Machine Learning Models as an Alternative to Standard Interpolation Techniques for Estimating Gaps in OMNI Data

Jasmine Kobayashi(1), Dogacan Ozturk(1), Hyunju Connor(2), Amy Keesee(3)
1: Geophysical Institute, University of Alaska Fairbanks, 2: NASA/GSFC, 3: Space Science Center, University of New Hampshire

MAGICIAN
Machine Learning Algorithms for Geomagnetically Induced Currents in Alaska and New Hampshire

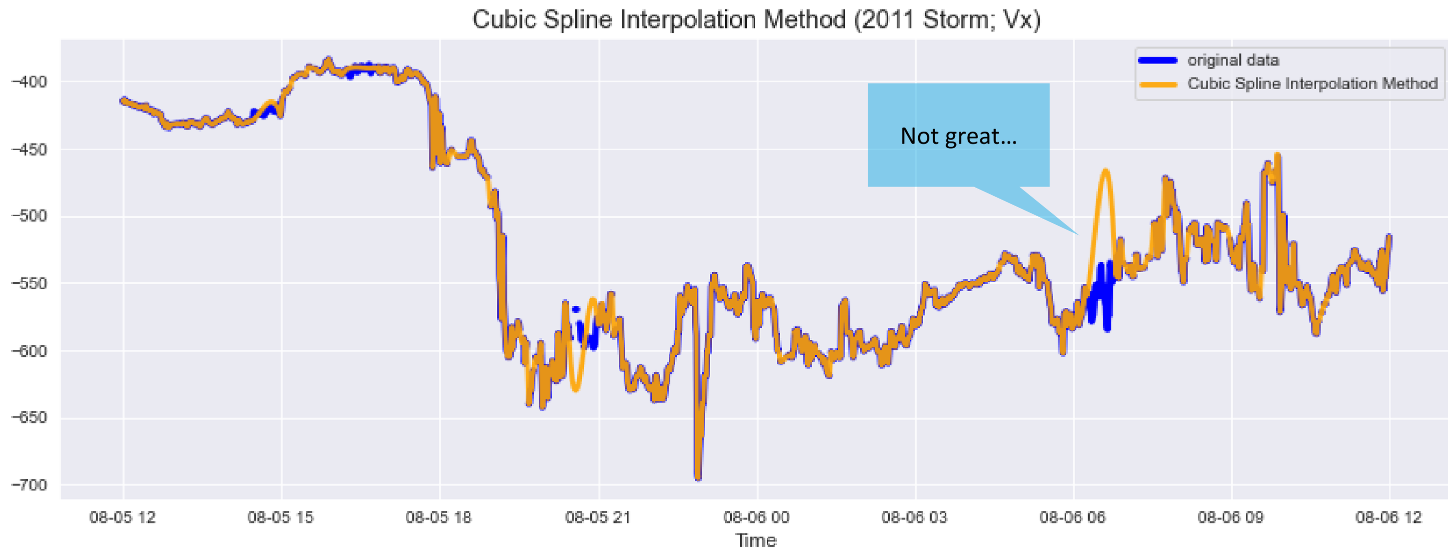
MOTIVATION & INTRODUCTION

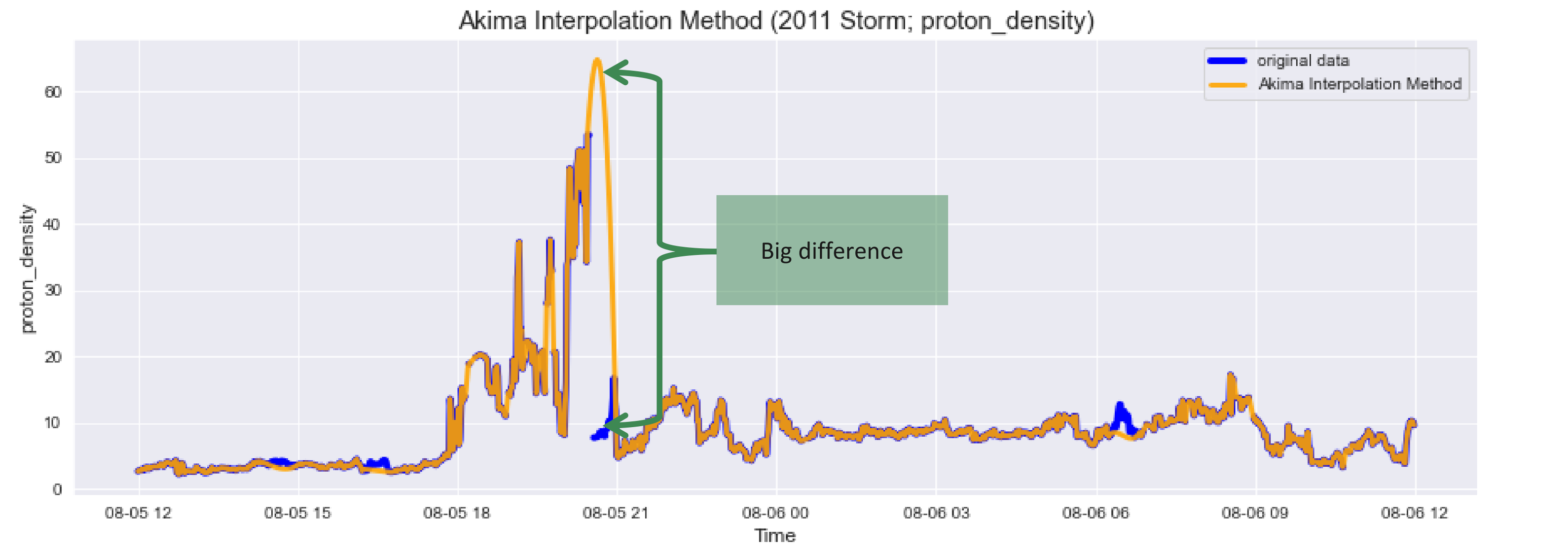
OMNI data provides conditions of the near-Earth environment and is widely used to drive numerical and machine learning models. However, especially during storm intervals, there are significant gaps in OMNI data.

OMNI Dataset:

- Contains approximately 20% of missing plasma parameter data
- Approximately 8% of missing IMF measurements

