Interpretable Machine Learning Methods for forecasting the SYM-H Index

Daniel Iong, Yang Chen, Enrico Camporeale*, Tamas Gombosi, Tuija Pulkkinen, Gabor Toth, Shasha Zou

University of Michigan, Ann Arbor, *University of Colorado

UNIVERSITY OF MICHIGAN

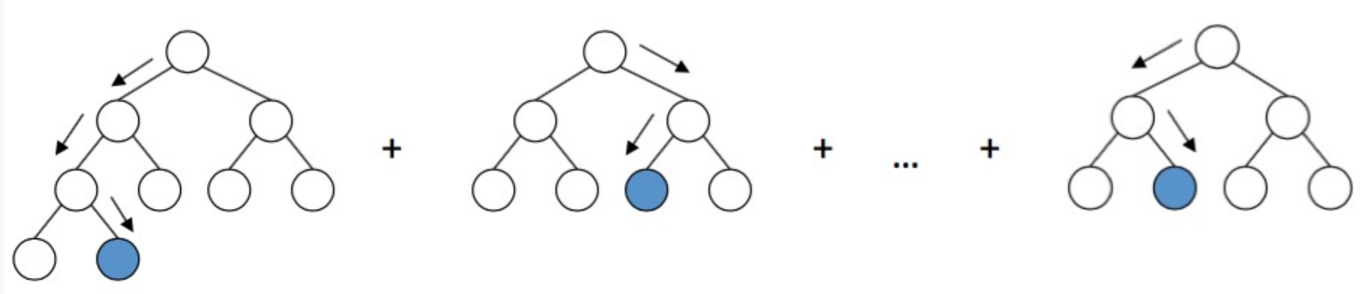
SOLSTICE Solar Storms & Terrestrial Impacts Center

An interactive web application showing our results is available at: **https://geomag-interpret.herokuapp.com/**

Summary

- We train **gradient boosting machines** (GBMs) to predict the SYM-H index **1-2 hours** ahead using past SYM-H values and solar wind/IMF parameters
- GBMs are a **highly accurate and interpretable** class of methods for forecasting the SYM-H index.
- GBMs achieve **higher accuracy** than current ML methods based on neural networks for forecasting the SYM-H index.
- Feature importance scores can be extracted to **interpret predictions** from GBMs and can **provide insight** into the complex relationship between the solar wind and Earth's ring current.

Gradient Boosting Machines (GBM)



- GBMs are a **sequential ensemble** of shallow **decision trees** constructed using gradient boosting.
  - Each successive tree improves upon the previous trees.
- GBMs have been shown to be **more accurate than neural networks** for prediction on tabular data in various domains. (Shwartz-Ziv et al. (2021))
- Like neural networks, GBMs can model **complex interactions** between features.
- GBMs are less susceptible to issues arising from correlated features.
- **Hyperparameters** tuned to optimize performance: number of trees, max. tree depth, min. child weight, learning rate, column subsampling ratio.
- We use the open-source framework **XGBoost** to train our GBMs.
- We compute **SHAP values** to obtain an individual feature's contributions to predictions made by GBM at any given time (*See Figure 1*).

Data

- To train our models, we use different combinations of the following features:
  - **Solar wind parameters** (*velocity, density, temperature*) from the *past 1 hour*
  - **IMF parameters** (*Bx, By, Bz*) from the *past 1 hour*
  - **Derived parameters** (*dynamic pressure, electric field*) from the *past 1 hour*
  - **Previous SYM-H** values from the *past 1 hour*
- The **time resolution** of both the SYM-H index and solar wind/IMF parameters is **5 minutes**.
- Solar wind and IMF parameters are from **ACE level 2 data** obtained from NASA CDAWeb. SYM-H are obtained from OMNI.
- Our models are trained, validated, and tested on a total of **42 strong geomagnetic storms** (DST < -100) between 1998-2018, which were used in Siciliano et al. (2021).

