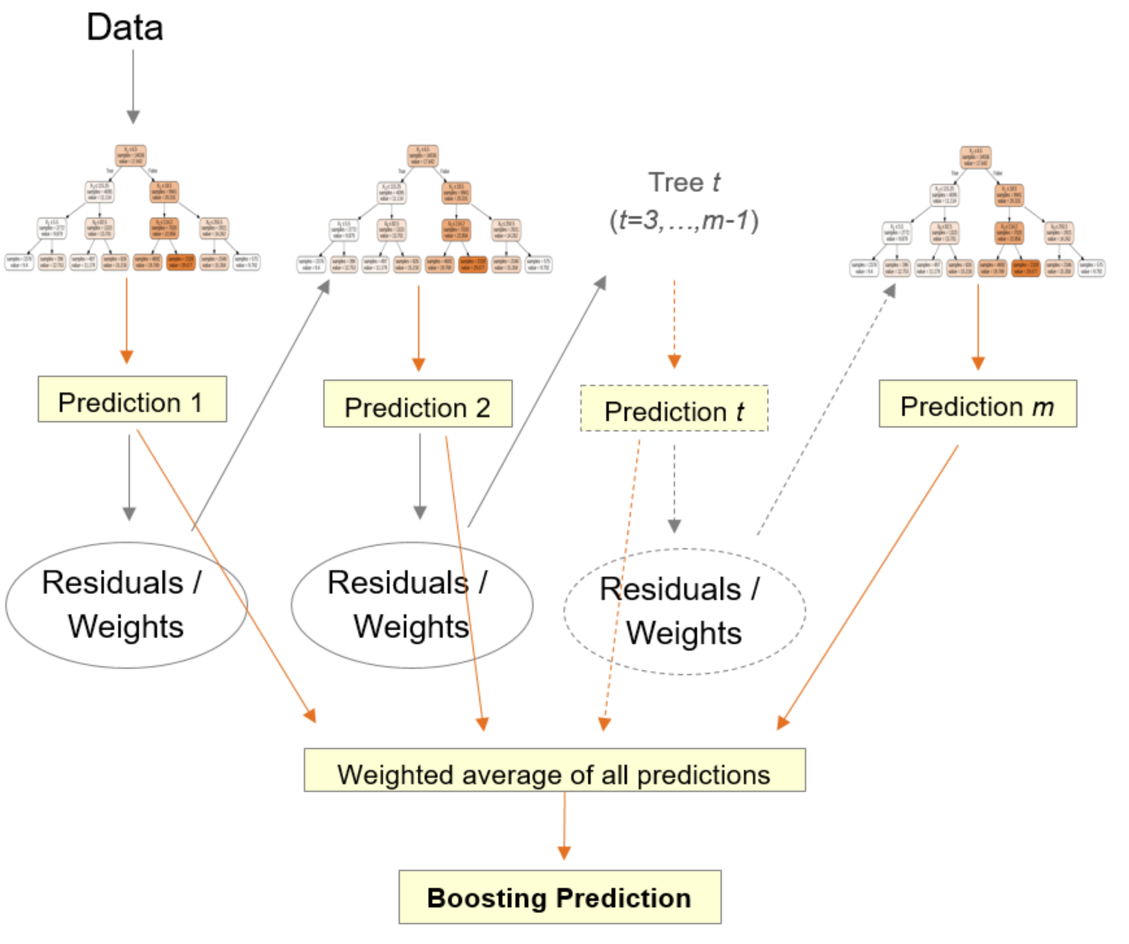


Bagging (Parallel Ensemble Learning):

- Random Forest (multiple randomized trees)

