JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Machine Learning Applications on Earth's Magnetotail

Savvas Raptis[1], Connor O' Brien[2,1], Kareem Sorathia[1], Viancheslav Merkin[1], Louis Richard[3], Simon Wing[1]

[1] APL/JHU, Laurel, MD, US
[2] Center for Space Physics, Boston University, Boston, MA, USA
[3] Swedish Institute of Space Physics, Uppsala, Sweden

Supported by John Hopkins University Applied Physics Laboratory independent R&D fund and NASA Drive Center CGS

savvas.raptis@jhuapl.edu
https://savvasraptis.github.io

CGS | CENTER FOR GEOSPACE STORMS

Goal of Presentation

Spark a discussion on problems we have with data-driven modeling.

🔹 **Evaluating Models**
What do our typical metrics really tell us about predictive power?

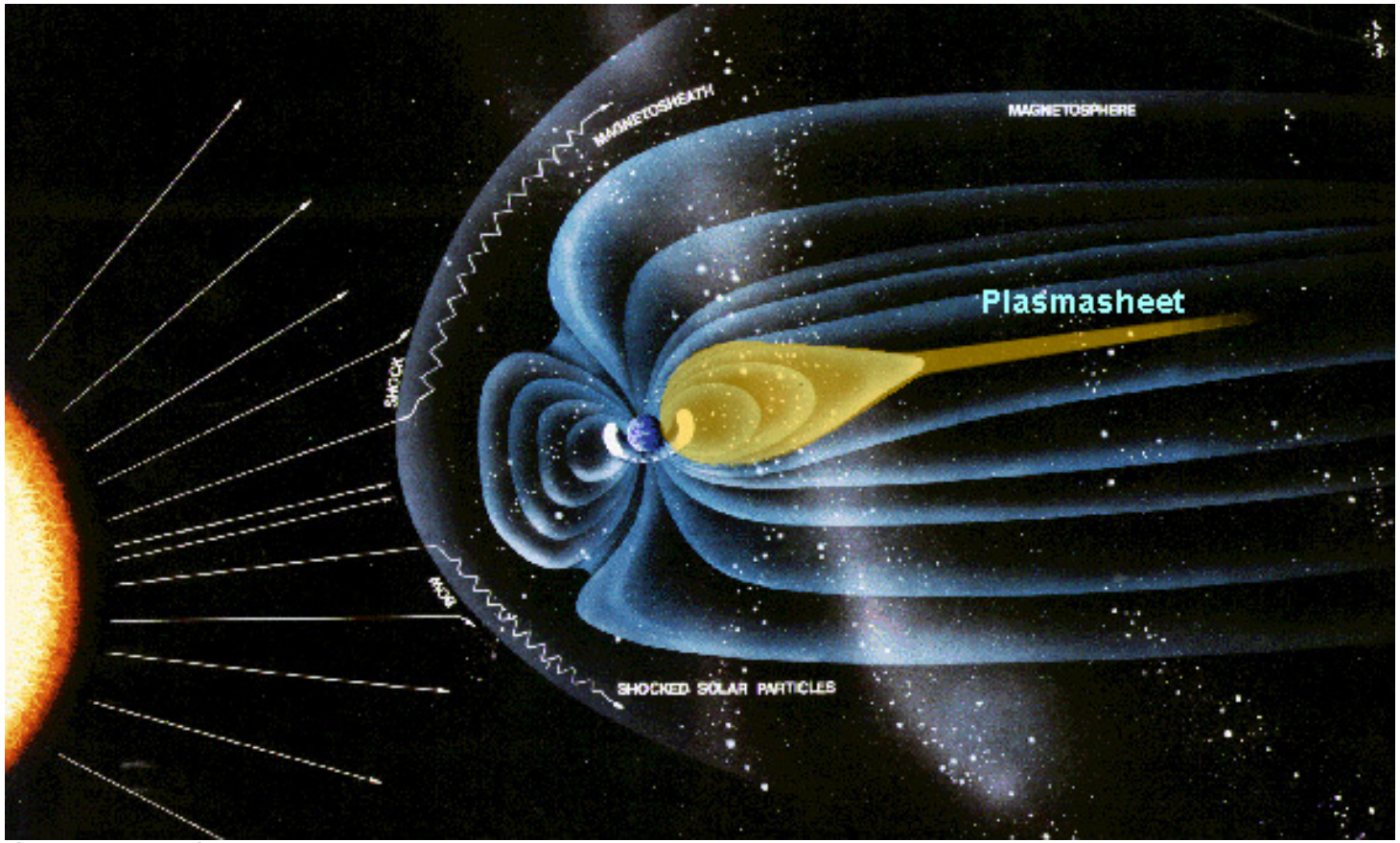
🔹 **Outliers & Sampling**
How should we treat rare events? Can we expand or restrict our dataset meaningfully?

🔹 **Ground Truth in Space**
How do we use in-situ data, and what counts as “truth” relative to model performance?

🔹 **Model Architectures**
Do architectures always drive performance? or do data and physics matter more?

All these questions will be discussed through the challenge of predicting Earth’s plasma sheet properties from solar wind and geomagnetic conditions

Earth's plasma sheet



Credits: NASA

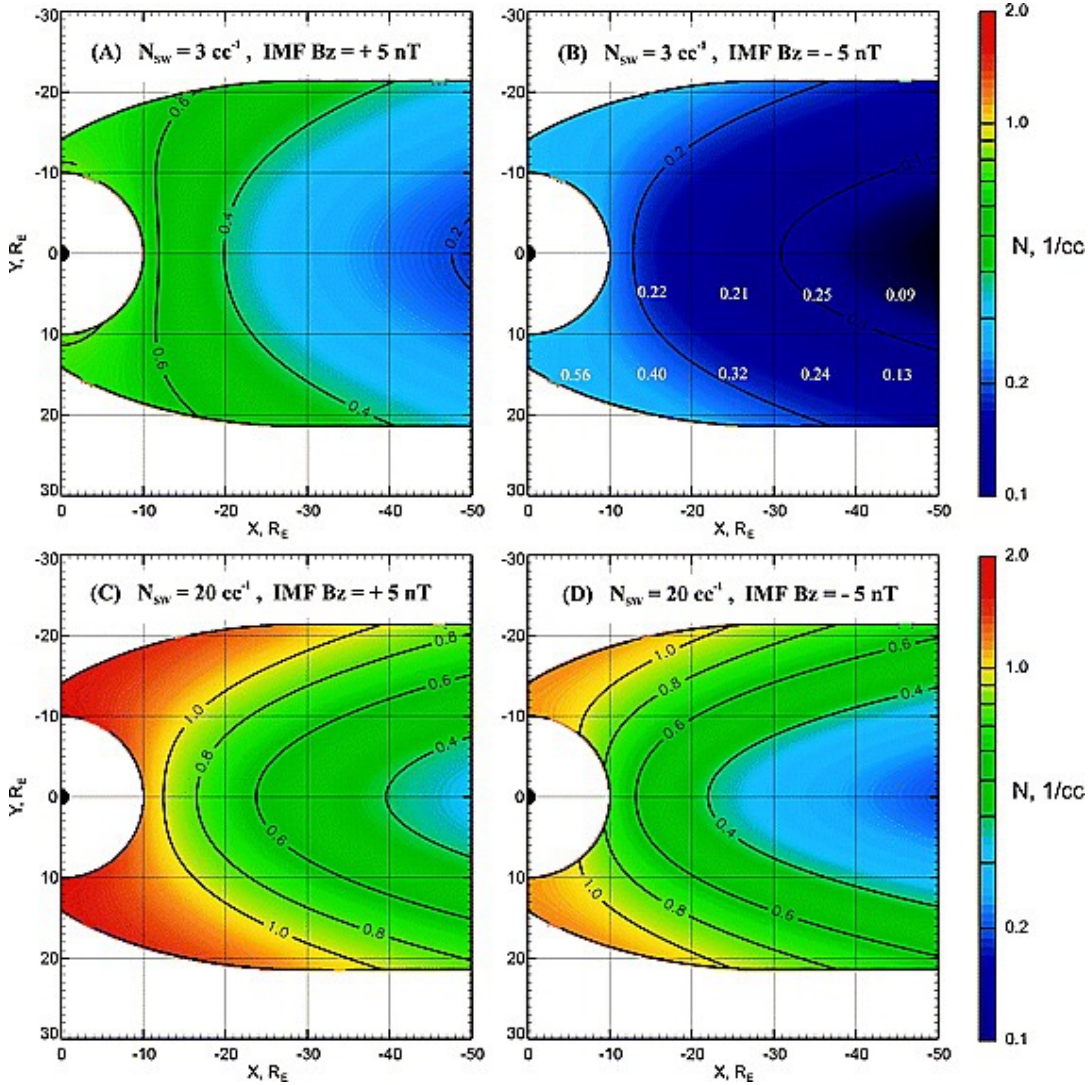
- Magnetotail reconnection
- Bursty Bulk Flows (BBFs)
- Global Convection Patterns
- Ring current

<u>Modeling PS is useful for:</u>

(a) Understanding storm/substorm dynamics
(b) Explain ring current configuration
(c) Facilitate space weather modeling
(d) Understand inner magnetosphere
(e) Source for radiation belts

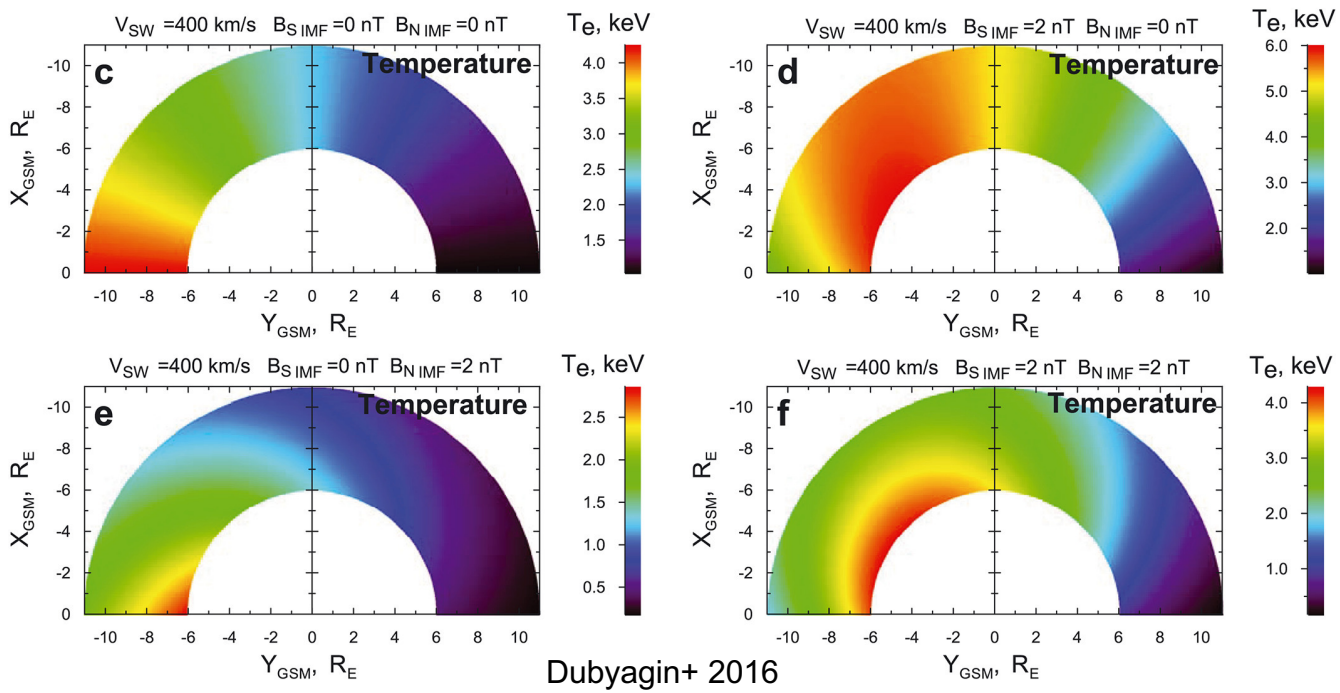
Baseline empirical models

Modelled with Geotail



Tsyganenko & Mukai 2003

Modelled with THEMIS

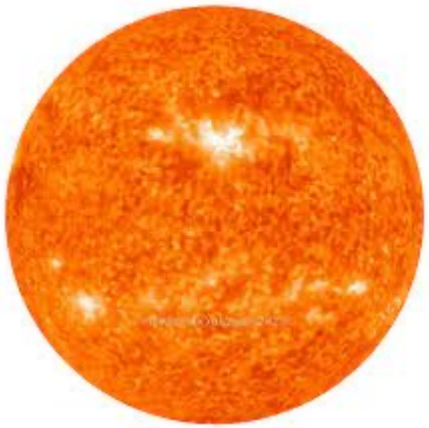


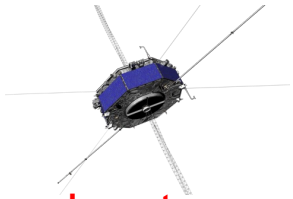
Dubyagin+ 2016

Why then work on this?