Results

The April 2001 storm (shown below) is the ***strongest storm*** *(min. DST < -400 nT) in the test set.*

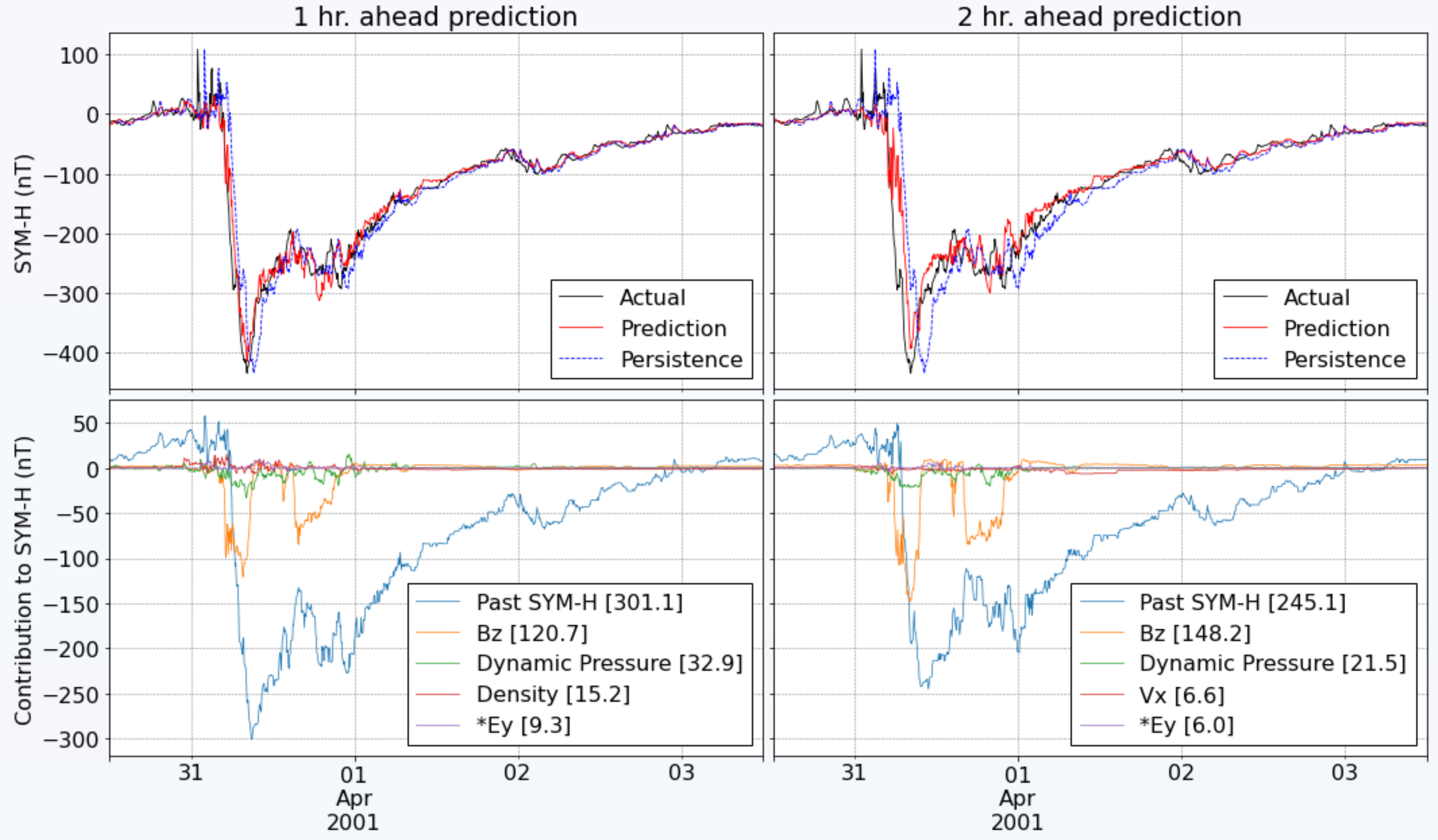


Figure 1: 1hr. and 2hr. ahead predictions from GBM trained on *all considered features* (**top**) and corresponding feature importance scores (**bottom**). These scores show how each feature contributes to the prediction made by GBM at any given time. A *larger* max. absolute value (**in brackets**) means the feature is *more* influential. The sum of contributions from all features is *approximately* equal to the prediction.