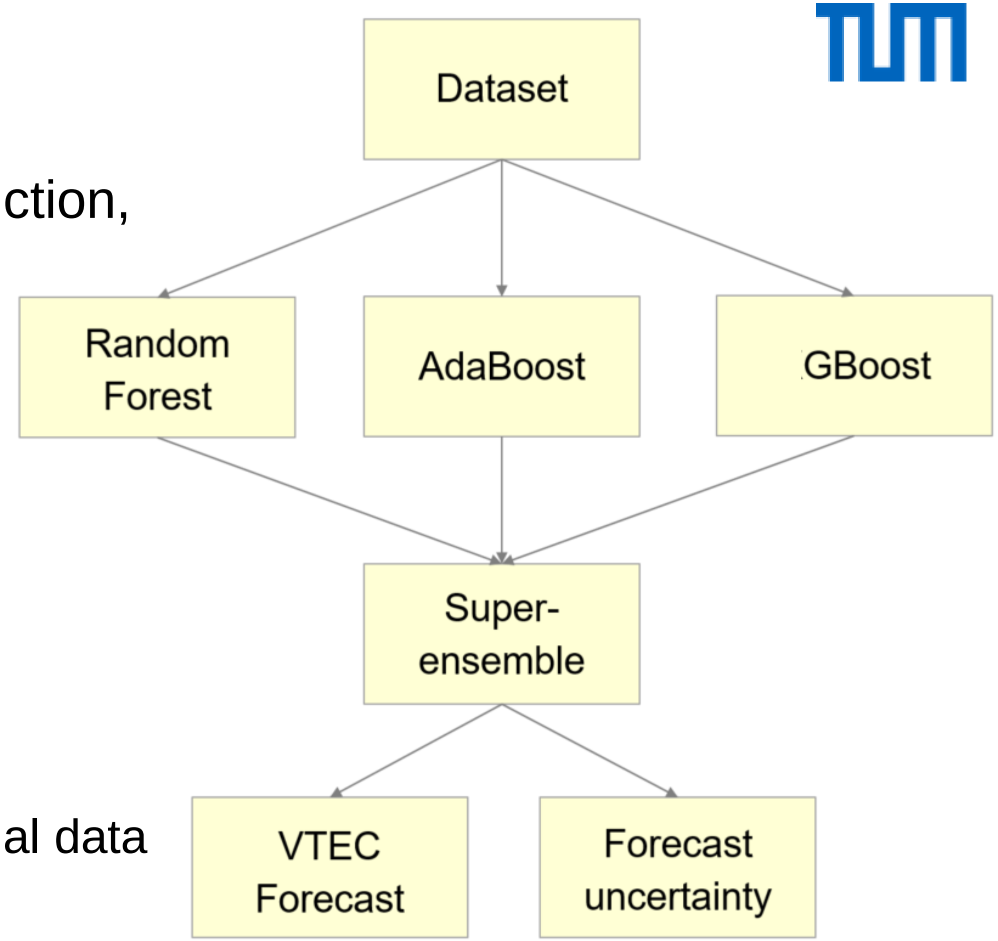
Boosting (Sequential Learning):

- Adaptive Boosting - AdaBoost (weighted obs.)
- Gradient Boosting - GBoost (residuals)



2. Quantifying uncertainties

- ✓ Define the accuracy and precision of VTEC prediction,
- ✓ Quantify the level of trust in VTEC prediction,
- ✓ Increase the reliability of VTEC predictions.



I. Multi-model and multi-data ensemble

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

II. Confidence interval

- Quantile objective loss function
- Applied for GBoost and 3rd dataset
- Quantiles: upper bound $\alpha = 0.95$, lower bound $\alpha = 0.05$

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

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

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

*Observations were preprocessed / cleaned before training.

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
- VTEC moving averages over previous 4 days and 30 days
- Derivatives of VTEC

Output data:

- VTEC (GIM CODE)
  - 10E 70N,
  - 10E 40N,
  - 10E 10N

Time: t+24h

Data split:

- Training
- Cross-validation

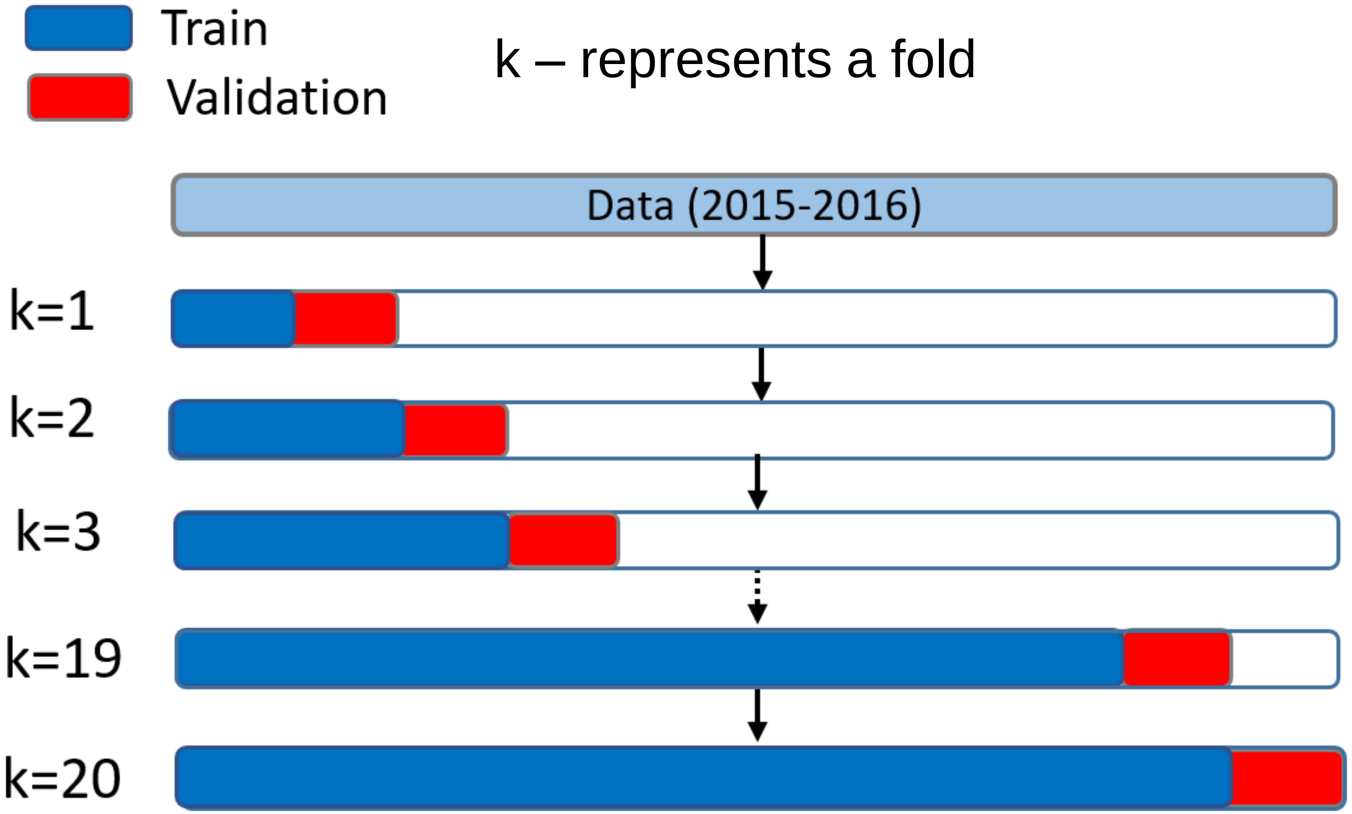
2015 - 2016

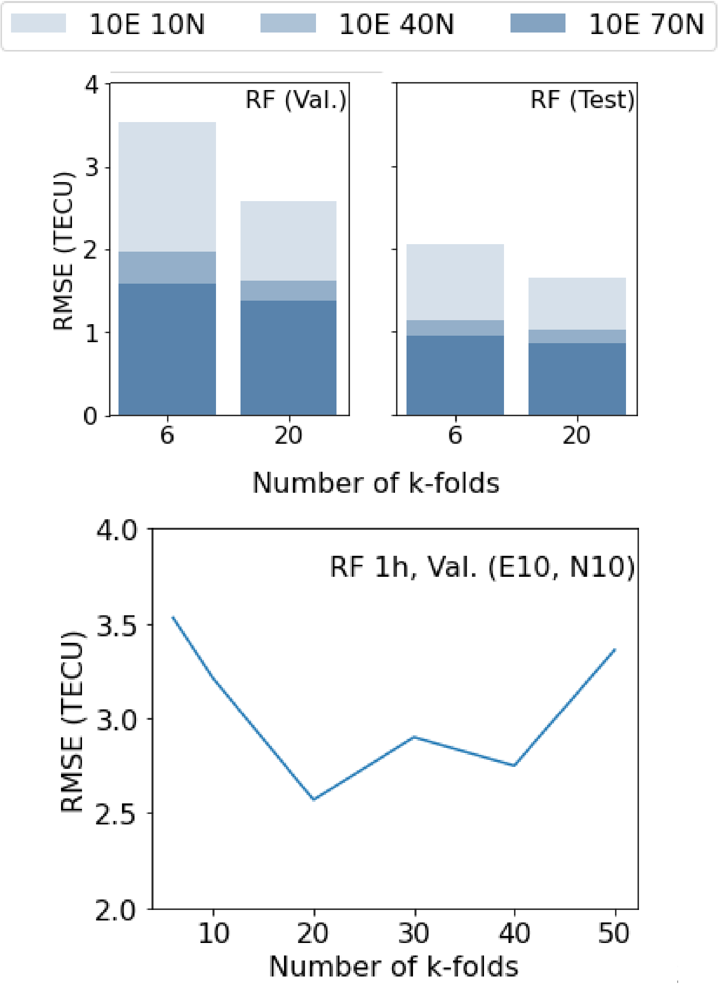
- Test

2017

Model training and optimization / evaluation