1. More data under different conditions
2. MMS was never used with its state of the art instrumentation
3. These models don't include time history
4. ML methods can reveal non-linear relationships easily

Where are we & what are we doing?





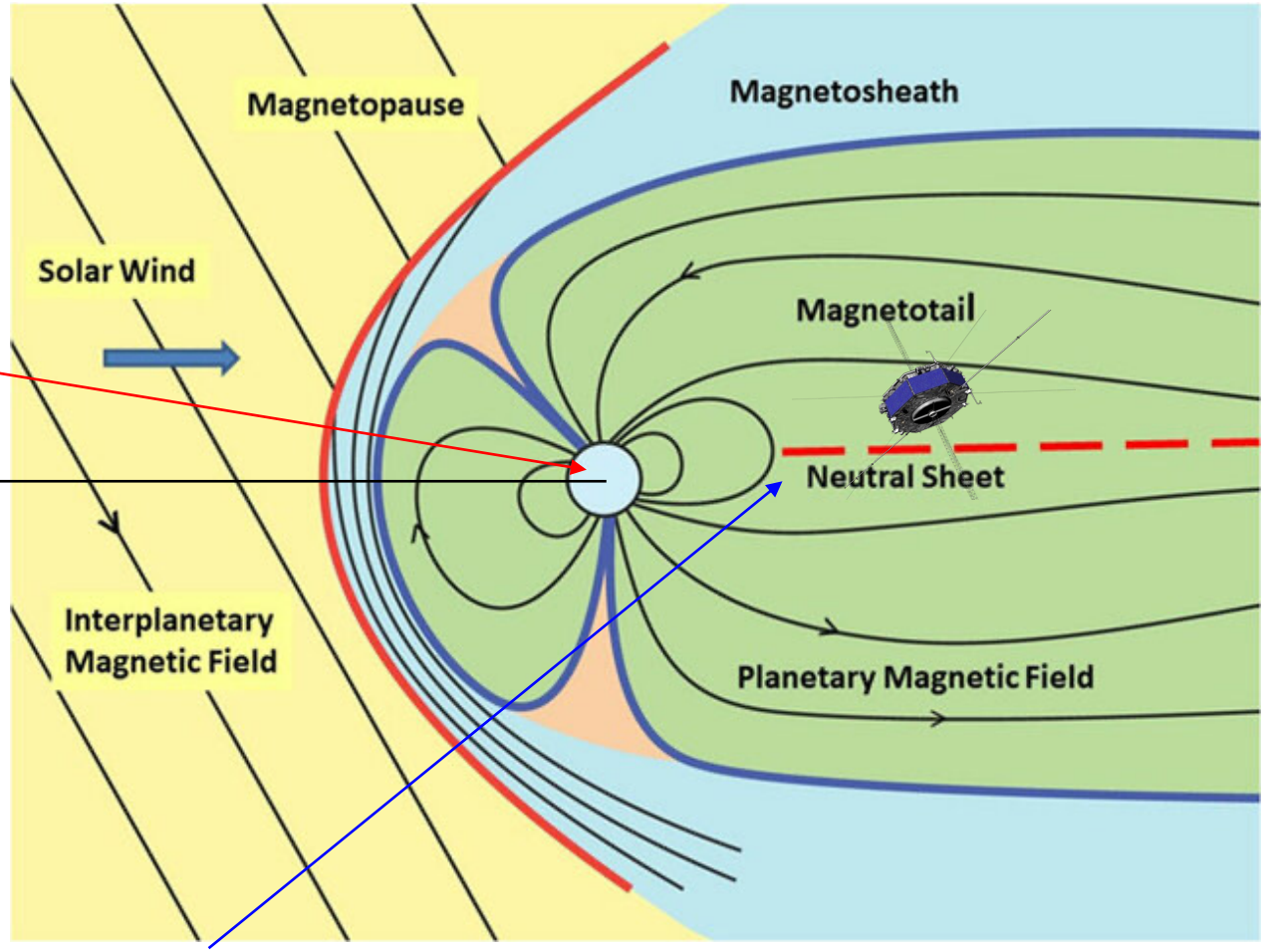
Input:

ACE/WIND/DSCVR

(L1)

$200R_E$

+Geomagnetic

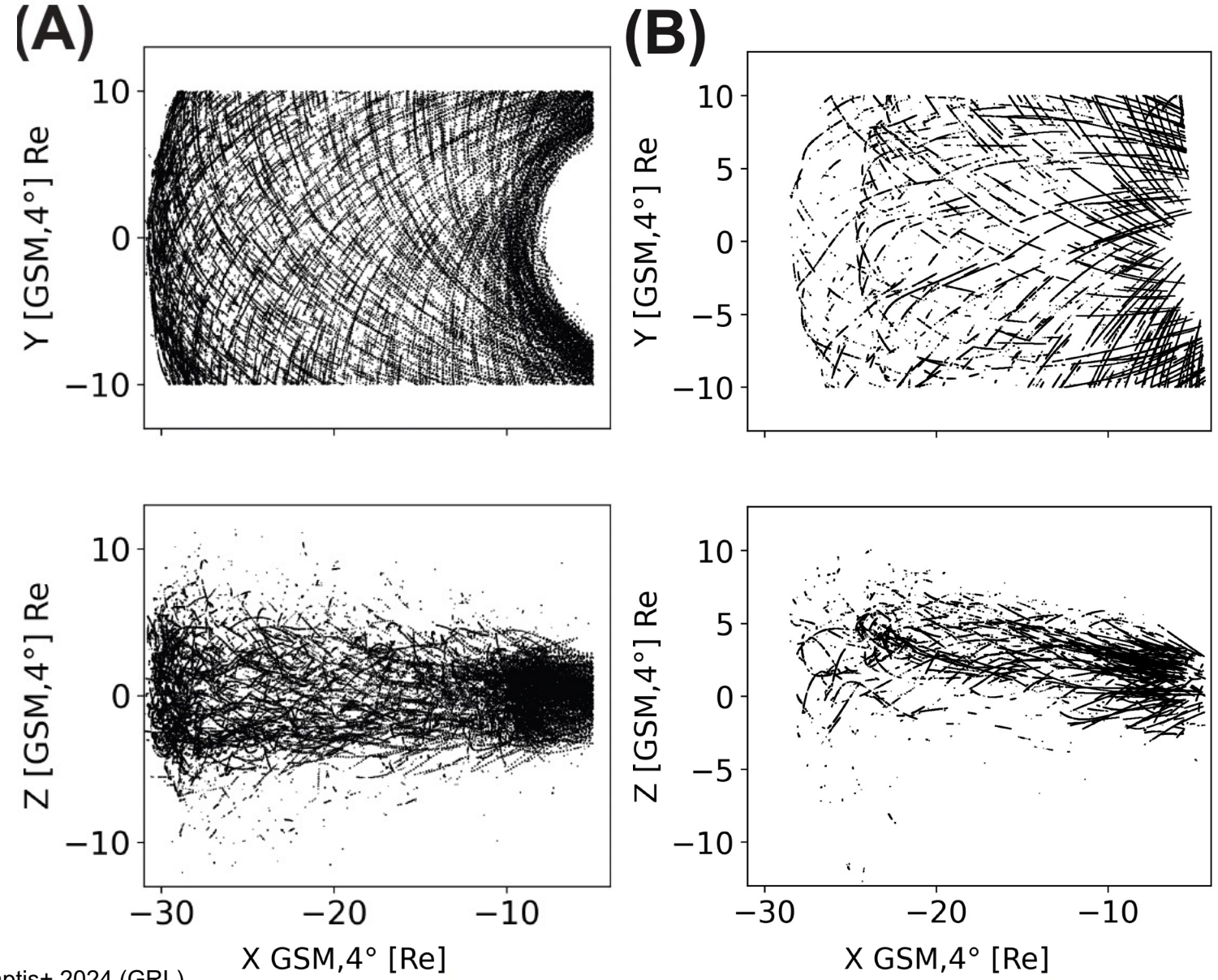


Output: *In-situ* properties from: MMS, Geotail, THEMIS

Goal: Model Plasmasheet properties based on driving (SW) and geomagnetic conditions

$R_E = 6371.2\ [km]$

The dataset (output – Central Plasma Sheet)



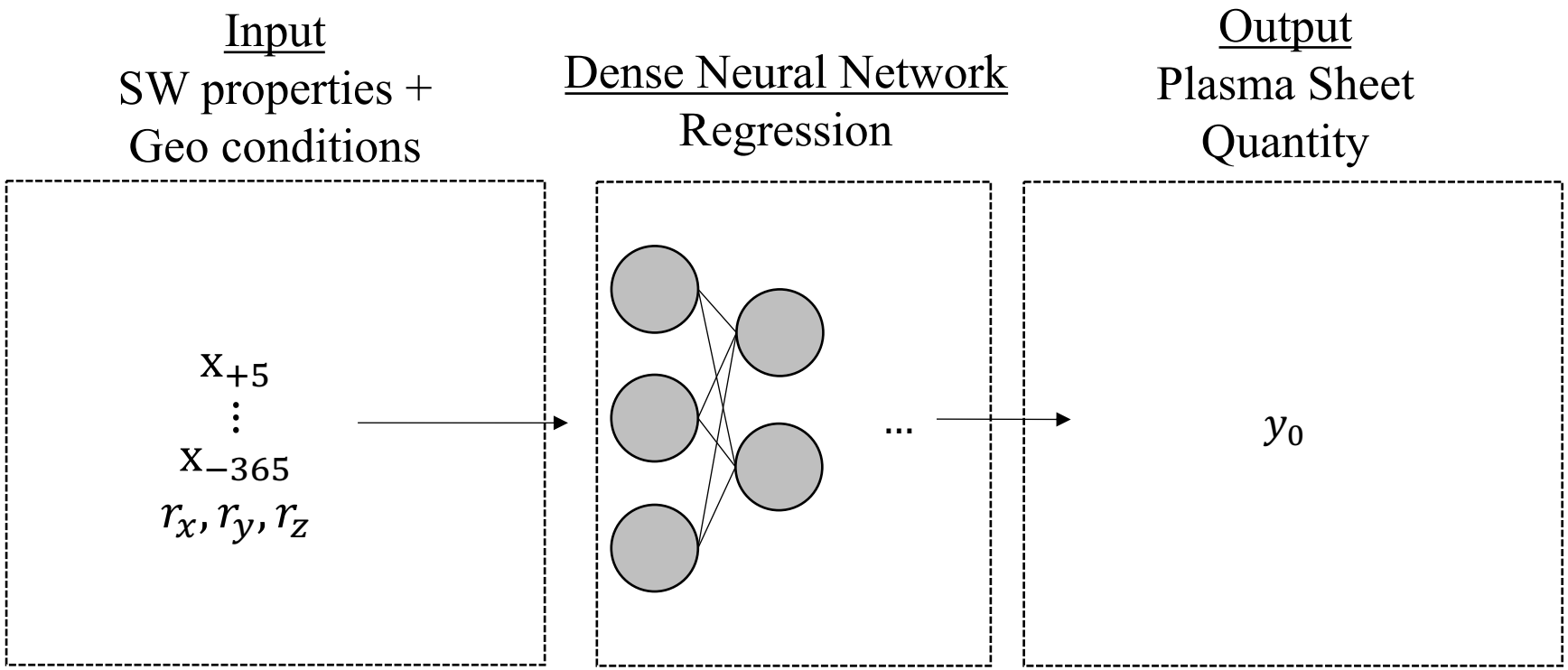
Raptis+ 2024 (GRL)

(A) Geotail (1994 - 2022)
>1 million points (~12s res)

(B) MMS (2015 – 2024)
~ 250k points (~12s res)

Output:
Anything locally measured
(In this example plasma moments)

Data Scientist POV
(i.e., Input, output & regression)



Input:
x: Different solar wind features (e.g., n, B, etc.) + geomagnetic indices including time history up to 6h
r: Location of SC measuring output

Output:
y: Different quantities at plasma sheet (e.g., n, B, T etc.)