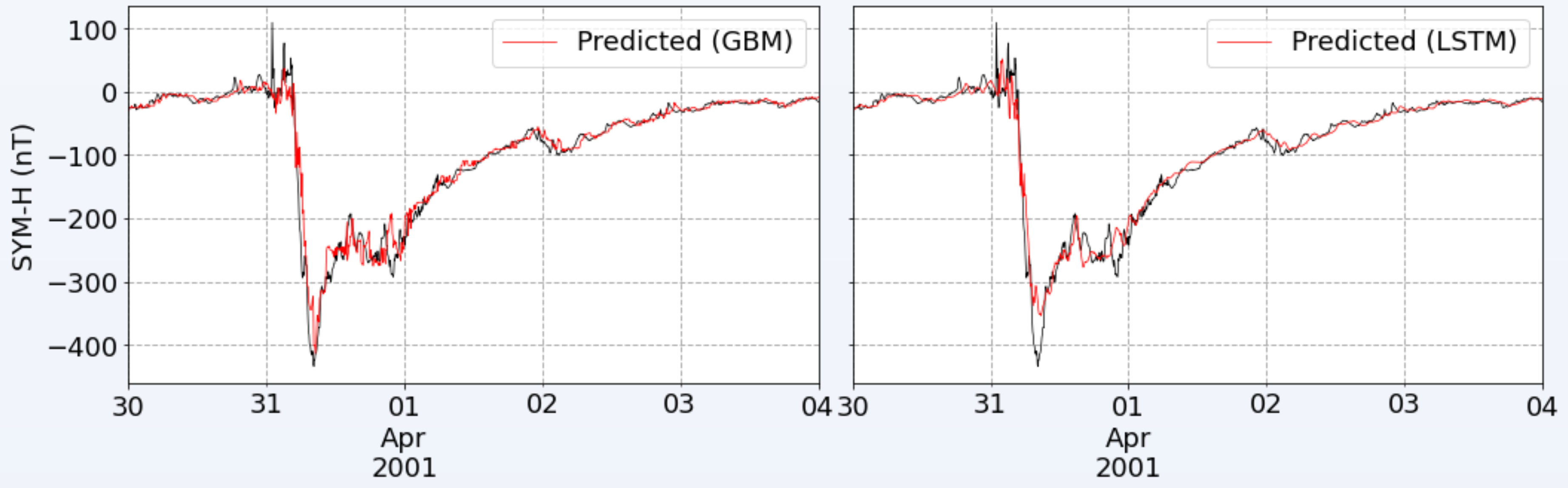


Figure 2: Comparison of **1 hr. ahead** predictions from GBM vs. LSTM developed in Collado-Villaverde et al. (2021), trained on **only past SYM-H and IMF parameters**.

	GBM	Collado-Villaverde et al.	Siciliano et al.	Persistence
RMSE (nT)	7.4	7.9	8.5	9.4

Table 1: RMSE values for **1 hr. ahead** predictions from GBM vs. existing methods on **17 strong test storms** trained on 25 strong storms using **only past SYM-H and IMF parameters** as features.