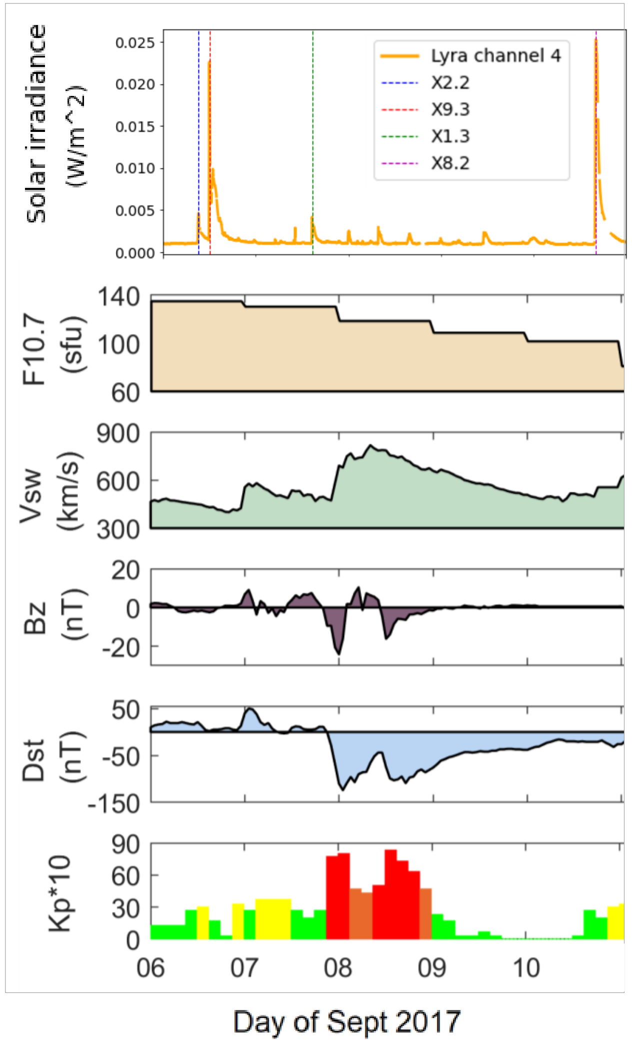
- Temporal structure of time series → Cross-validation on a rolling basis
- Evaluate model performance in a robust way