>> Both first-principles and ML models require **continuous input.**





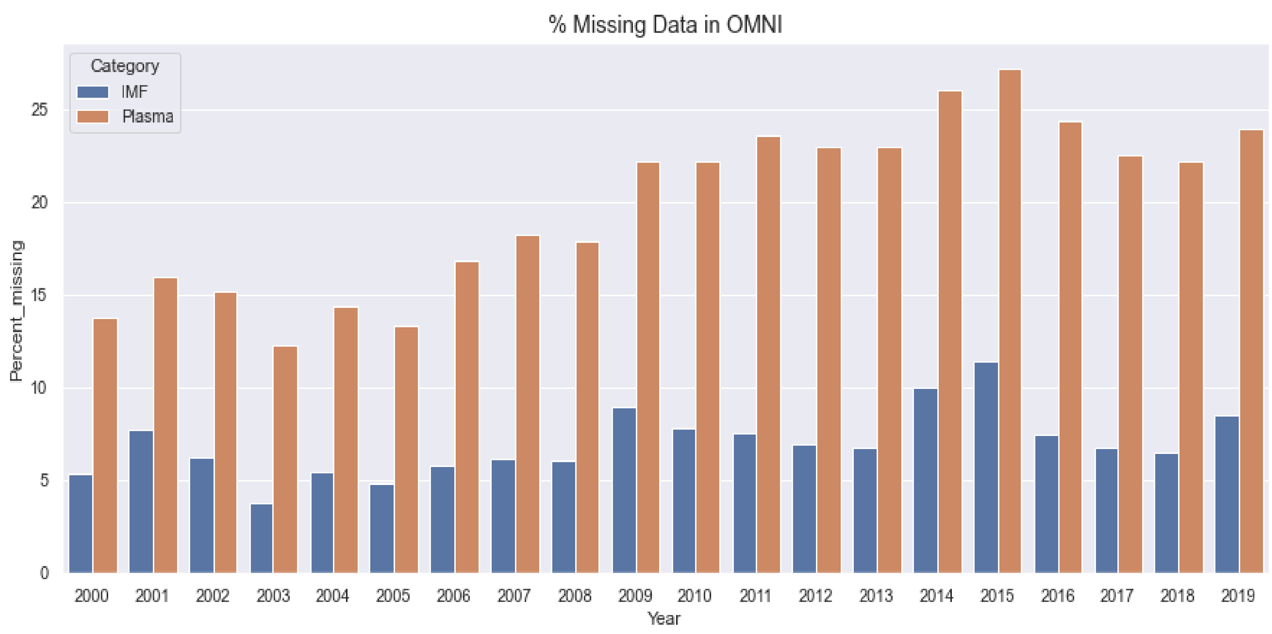


Fig. 1 (Left): Histogram of the percentage of missing plasma & IMF data in OMNI for each year. Starting with 2000 up to 2019 (left to right)

Table 1 & Table 2 (Right): Evaluations of the different interpolation methods based on three forms of randomly created data gaps (all adding up to 2 hours).

2011 Storm Vx Data gap Interpolation	Linear/Time (R2 Score)	Linear/Time (RMSE)	Nearest (R2 Score)	Nearest (RMSE)	Spline (order:2) (R2 Score)	Spline (order:2) (RMSE)	Spline (order:3) (R2 Score)	Spline (order:3) (RMSE)	C. Spline (R2 Score)	C. Spline (RMSE)	Akima (R2 Score)	Akima (RMSE)
4 gaps; 30 mins	0.98452	10.43362	0.97870	12.23734	0.87309	29.86986	0.86657	30.62770	0.81142	36.41178	0.97626	12.91911
2 gaps; 60 mins	0.89895	4.60000	0.67223	8.24666	0.75246	7.19959	0.85432	5.52322	-1.46393	22.71431	0.90637	4.42780
1 gap; 120 mins	-0.14201	22.68688	-1.10566	30.80594	-0.81993	28.63971	0.19108	19.09389	0.11698	19.94918	-0.17574	23.01951

2011 Storm Np Data gap Interpolation	Linear/Time (R2 Score)	Linear/Time (RMSE)	Nearest (R2 Score)	Nearest (RMSE)	Spline (order:2) (R2 Score)	Spline (order:2) (RMSE)	Spline (order:3) (R2 Score)	Spline (order:3) (RMSE)	C. Spline (R2 Score)	C. Spline (RMSE)	Akima (R2 Score)	Akima (RMSE)
4 gaps; 30 mins	-0.53291	8.62459	-1.27086	10.49725	-18.07562	30.42421	-11.99745	25.11361	-53.64547	51.49406	-5.60573	17.90361
2 gaps; 60 mins	0.12596	0.47848	0.16050	0.46893	-0.53023	0.63311	0.12853	0.47777	-7.53889	1.49335	-0.31160	0.58614
1 gap; 120 mins	-3.06233	6.84591	-6.26026	9.15208	-0.32863	3.91513	-6.00046	8.98684	-12.31003	12.39177	-4.57167	8.01745

Fig 2 (Above Left) & Fig 3 (Above Right): Sample plots of the Vx and proton density data over time for the 2011 storm with randomly generated data gaps seen in blue. These plots are samples where the interpolation methods performed poorly, even for small data gaps.

- Linear interpolation has the lowest RMSE and highest R2 score.
- All interpolation methods generally do not perform well with large data gaps (especially with large variation).
- There is no interpolation method that performs well for Vx and consistently performs well with proton density.

METHODOLOGY

~20 years of OMNI data.
We tested the performance of traditionally used interpolation techniques vs ML models to fill plasma data gaps.