** Ey = max(Bz, 0) * Vx*

Results

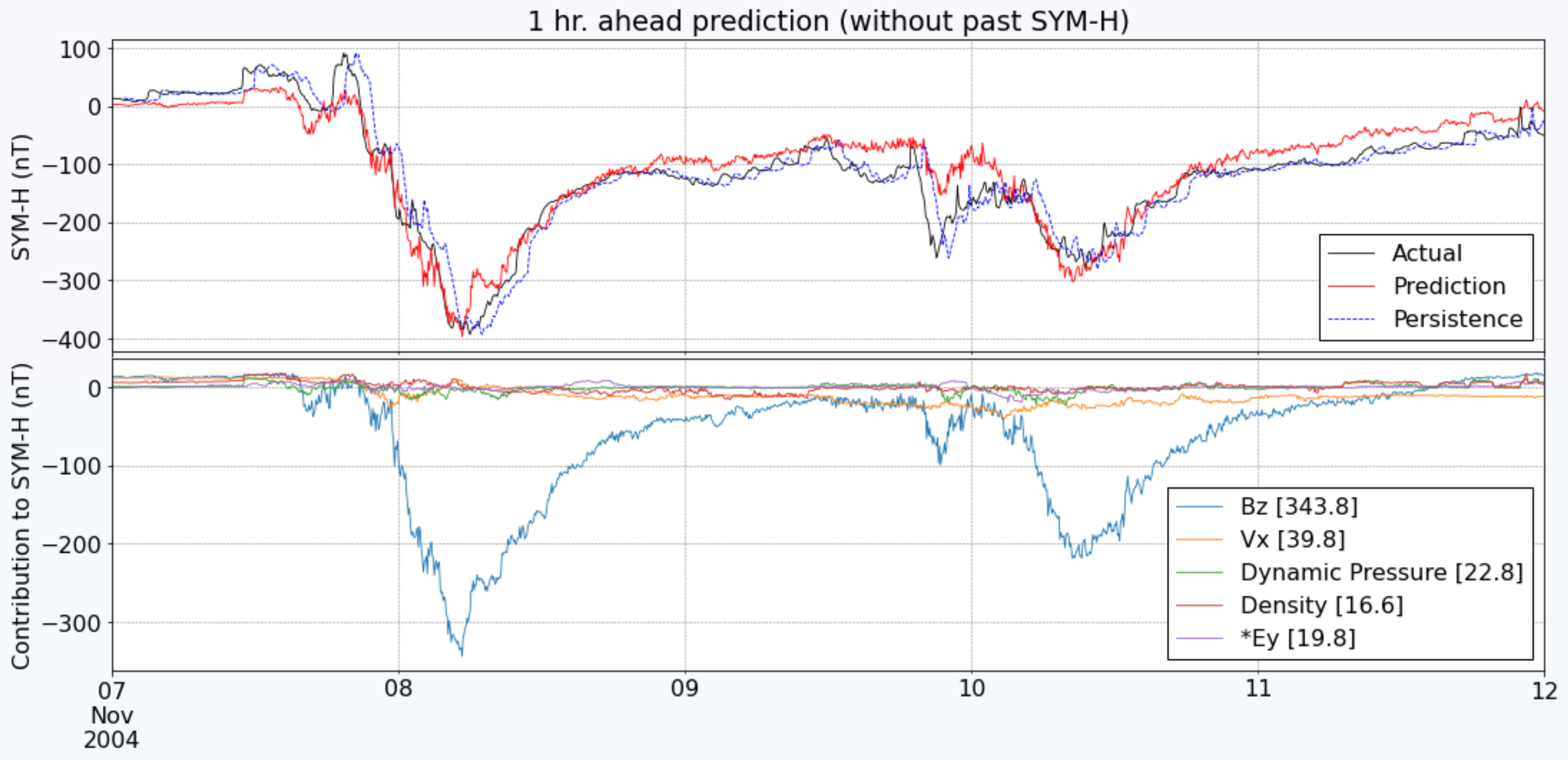


Figure 3: 1hr. ahead prediction from GBM trained on solar wind/IMF parameters only (**top**) and corresponding feature importance scores (**bottom**).

Discussion

- The combined contribution from solar wind speed, motional electric field, and dynamic pressure is only a few percent in Figure 1, indicating that these effects are implicitly included in the contribution from past SYM-H.
- This is the first time an interpretable ML method has been applied to SYM-H forecasting, which opens **new possibilities for detailed investigation**.
- GBMs for forecasting SYM-H are highly accurate for **operational** use (*trained with only past SYM-H and IMF parameters*) and can be used to extract **novel insight** into the relationship between solar wind and Earth's ring current.
- **Work in progress** includes:
  - Studying how feature contributions vary for different types of storms
  - Investigating how correlation among features affects SHAP values
  - Evaluating GBM predictions using other metrics besides RMSE

References

A. Collado-Villaverde et al., Deep neural networks with convolutional and LSTM layers for SYM-H and ASY-H forecasting. **Space Weather, 19,** e2021SW002748 (2021). doi: 10.1029/2021SW002748

F. Siciliano et al., Forecasting SYM-H index: A comparison between long short-term memory and convolutional neural networks. **Space Weather, 19,** e2020SW002589 (2021). doi: 10.1029/2020SW002589

R. Shwartz-Ziv et al., Tabular Data: Deep Learning is Not All You Need. (2021) **arXiv**:2106.03253 [cs].

Acknowledgement

This work was supported by a NASA Heliophysics DRIVE Science Center (SOLSTICE) at the University of Michigan under grant NASA 80NSSC20K0600

Contact: Daniel Iong (*daniong@umich.edu*)