SW Input tried: Wind (1min res) – OMNIweb (1/5 min res)

Results here shown for 1min averaged quantities for output

Metrics & Evaluations

Quick Reminders

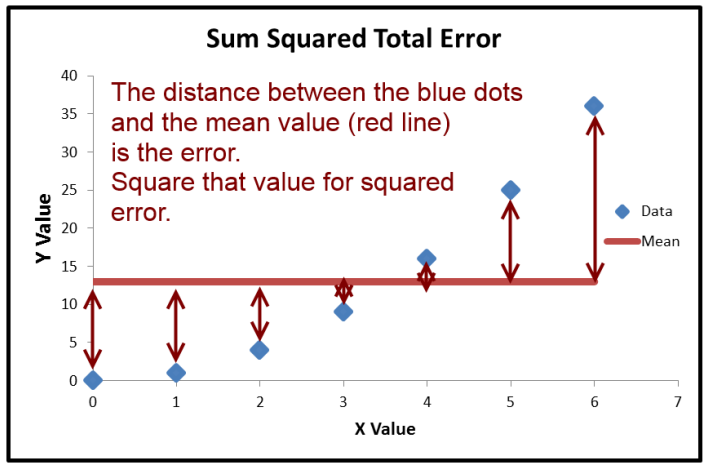
What are the metrics we need to use

```
...  99/99 ━━━━━━━━━━━━━━━━━━━━ 0s 808us/step
     explained_variance:  0.019
     median absolute error:  0.11
     r2:  -0.01
     MAE:  0.157
     MSE:  0.055
     RMSE:  0.235
     Cor:  0.533
```

A complex and intriguing model

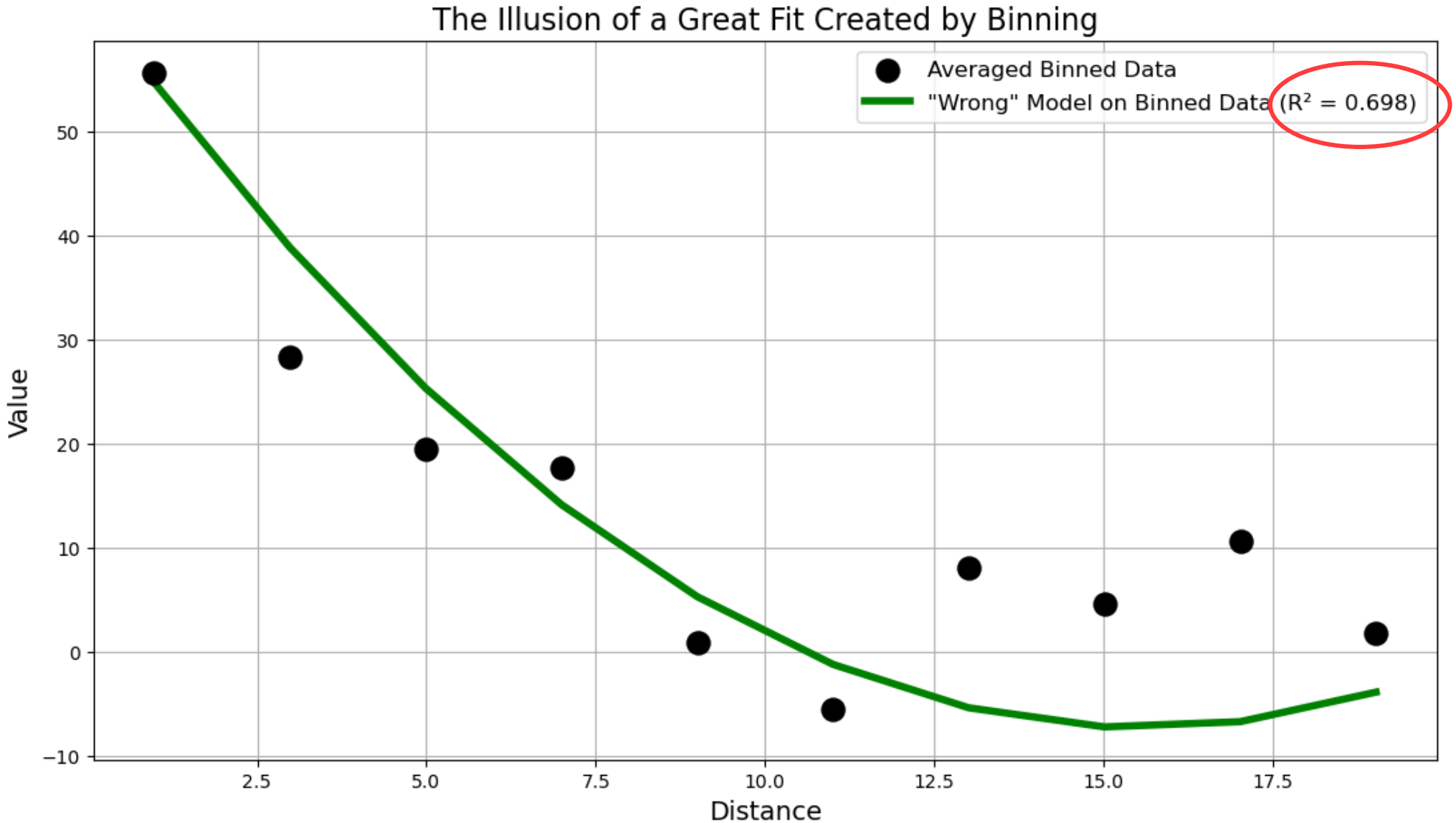

```
...  99/99 ━━━━━━━━━━━━━━━━━━━━ 0s 778us/step
     explained_variance:  0.0
     median absolute error:  0.14
     r2:  0.0
     MAE:  0.17
     MSE:  0.055
     RMSE:  0.234
     Cor:  0.0
```

np.mean()



When $R2<0$, the horizontal line explains the data better than your model (i.e., mean of observed).

The illusion success by spatially binning our data



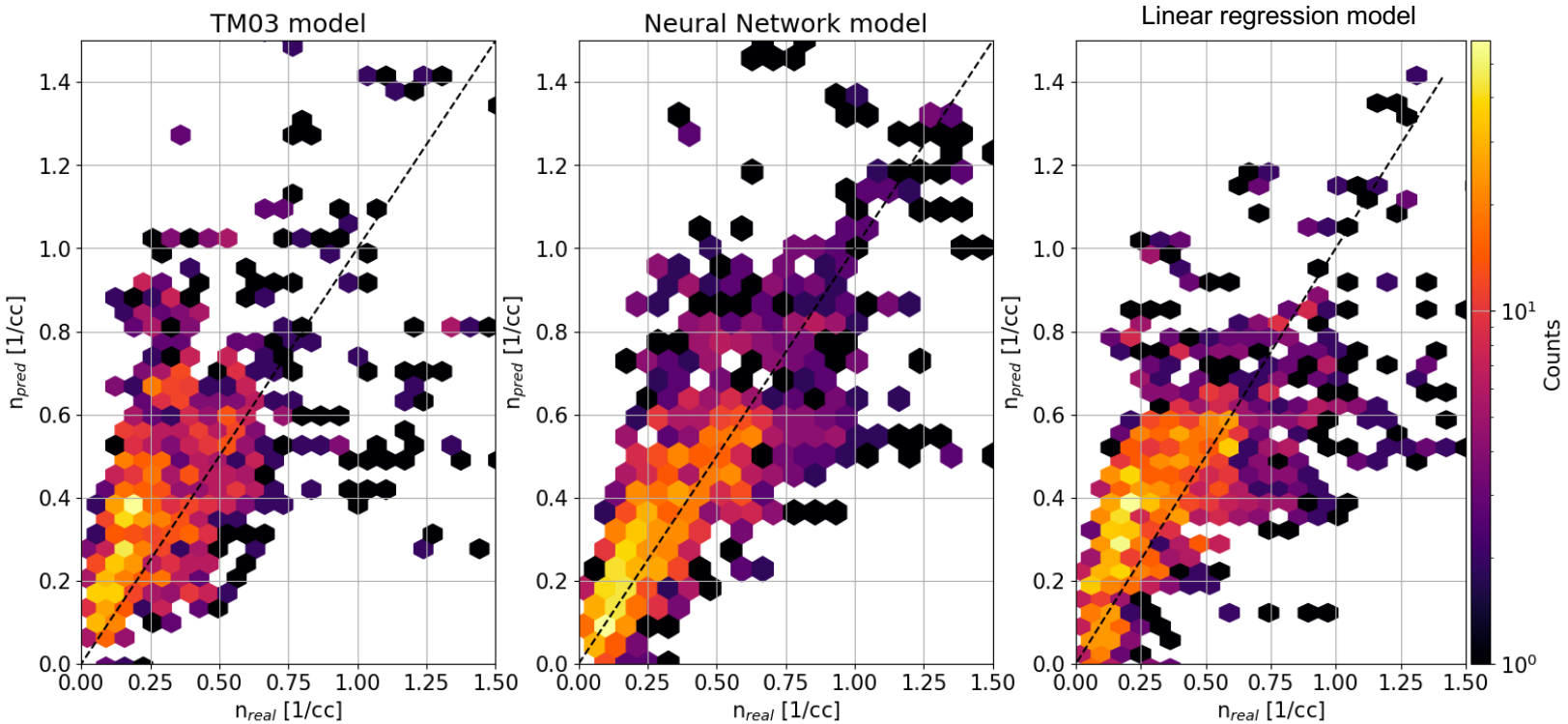
Underlying process is structured, but the measurements or local variations are noisy.

Actual Results

Modeling Density | Predictions vs Observations

Model maximizing correlation for input and output (replace for linear regression)

Key Message: NN > Base ≥ TM03



	TM03	**NN**	Base
R2	0.17	**0.68**	0.32
MAE	0.19	**0.11**	0.18
RMSE	0.27	**0.2**	0.27
r (cor)	0.58	**0.83**	0.57

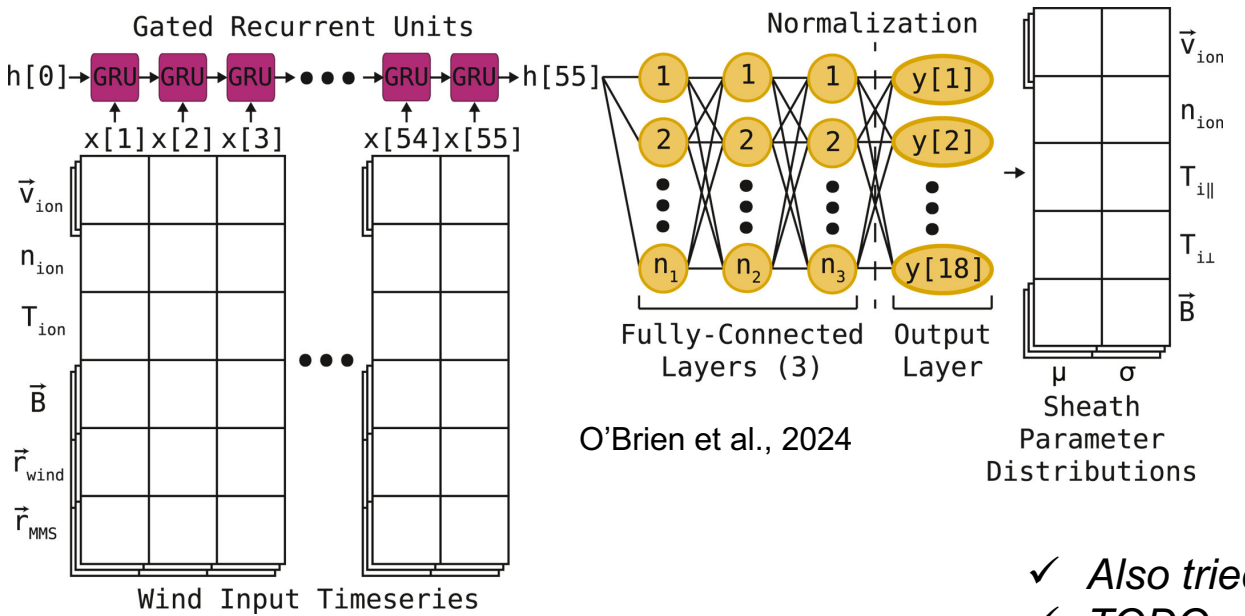
Geotail data

- Presented Testing of NN → Prone to data leakage
- Harder test set (i.e., 5 years of out of sample test data) gives R2 ~0.3-0.4

Raptis, O'Brien et al., 2024 (under prep.)

More methodologies & input space

- **PRIME**: GRU architecture, non-propagated Wind values tried up to several hours of history time



Time History	Type of Input	Architectures
1-10h	Wind (L1)	Linear Reg
	OMNIweb	Gradient Boosting
		Neural Network
		RNN/LSTM/GRU (PRIME-PS)

- ✓ *Also tried different error functions, optimizers, hyperparameters etc.*
- ✓ *TODO some different imbalanced techniques*

Key Takeaway:

To quantify our method's impact, we tested diverse variations of the problem.

Raptis, O'Brien et al., 2024 (under prep.)

Updated results (Test set, last 20% of data)

Method	Strict CPS			
	MAE	R^2	r	CRPS
LightGBM	0.145	0.242	0.631	–
Neural Net	0.152	0.325	0.603	–
Linear Reg	0.173	0.265	0.620	–
PRIME-PS	**0.113**	**0.453**	**0.707**	0.083
TM03	0.163	0.208	0.570	–

Key Results:

- PRIME-PS demonstrates a performance edge (~30% MAE from TM03 and ~15% from other ML approaches).
- This advantage can get quite low (from cross-validation | not shown).
- Different input, method, time-history, and hyperparameter tuning etc. had overall a statistically marginal effect.
- Why is this the case?