Interpolation

- Methods evaluated: Linear, Time, Nearest, Spline, Cubic Spline, Akima, PCHIP

ML Regression Models

- Linear, Polynomial, Random Forest (n_estimators = 10)
- Split types (train : test = 0.8:0.2)
  - Random: SciKitLearn random train_test_split method
  - Sequential: Manually take first 80% of the data as the training set, and the remaining 20% as the test set

Target

- Plasma parameters
  - Velocities (x,y,z)
  - Proton Density (Np)
  - Temperature

Features

- IMF vectors
- Auroral Indices
- SYM/ASY H

Performance evaluation

- August 2011 Storm

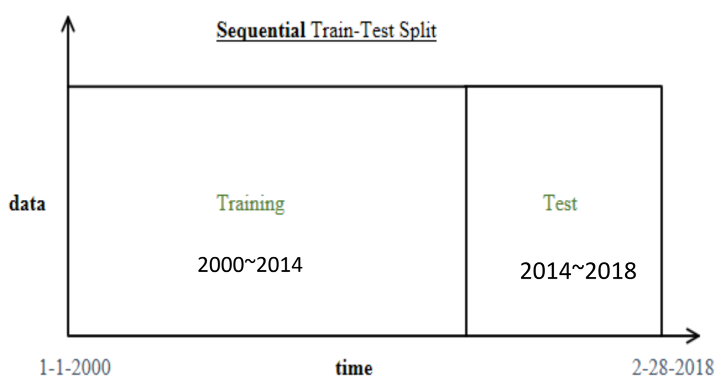


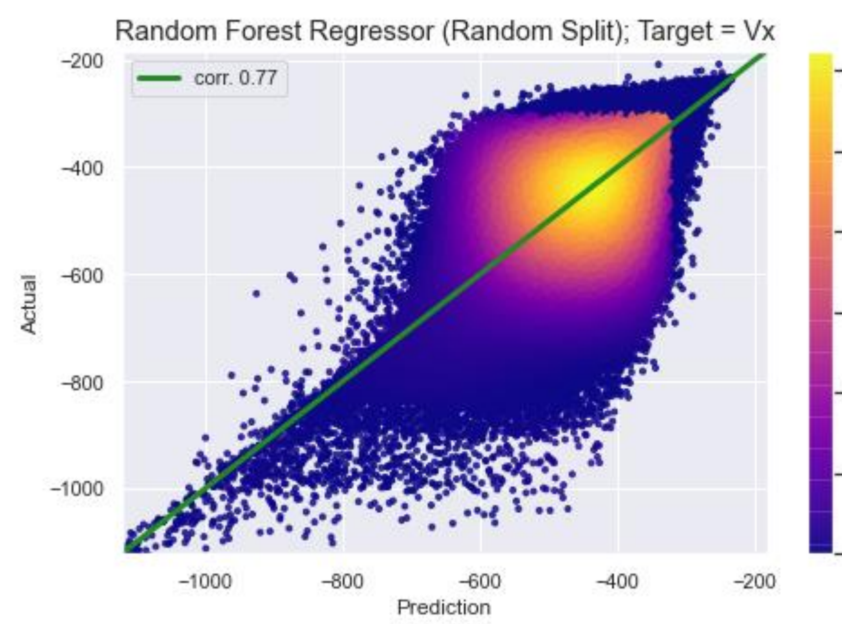
Fig 4: Figure to picture idea of sequential split.

Random split + time history included for the input parameters.

RESULTS

ML Regression Types	Overall Model (R2 Score)	Vx (R2 Score)	Vx (RMSE)	Vy (R2 Score)	Vy (RMSE)	Vz (R2 Score)	Vz (RMSE)	Np (R2 Score)	Np (RMSE)	Temp (R2 Score)	Temp (RMSE)
Linear/Polynomial (Random)	0.11719	0.25099	91.32519	0.00931	25.47781	0.00135	22.66787	0.16680	4.54221	0.15748	90282.11515
Linear/Polynomial (Sequential)	0.10641	0.23991	85.42058	0.01236	25.26472	-0.00603	21.68600	0.14152	4.94505	0.14430	85303.40491
Random Forest (Random)	0.51088	**0.59041**	67.53392	0.40814	19.69264	0.39475	17.64711	**0.63925**	2.98879	0.52186	68012.33580
Random Forest (Sequential)	0.06634	0.20068	87.59700	-0.03861	25.90848	-0.08202	22.49159	0.24934	4.62115	0.00233	92108.31224

Table 3 (Left): Performance evaluation of the Machine Learning models with the two different train-test split types.