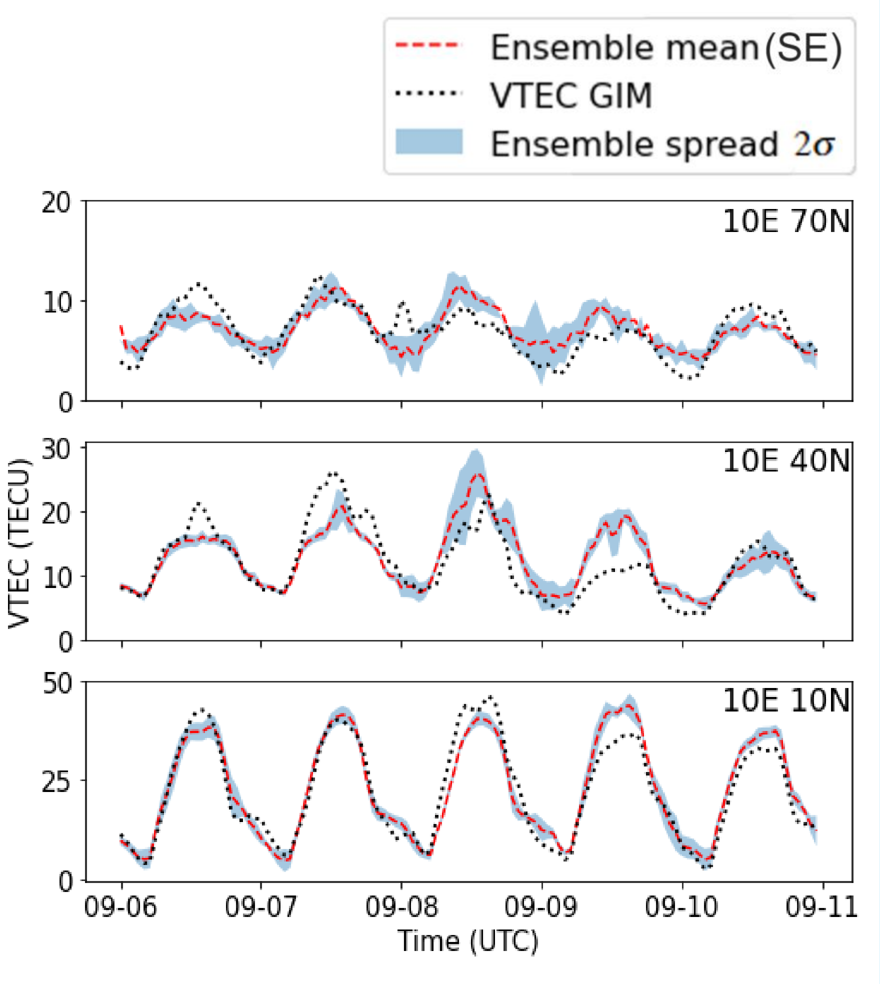


Results: September 2017 space weather events

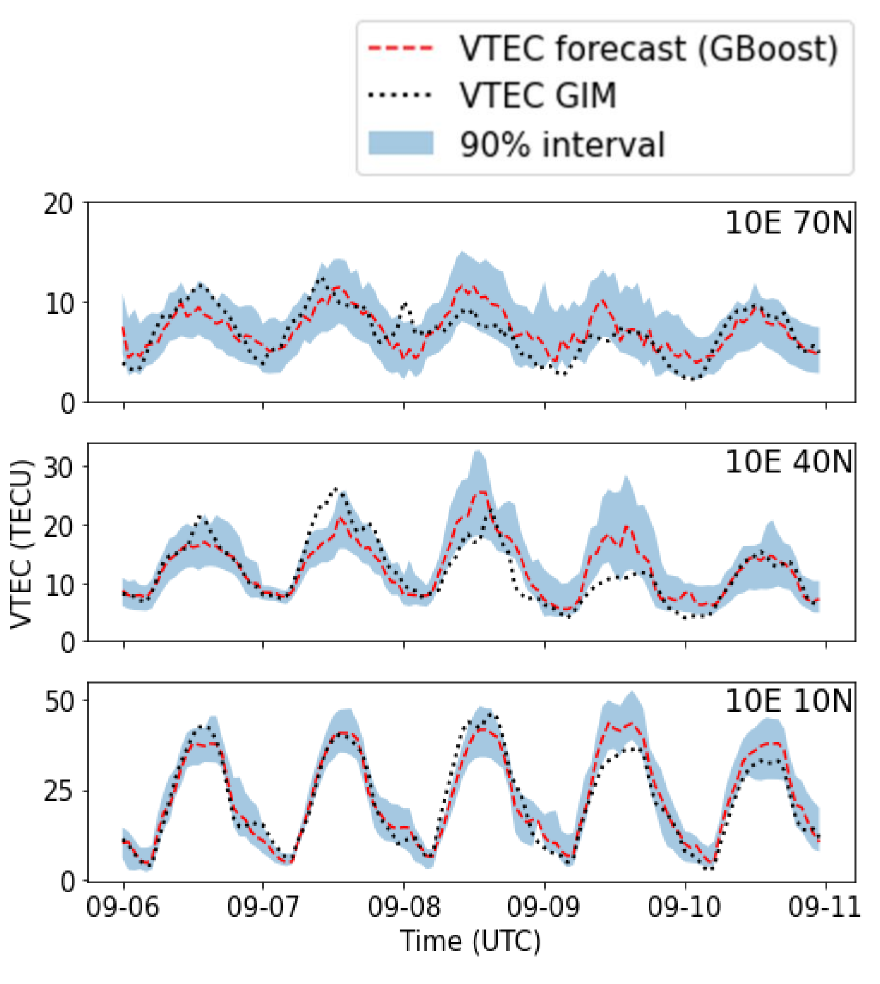
Space weather overview



UQ: Ensemble method



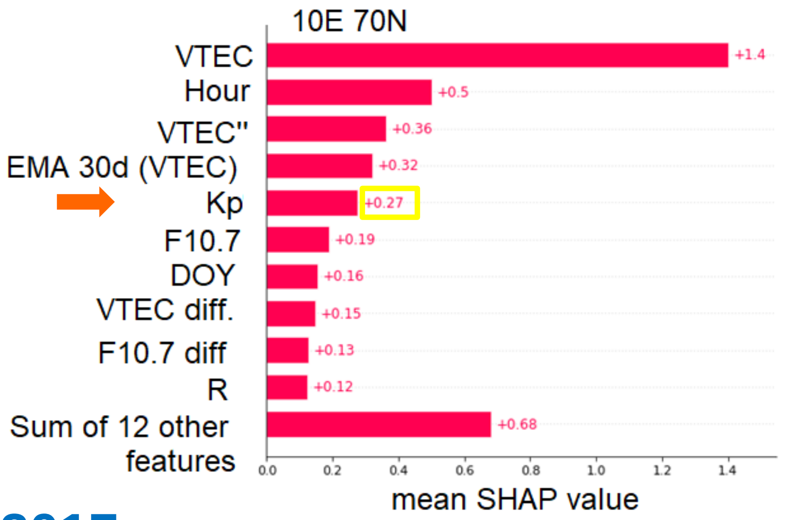
UQ: Confidence intervals

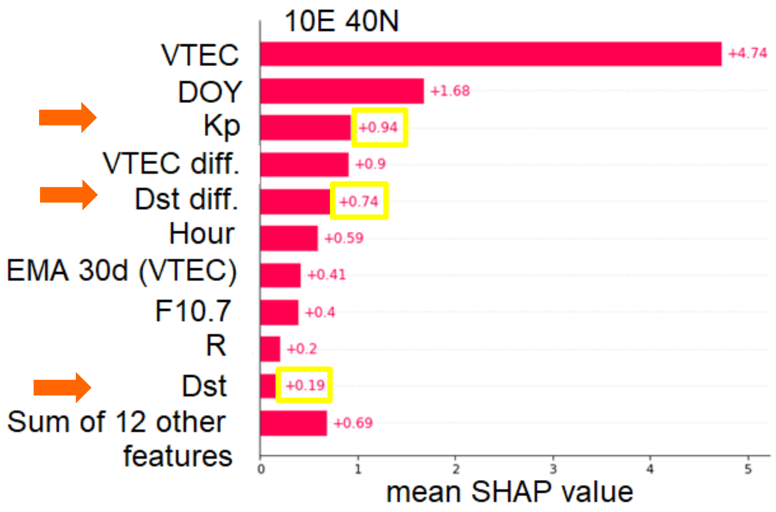


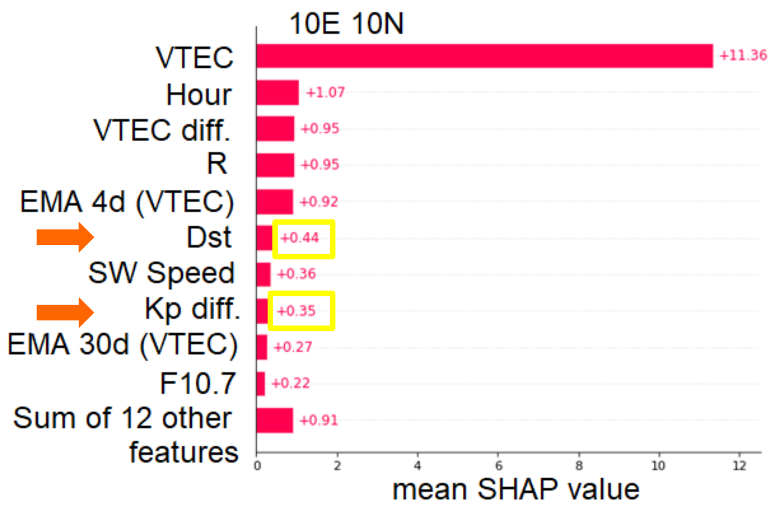