Raptis, O'Brien et al., 2024 (under prep.)

Feature Importance Analysis

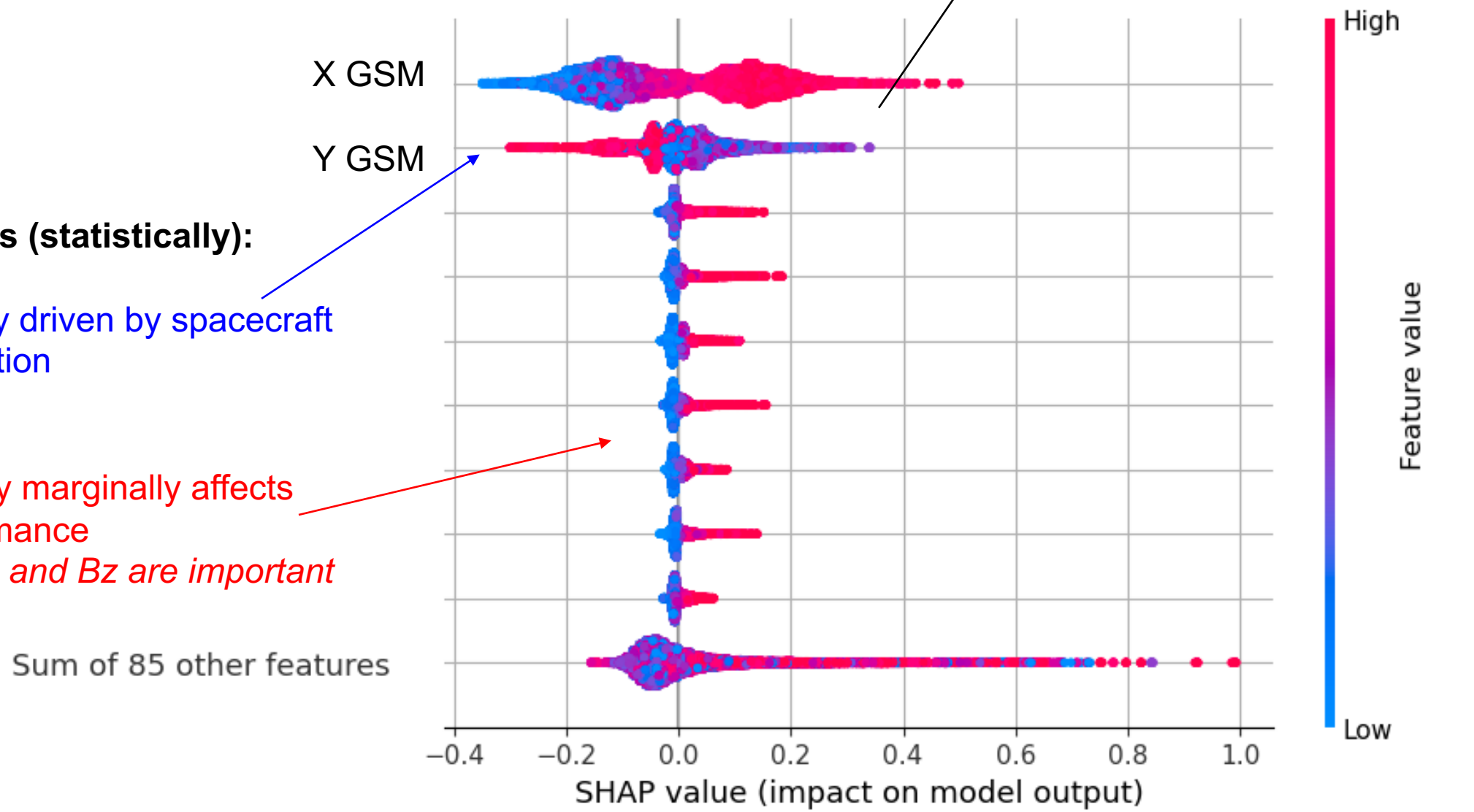
Higher density close to earth and at dawn

Answer: In most cases (statistically):

Model is predominantly driven by spacecraft location

Solar wind input only marginally affects performance

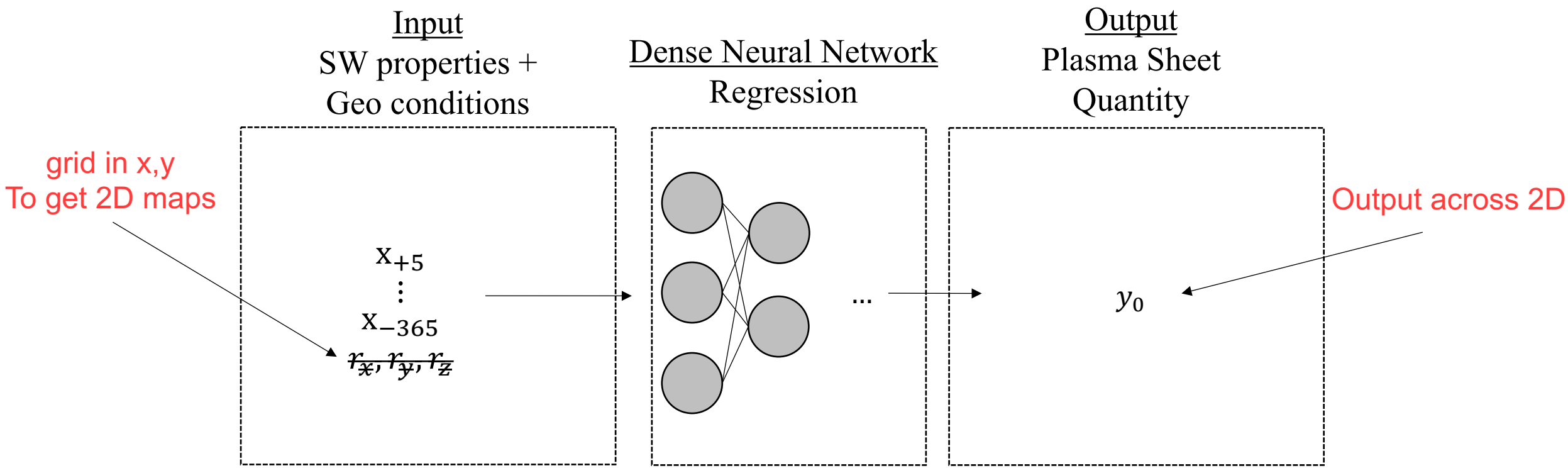
Storms: 1h history of n and Bz are important



SHAP Values explain why a model made a specific prediction, by showing each feature's impact.

Modeling Efforts

Next step: 2D modeling



Input:
x: Different solar wind features (e.g., n, B, etc.) + geomagnetic indices including time history up to 6h
r: Location of SC measuring output

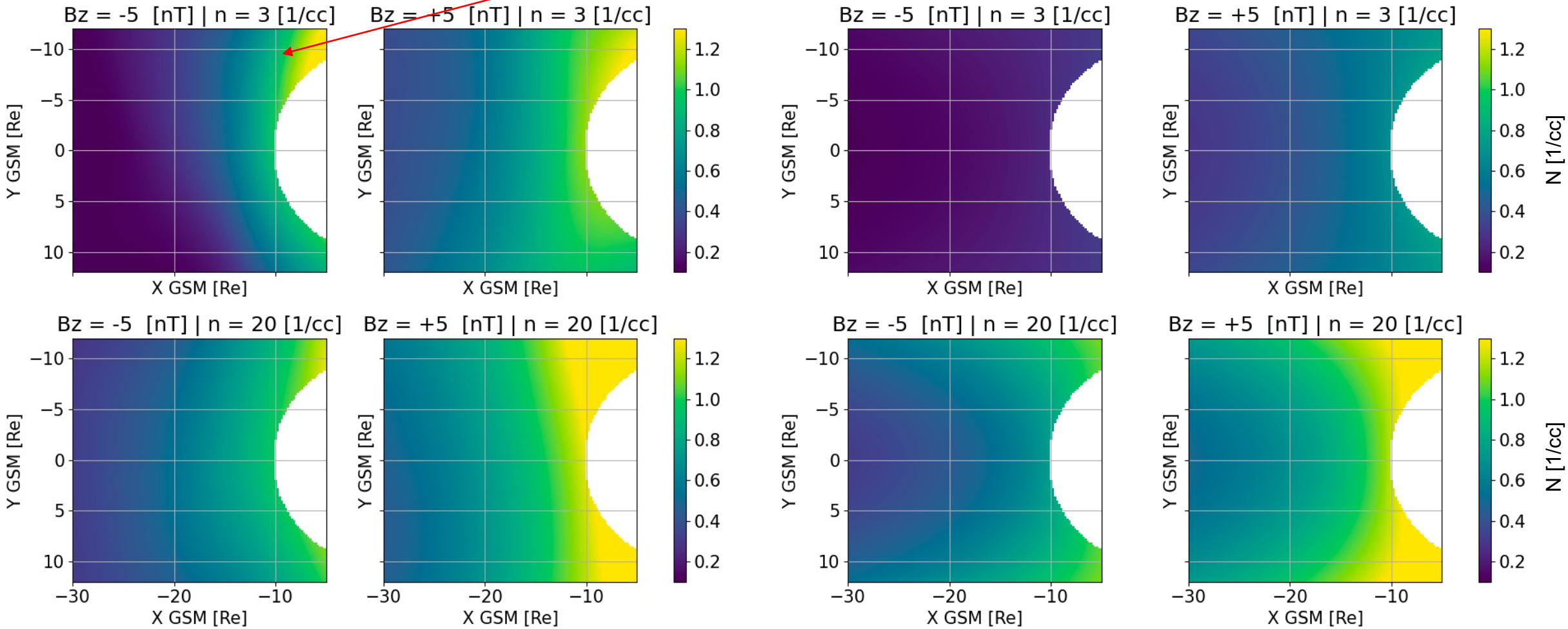
Output:
y: Different quantities at plasma sheet (e.g., n, B, T etc.)

SW Input tried: Wind (1min res) – OMNIweb (1/5 min res)

Results here shown for 1min averaged quantities for output

Modeling Density | 2D Maps

Asymmetries introduced



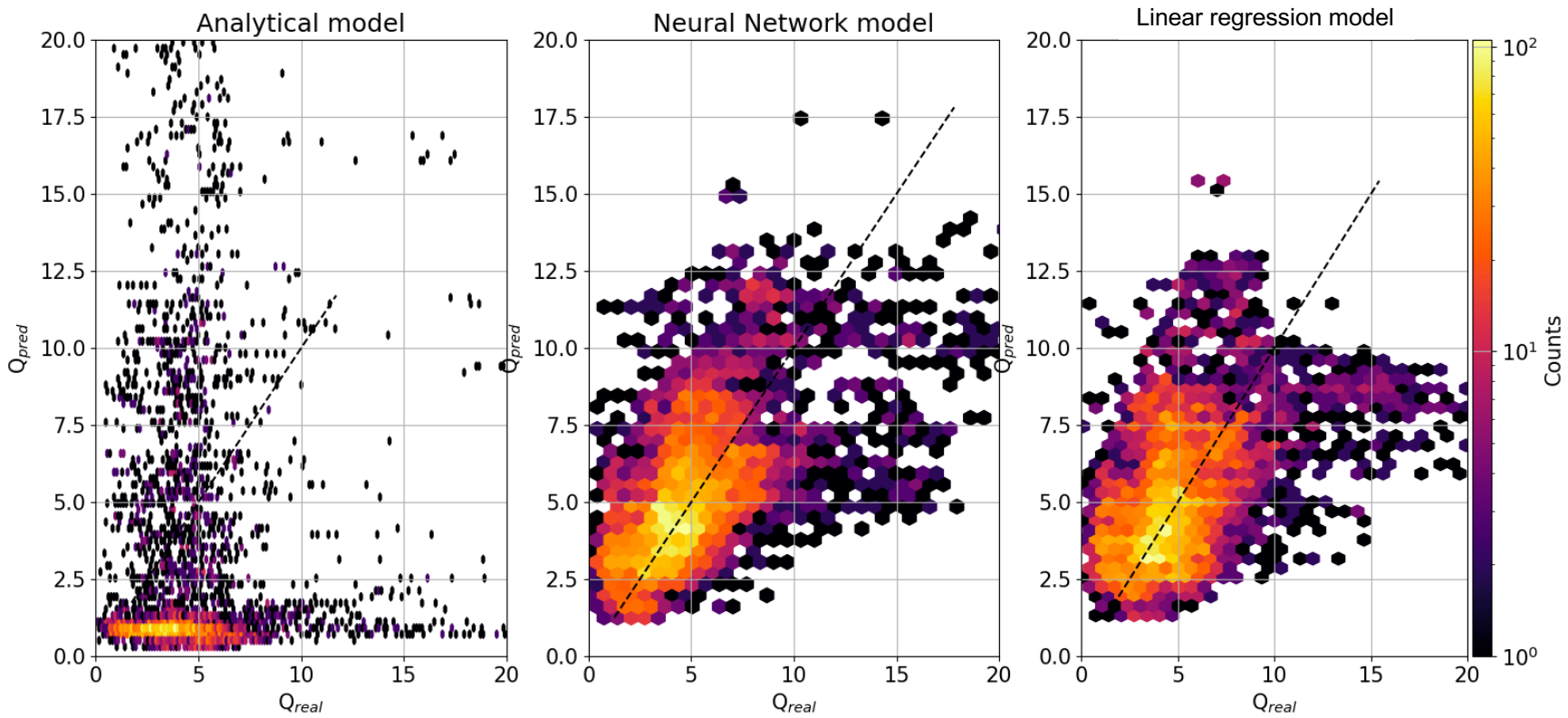
Neural Networks modeling

Empirical modeling (TM03)

Raptis, O'Brien et al., 2024 (under prep.)