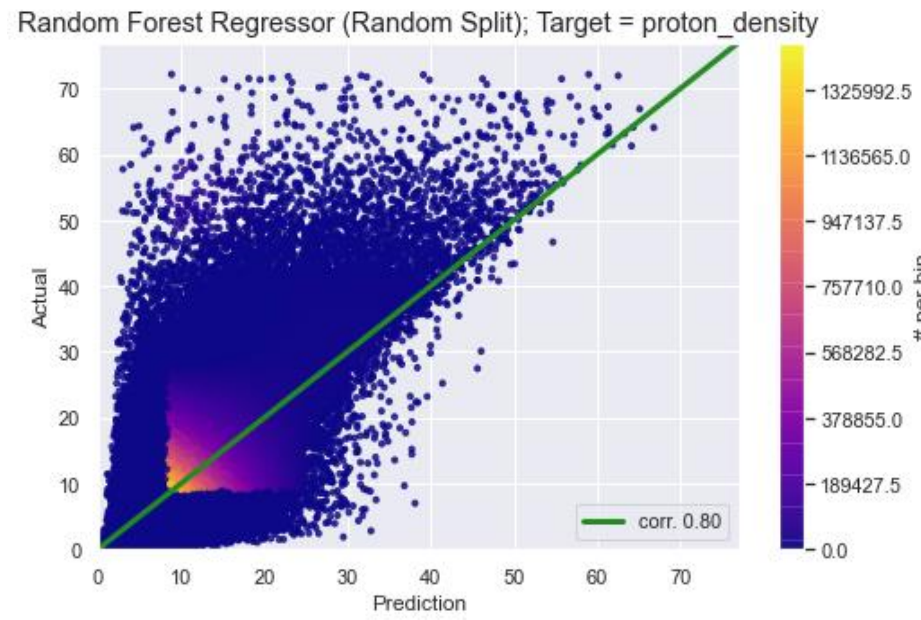
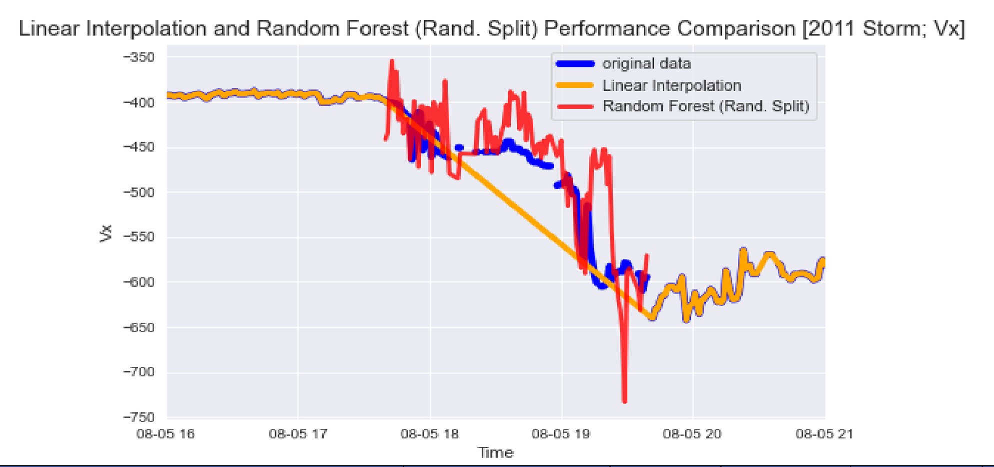


Fig 5 & Fig 6 (Right): Scatter plots for Vx and proton density. The plots evaluate the performance of the Random Forest model, trained with the (randomly split) ~20yrs of data, by comparing the predicted results with the actual data values of each plasma parameter. Pearson correlation coefficients (R) are also printed in the legends of the plots.