Average features impact on the output

GBoost model

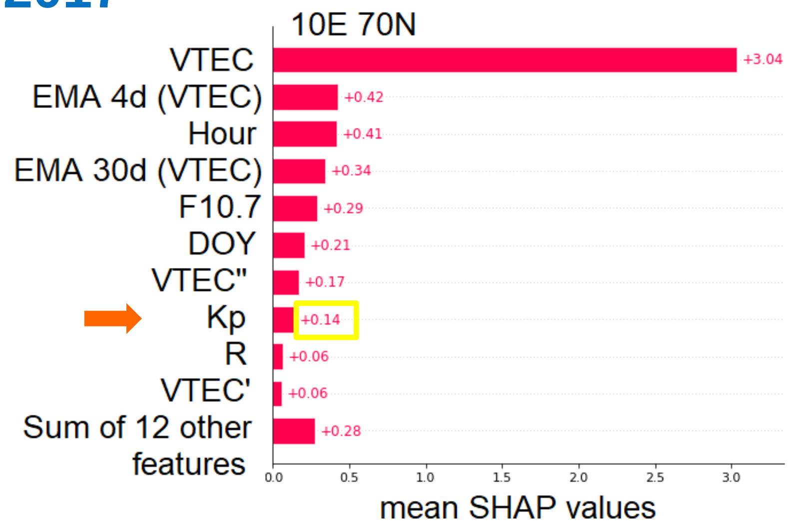
Sept 8-9, 2017

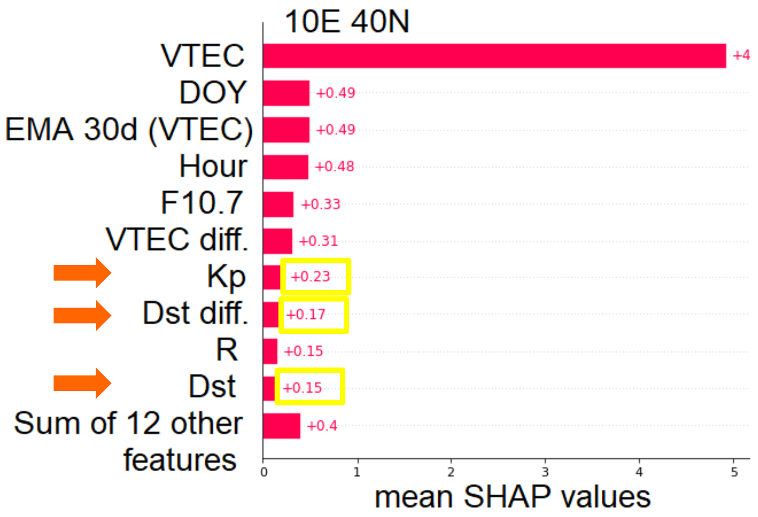


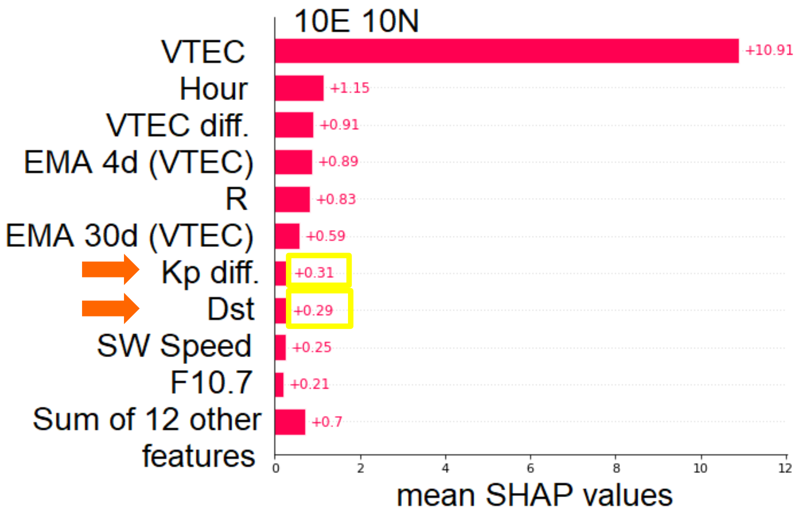




2017

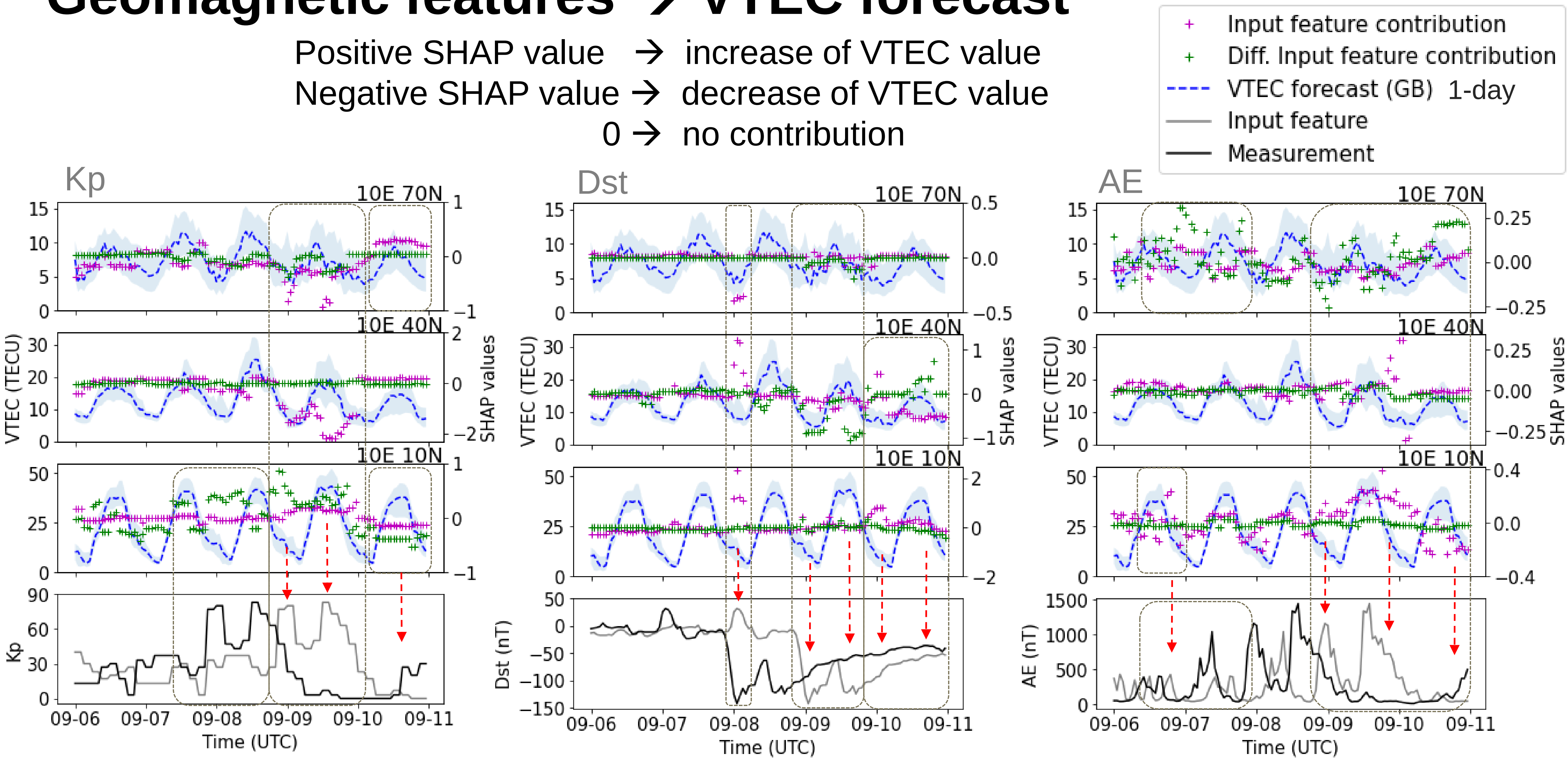






Geomagnetic features → VTEC forecast

Positive SHAP value → increase of VTEC value
Negative SHAP value → decrease of VTEC value
0 → no contribution



Conclusion

- Learning algorithms: interpretability / performance trade-off, amount of data
- The uncertainty information defines the reliability and precision of VTEC predictions
- Uncertainty quantification allows to assess the trustworthiness of predictions
- Ground-truth VTEC within predicted confidence intervals for space weather events
- Higher contribution from geomagnetic-related input features during the storm

Thank you for your attention!

Randa Natras

Technical University of Munich

randa.natras@tum.de