How does a more complicate quantity looks like (Ti/Te)

Trying to model Ti/Te in the plasmasheet (Output MMS)



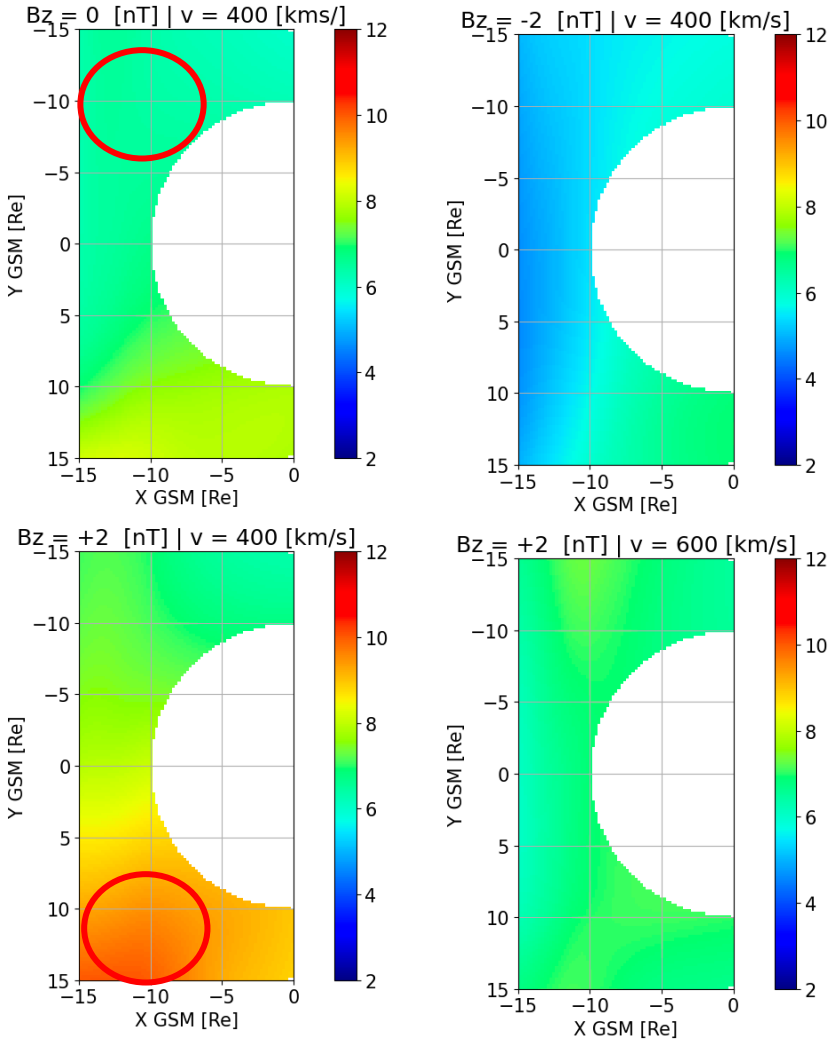
	NN	Linear
R2	**0.32**	0.12
MAE	**1.8**	2.1
RMSE	**2.5**	2.9
r (cor)	**0.6**	0.45

TLDR: Analytical model bad, neural network good? Yes kind of.

Note: To be fair to the analytical models, they are not trained on similar radial distances or datasets (Ti is from Geotail, Te is from THEMIS)

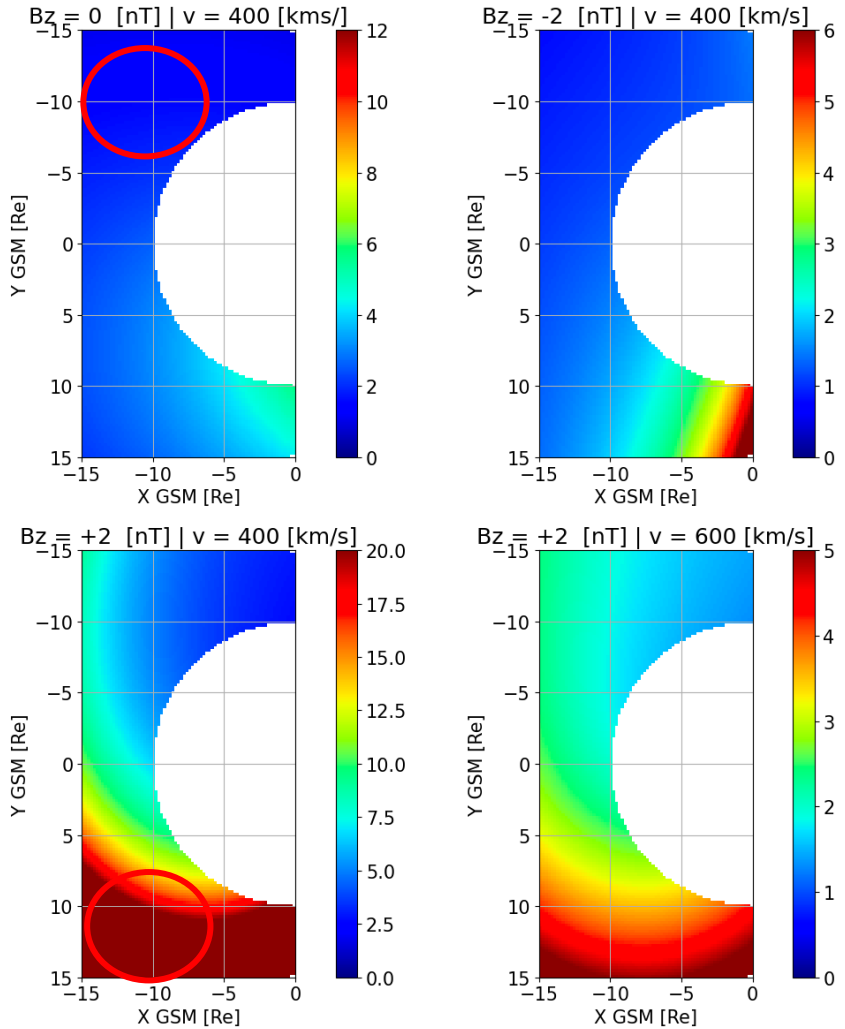
Modeling Temperature Ratios with MMS | 2D Maps

Note: Wang et al., 2009 with dusk Ti/Te much higher than dawn (Using THEMIS data)



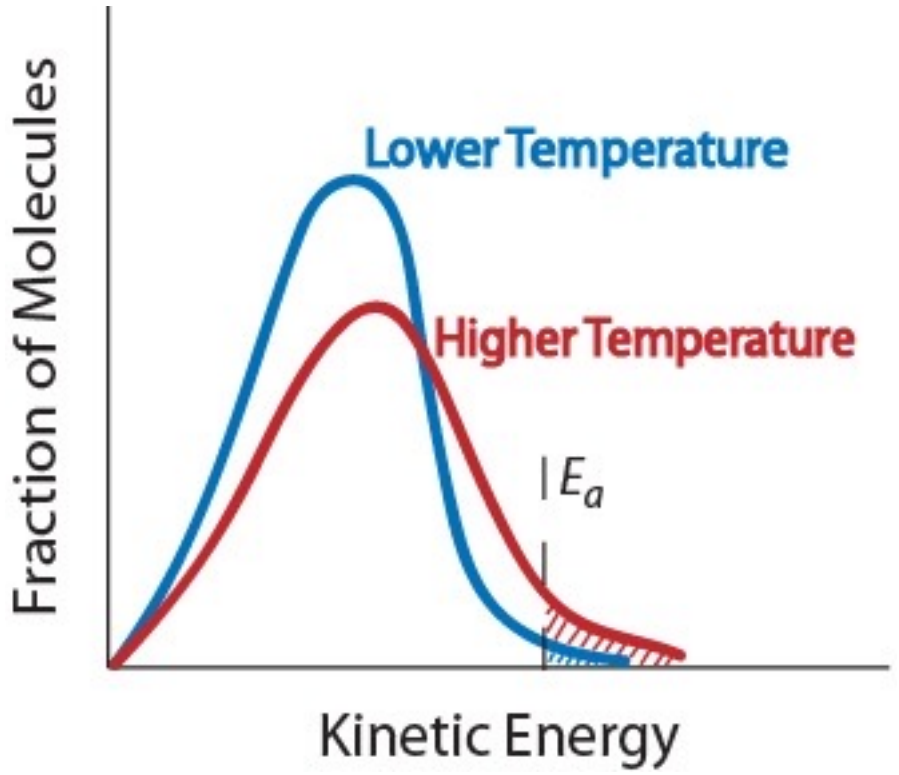
Neural Networks modeling

+No extreme values
+Asymmetries are shown
+ Coherent physical picture



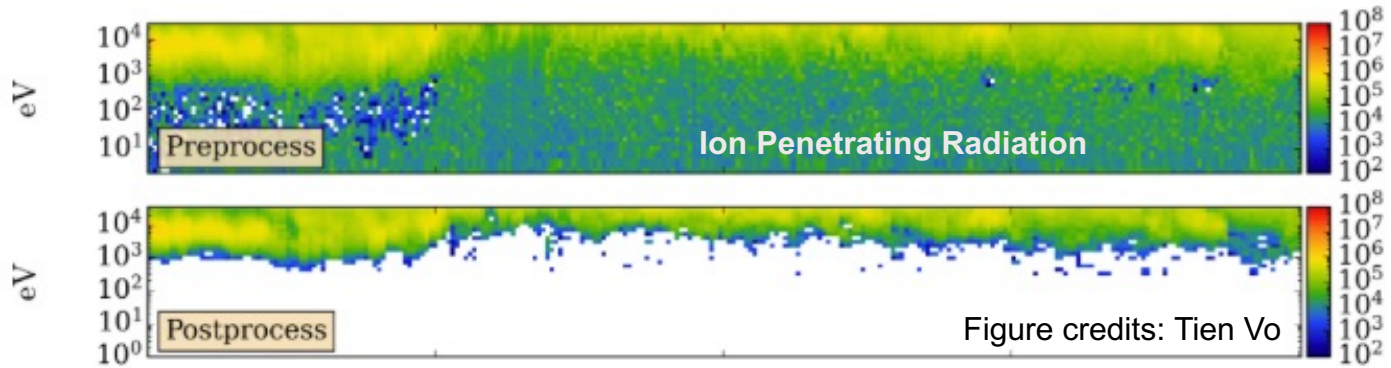
Empirical modeling (TM03/DSGR16)

Community Reminder on Temperature



$$T = \frac{m}{3k_B n}\int (\mathbf{v}-\mathbf{v}_b)\cdot(\mathbf{v}-\mathbf{v}_b) f(\mathbf{v}) d^3v$$