"Overall Model" refers to the performance of the model to predict *all* the parameters. The Linear and Polynomial models for each split type produced the same results for all parameter estimations.

```
Linear Interpolation (Vx) R2 score: 0.5385668551136159
Linear Interpolation (Vx) RMSE: 43.730289522327794
Random Forest (Rand. Split) (Vx) R2 score: 0.3583886079197973
Random Forest (Rand. Split) (Vx) RMSE: 51.56606508940755
```




```
Linear Interpolation (proton_density) R2 score: 0.32349736607821433
Linear Interpolation (proton_density) RMSE: 5.2632495695487735
Random Forest (Rand. Split) (proton_density) R2 score: -0.2859590546557518
Random Forest (Rand. Split) (proton_density) RMSE: 7.256593376668126
```

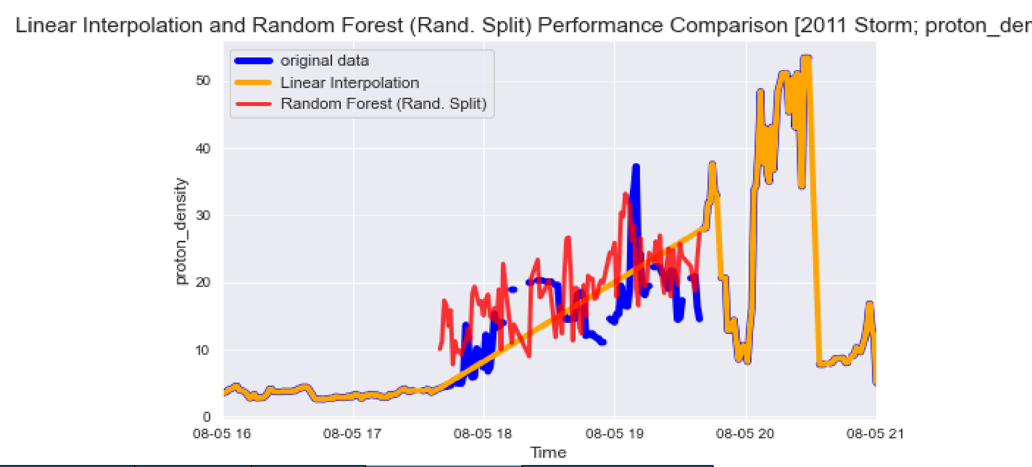


Fig 7(Left) & Fig 8 (Right): Figures to compare the performance of the Random Forest model (with random split) with the Linear Interpolation method for a large data gap (120mins) with large variation.

Random Forest: Random Split	Overall Model (R2 Score)	Vx (R2 Score)	Vx (RMSE)	Vy (R2 Score)	Vy (RMSE)	Vz (R2 Score)	Vz (RMSE)	Np (R2 Score)	Np (RMSE)	Temp (R2 Score)	Temp (RMSE)
No Time History	0.51088	**0.59041**	67.53392	0.40814	19.69264	0.39475	17.64711	**0.63925**	2.98879	0.52186	68012.33580
With Time History	0.54865	**0.65078**	62.63838	0.41640	19.30022	0.42198	16.95320	**0.63860**	3.13208	0.61550	60288.46100

Table 4 (Left): Performance evaluation of the Random Forest-random split model with and without time history