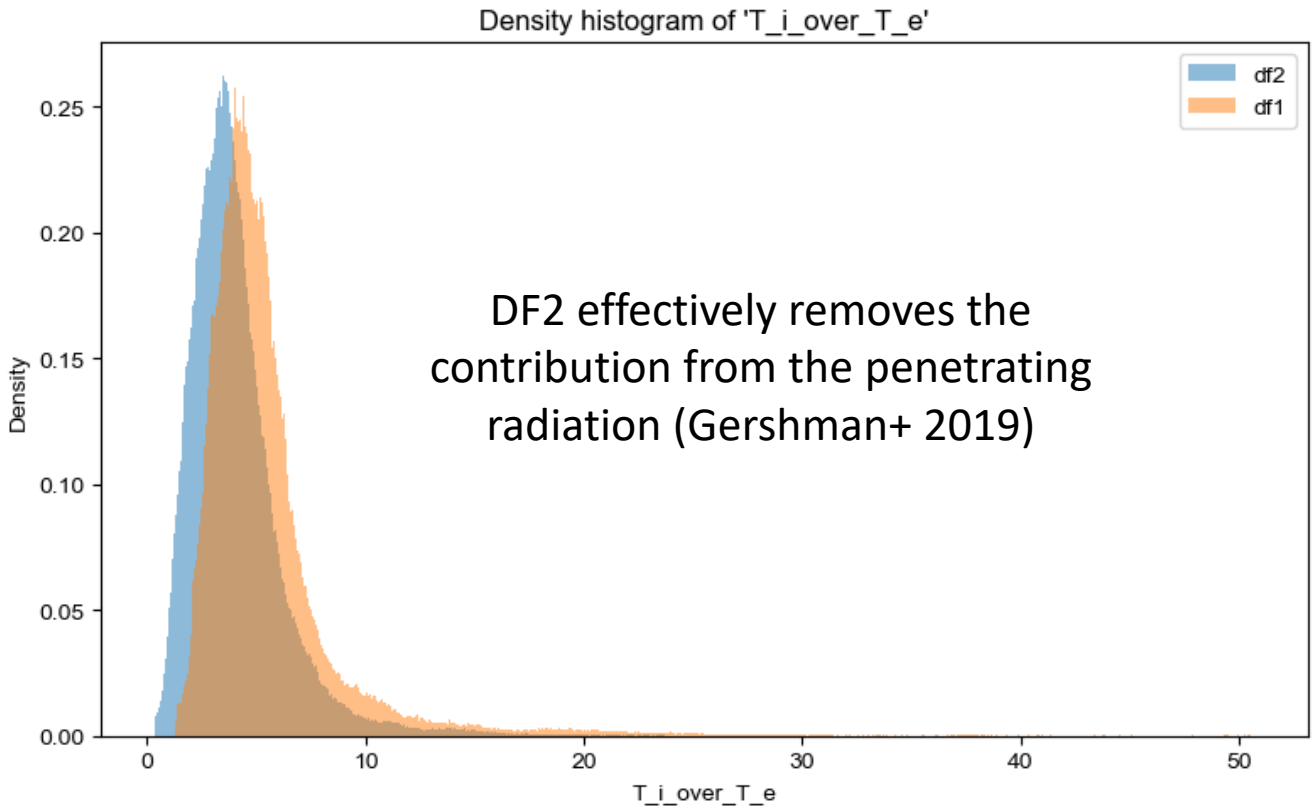
- Temperature is the 2nd plasma moment
- The higher the moment, the more uncertain because you rely more on the poorly sampled tails of the distribution.
- So, 0 and 1st moment (**Density and Velocity**) are usually ok, but Temperature, we got to be careful



The temperature here was completely incorrect, and the velocity increased from about 200 km/s to over 1500 km/s.

MMS Ti/Te plasmasheet ratio example (Full vs Partial moments)

mean df1 (full distribution moments): 5.5513
mean df2 (partial distribution moments): 4.0797



Key Message1: The mean differenced changed by 1.5 (>30%) simply by recalculating moments
Key Message2: A model with +30% is exciting, but we need to know if "*ground truth*" vary by the same magnitude

Criterion	Strict CPS	Flexible CPS	High density
$\beta > 1$	yes	—	—
$\beta > 0.5$	—	yes	—
$\sqrt{B_x^2 + B_y^2} < 2\lvert B_z \rvert$	yes	—	—
$N < 6$	yes	—	—
$N > 6$	—	—	yes
EA1SW0 = EA	yes	yes	yes
$-31 < R_x < -5$	yes	yes	yes
$\lvert R_y \rvert < 15$	yes	yes	yes
$\lvert R_z \rvert < 10$	yes	yes	—
$V_x > -20$	—	—	yes

Table 1. Plasma sheet classification thresholds for the strict CPS, flexible CPS, and high-density subsets. *beta* is the ion plasma beta parameter, density (N) is in 1/cc units, V_x is in km/s, and all the locations ($R_{x,y,z}$) are in Earth radius. The coordinate system for all vectors is the aberrated Geocentric Solar Magnetospheric (GSM) coordinates

Storm Time Behavior and Importance of Outliers

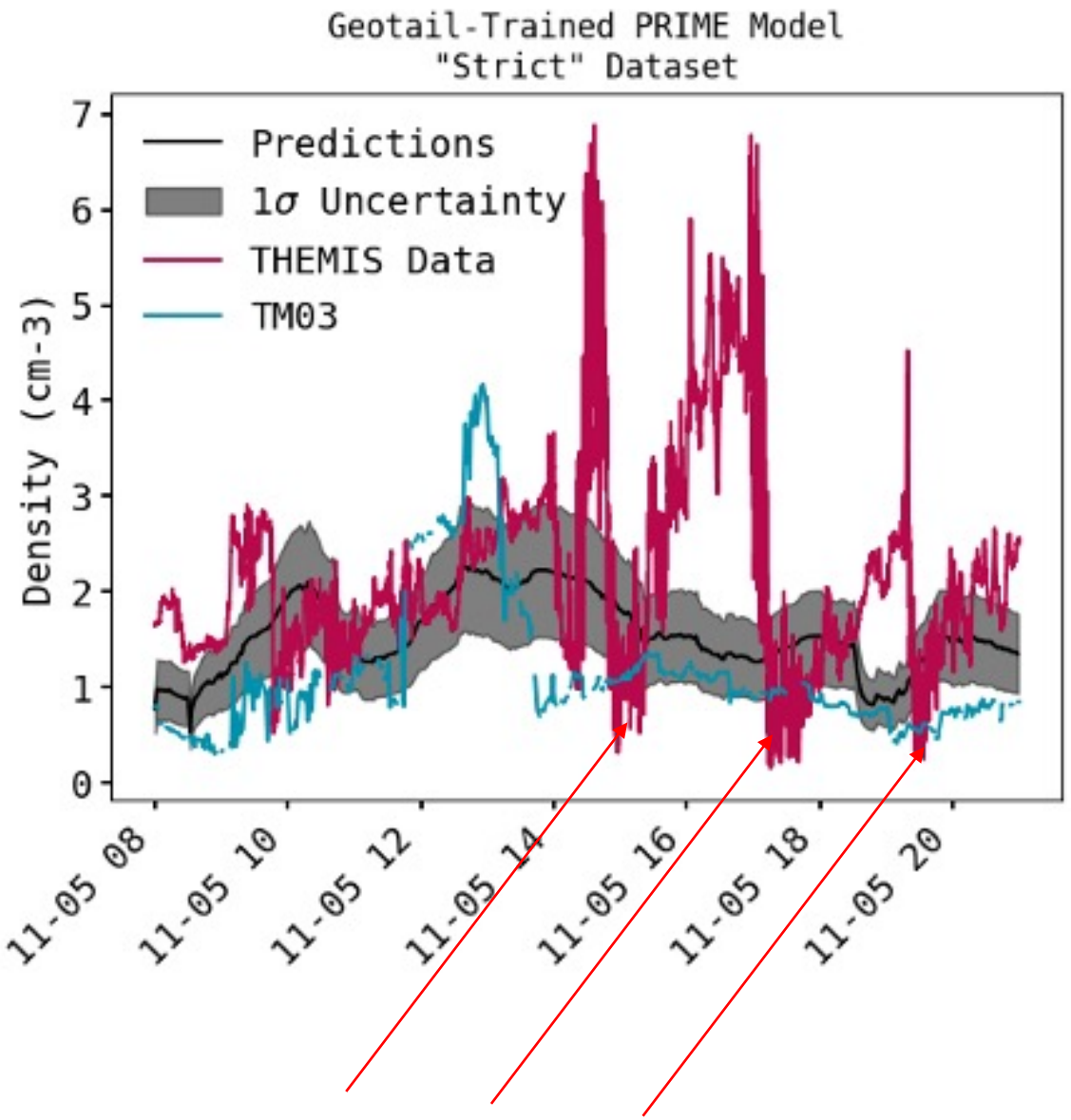
The Problem: We use static thresholds for dynamic environents.

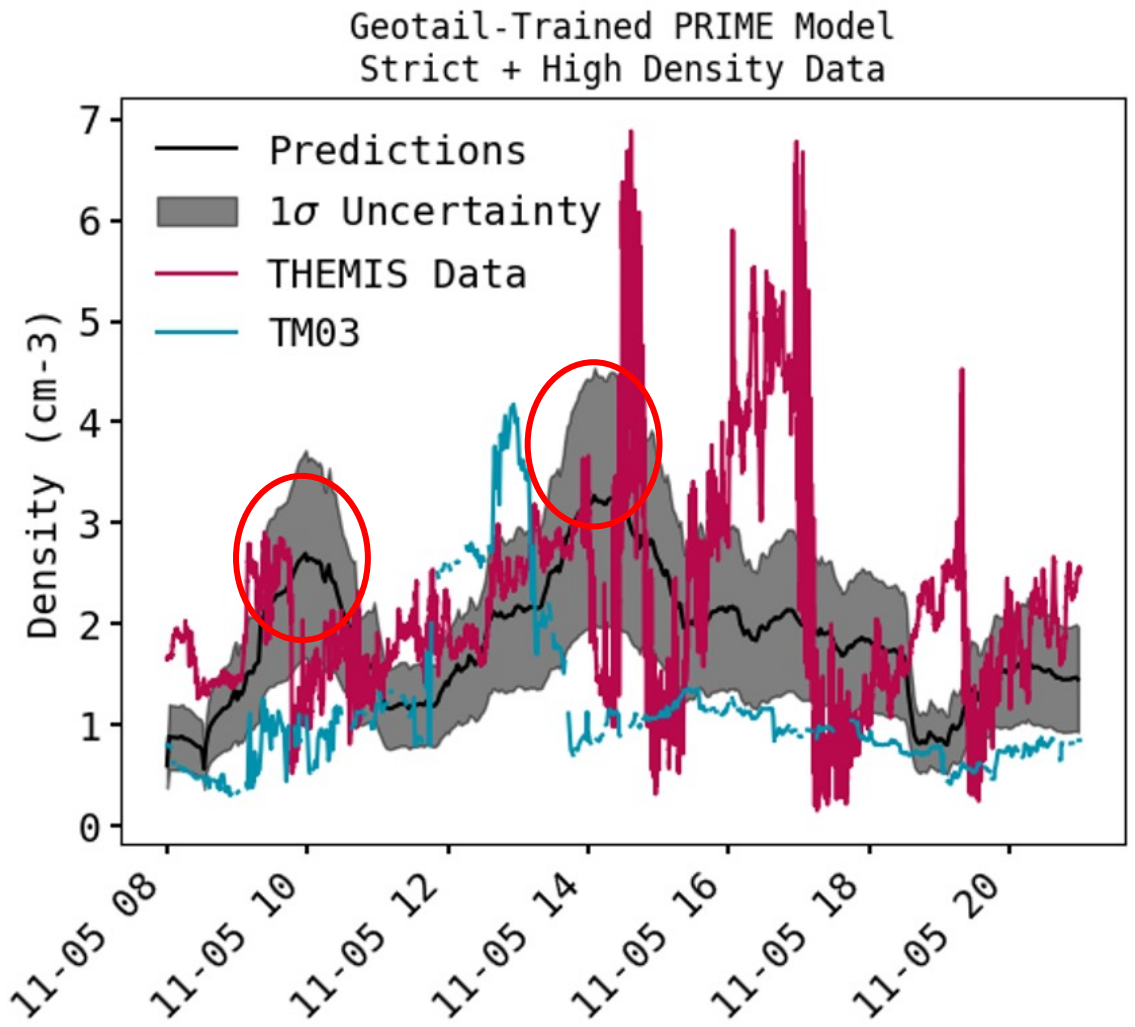
The Risk: Therefore we can mistakenly remove the crucial "stormtime plasmasheet."

The "Solution": Manually find the missing data and add it to the dataset.

Strict CPS (e.g., Ohtani et al., 2008 Raptis et al., 2024) & Flexible CPS (e.g., Richard et al., 2022)

Test case of a storm (05 Nov 2023)



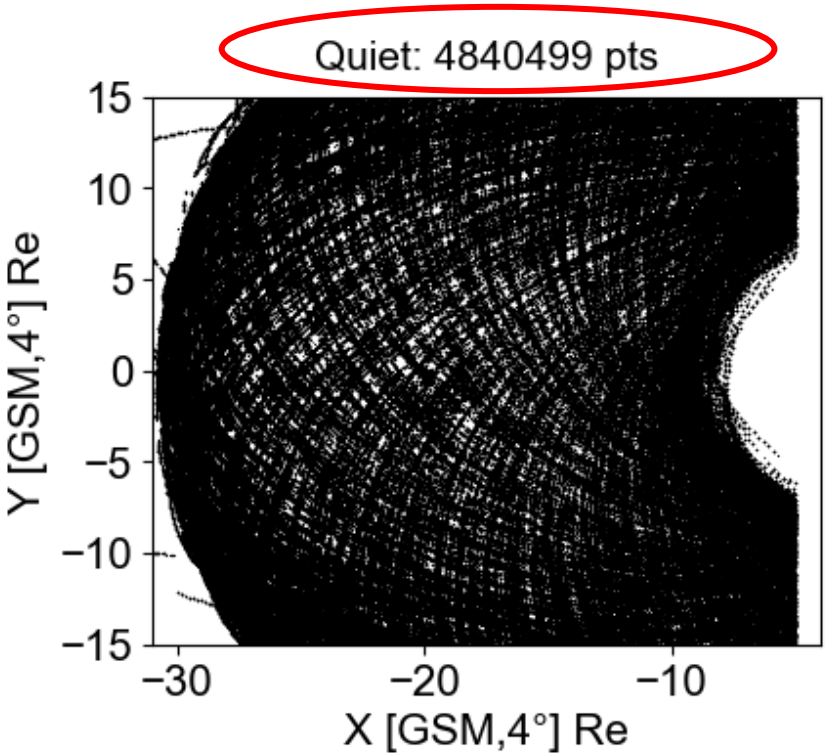


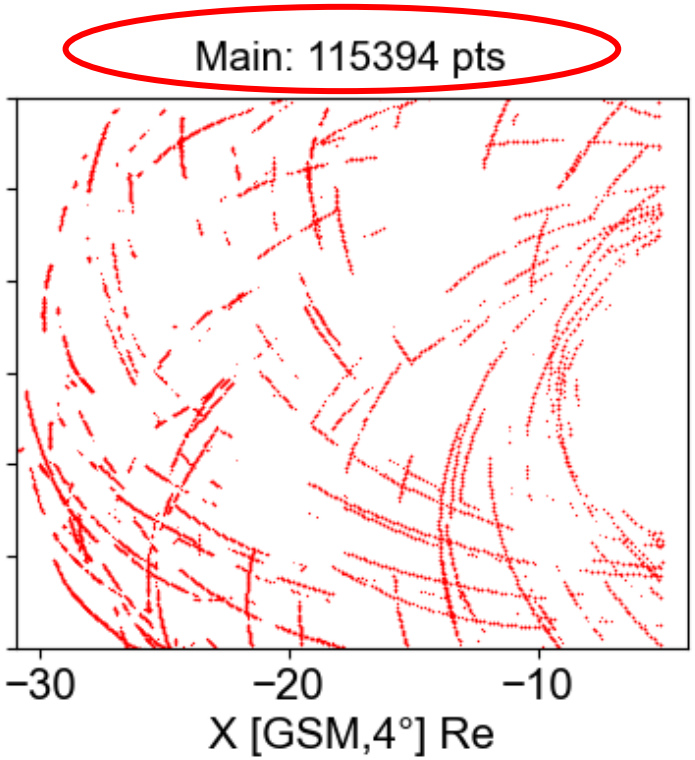
Note: values <1 cm^{-3}, are transitions to the lobe/BL (will filter them out).

<u>MAE (>40% improvement)</u>
ML model: 0.7 [1/cc]
TM03: 1.22 [1/cc]

Data Sparsity During Extreme Events

Geotail data (1994 – 2022), time resolution: 12 seconds



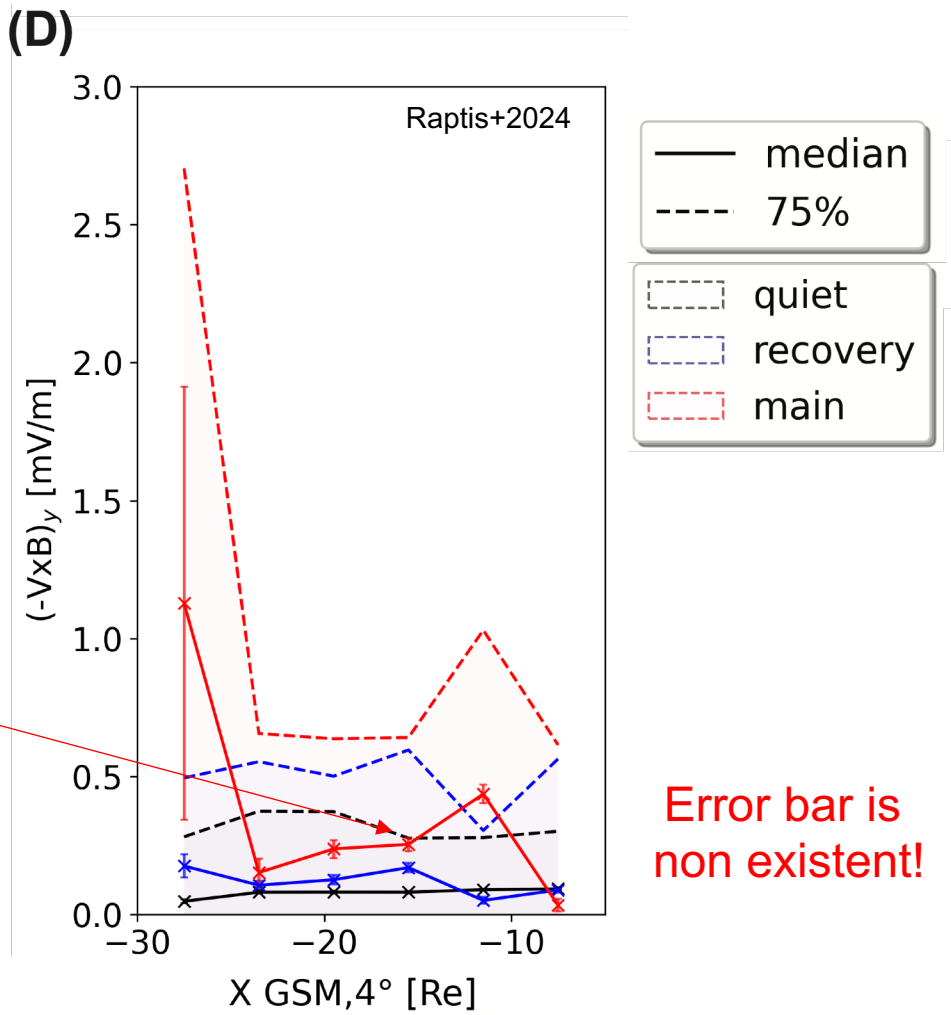


Number of data points can be misleading, not showing the actual coverage

Data number & storms

Table S1. Distribution of points for Geotail and MMS, used to generate Figure 2 of the main text. Each cell indicates the number of points along with the number of unique days and unique storms (#days | #storms). The bins used for the X axis are shown in the last row.

Geotail						
Main	7377 (36|32)	3006 (22|19)	2715 (17|17)	6240 (22|19)	6009 (21|19)	2450 (24|23)
Recovery	27254 (77|66)	19855 (68|58)	17770 (60|53)	14892 (58|52)	26234 (79|68)	35488 (102|93)
Quiet	220027 (553)	122734 (425)	104663 (401)	128919 (438)	234532 (666)	382446 (833)
MMS						
Main	8 (1|1)	573 (3|3)	1476 (3|3)	1512 (3|3)	1299 (3|3)	1987 (7|5)
Recovery	781 (3|2)	4412 (8|6)	3889 (10|8)	4284 (12|10)	7907 (13|13)	7451 (18|16)
Quiet	21036 (52)	53739 (135)	40825 (126)	37538 (132)	46777 (166)	72190 (195)
Bins x	**[-30, -25]**	**[-26, -21]**	**[-22, -17]**	**[-18, -13]**	**[-14, -9]**	**[-10, -5]**



Error bar is non existent!

What does it mean to have 1500 data points if they originate from 3 storms/3days?

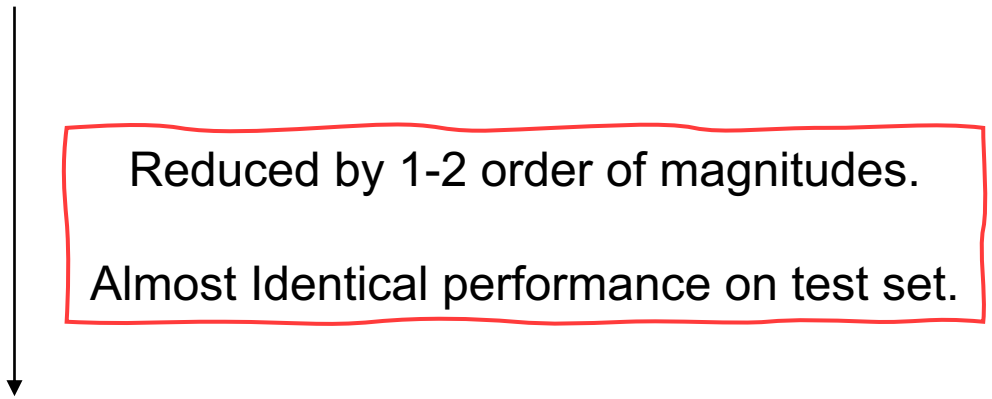
Key Message: #of unique days, # unique storms, and distribution of upstream conditions is more important than SE

Is all this data needed? (Discussion point – Imbalanced learning)

1. **Identify Outliers:**
   - **Find Outliers:** Detected using *Isolation Forest* (unusual feature patterns)

2. **Build a Diverse Core (Farthest Point Sampling):**
   - Selects a *core subset* where points are maximally distant
   - Ensures broad coverage and *high diversity* of the original data

3. **Balance with Rare Samples (Kernel Density Estimation):**
   - Adds points from *under-represented regions*




```
Original dataset size: 241311
Reduced dataset size: 6978
```

Summary & Discussion

Results

✅ **Marginal Gains:** ML models overall outperform analytical methods and show hidden asymmetries.

❌ **Mediocre Predictability:** We mainly capture "boring" conditions, not the critical rare events.

🧠 **Core Problem:** Our training data is biased. **Extreme events, are not always captured by simple thresholds, are very rare, and including them is methodologically challenging**

Future Work

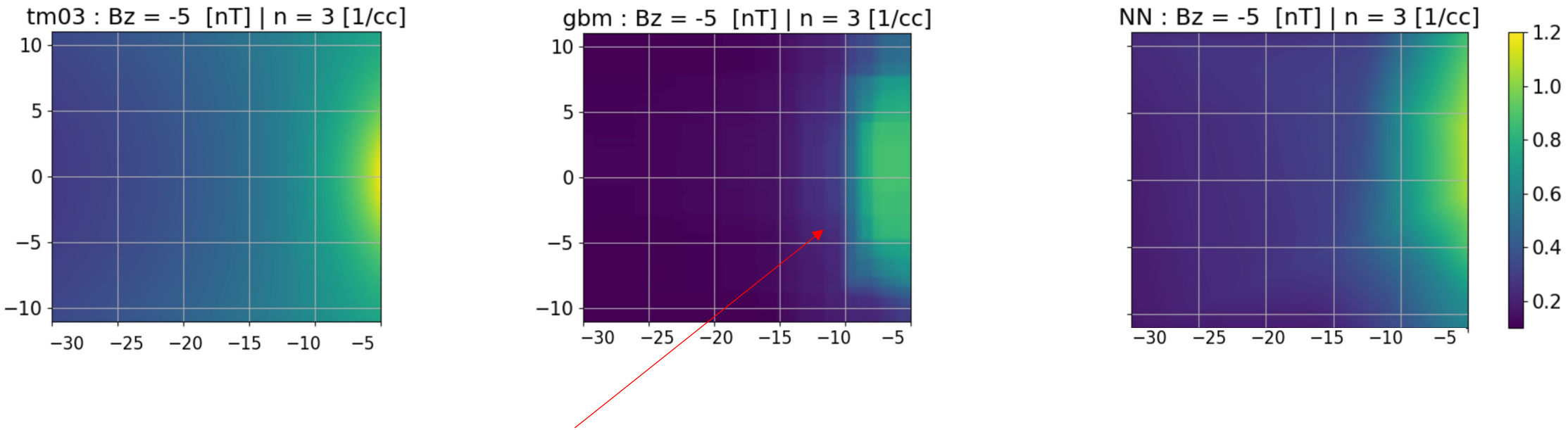
- **Understand the output:** How can we use these output to understand the physical processes?
- **Simulations to the Rescue (?):** Use simulations to generate extreme events we lack in data.

Discussion Points/Reminders

- A model with R^2 ~0 can have a r of 0.7 and tiny MAE depending on the problem.
- "Better" data can yield ~30% difference which is in the same order of improvements that was found from a simple linear regression to the most complicated model (~40%)
- "Unique and extreme" events and parameter distribution can be more important than metrics.