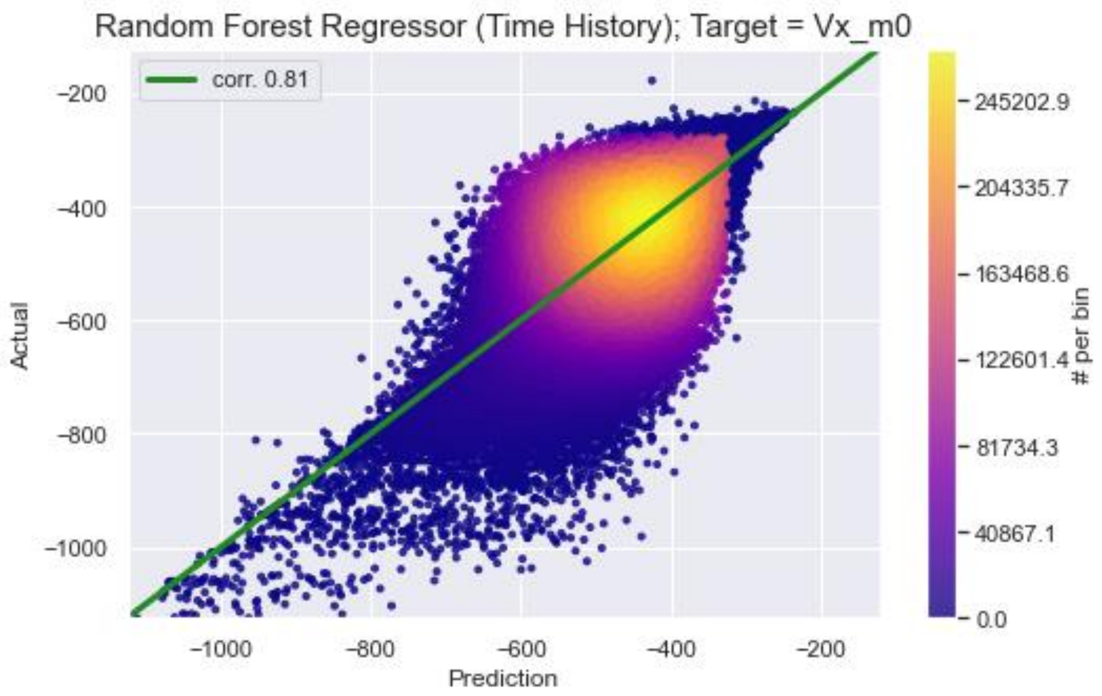


Fig 9 & Fig 10 (Left): "Actual vs. Prediction" scatter plots of Vx and proton density for the model that includes the time history for the input parameters

CONCLUSIONS

Machine Learning Model Results

- The Random Forest Model with random split performed the best.
  - Not only for Vx and proton density, but with all parameters.
  - Machine Learning models (Random Forest) do better for larger (120 mins) data gaps with more extreme variation. Which are important for prediction models that rely on data from large geomagnetic storms.
  - There were improvements to the model with the inclusion of time history data

Interpolation vs. Random Forest Model

- While the scores and RMSE may not necessarily be better than those of linear interpolation in Fig 7 & 8, we can see the model better simulates the dynamic nature of the target parameters

FUTURE WORK & IMPACT

- Working on providing the code open source for community use through GitHub
- Improvements to the model
  - Optimized Random Forest hyperparameters
  - Neural Network
- Used the improved input for GIC prediction models
- This tool would have immediate impact on
  - case studies with numerical models
  - machine learning models

REFERENCES

[1] Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 51–6.
[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay; 12(85):2825−2830, 2011.
[3] Keesee AM, Pinto V, Coughlan M, Lennox C, Mahmud MS and Connor HK, Comparison of Deep Learning Techniques to Model Connections Between Solar Wind and Ground Magnetic Perturbations. 2020. Front. Astron. Space Sci. 7:550874. doi: 10.3389/fspas.2020.550874
[4] King, J.H. and N.E. Papitashvili, Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data, J. Geophys. Res., 110, A02104, 2005.
[5] Lotz, S. I., and P. J. Cilliers (2015), A solar wind-based model of geomagnetic field fluctuations at a middle latitude station, Adv. Space Res., 55, 220–230, doi:10.1016/j.asr.2014.09.014.

Acknowledgements and Data statement
We gratefully acknowledge the following Python libraries: Nasaomnireader, Scipy, Matplotlib, NumPy, Sklearn, Pandas, Seaborn
The data used in this study is taken from the OMNI and Kyoto World GDC Databases.
This project is funded by NSF-EPSCoR Award OIA-1920965.

contact: jrkobayashi@alaska.edu