Extras

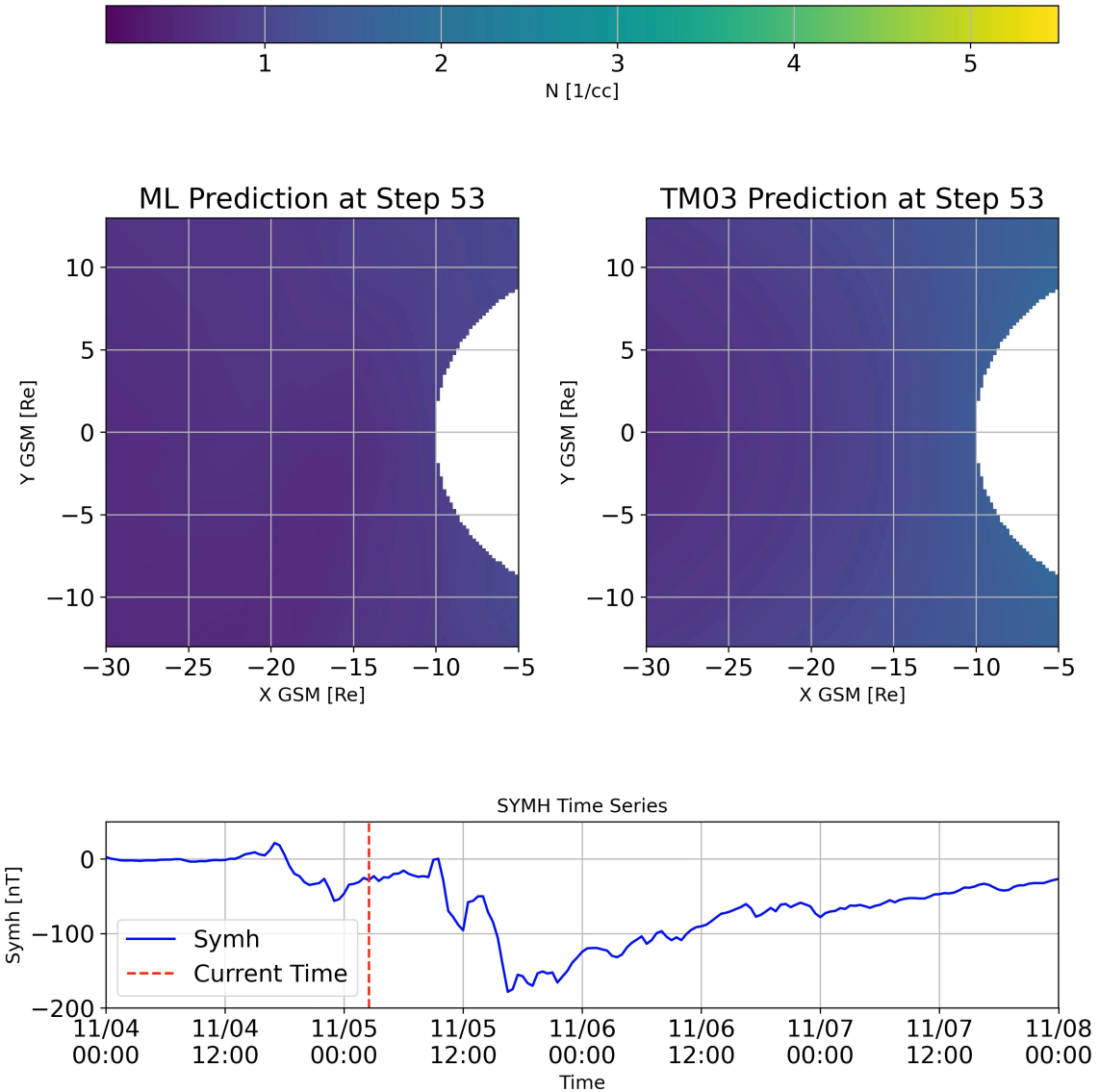
Community Reminder via Modeling Bz



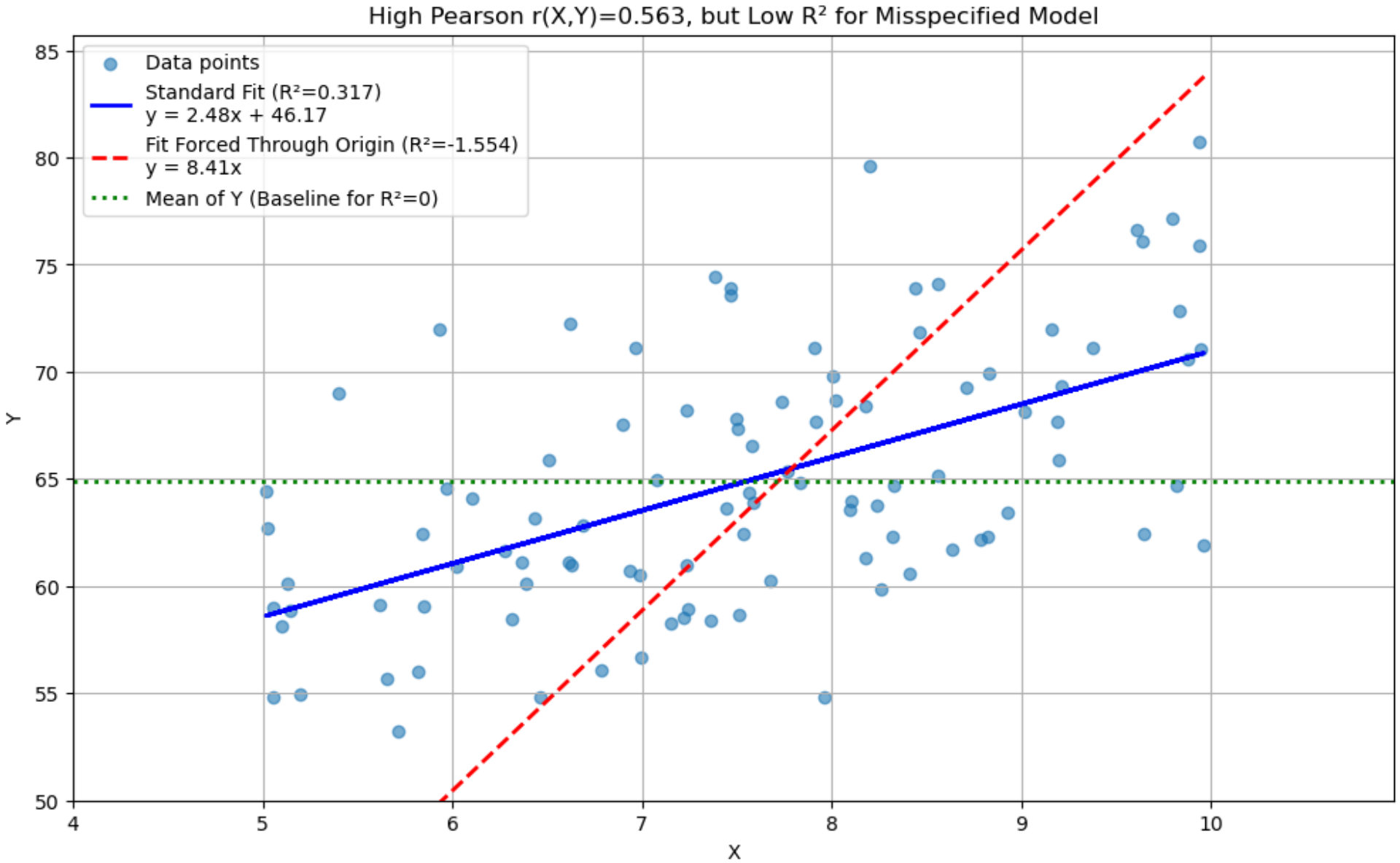
Gradient boosting is powerful, but it's a poor extrapolator.

Since it's built on decision trees, it predicts from nearby data rather than extending patterns like a neural network.

ML storm time density modeling

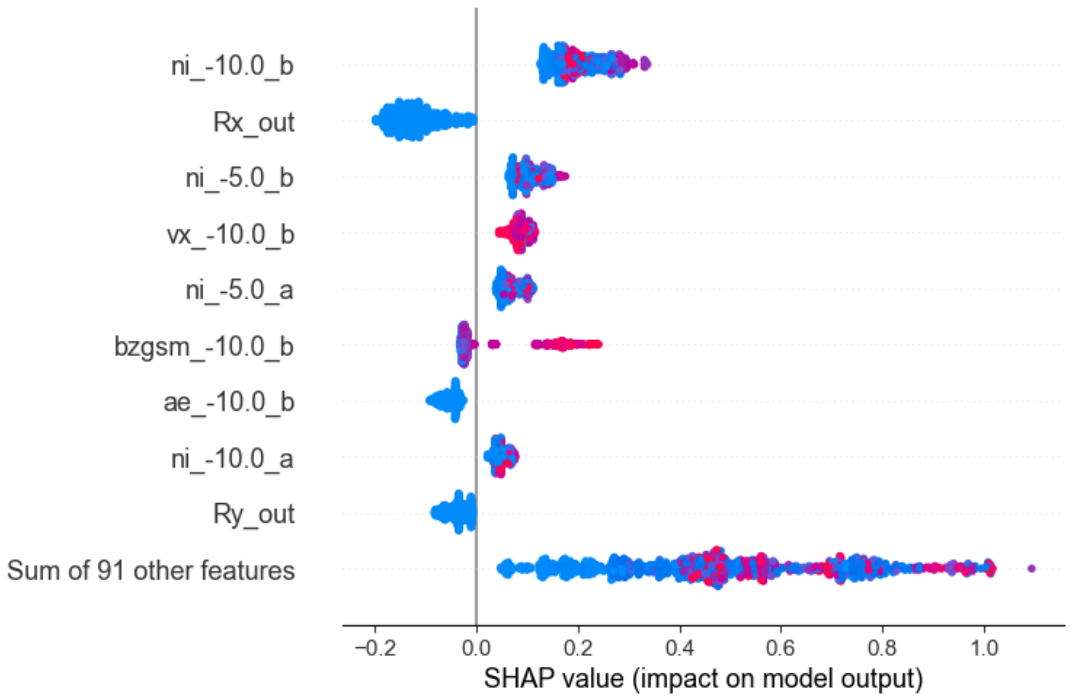


Correlation and Rsquared difference

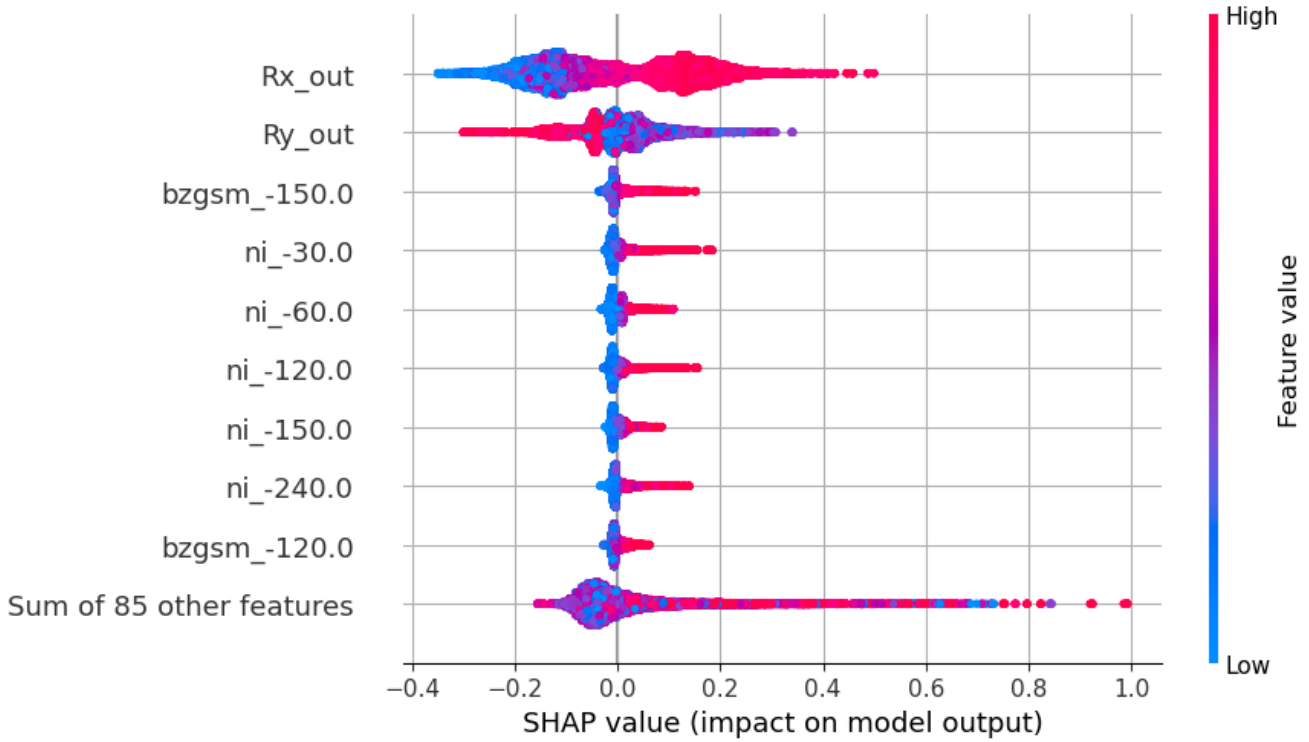


Model Feature importance storm vs quiet

In other words, the increased upstream density (-1h) had a greater impact during the storm than the SC location.

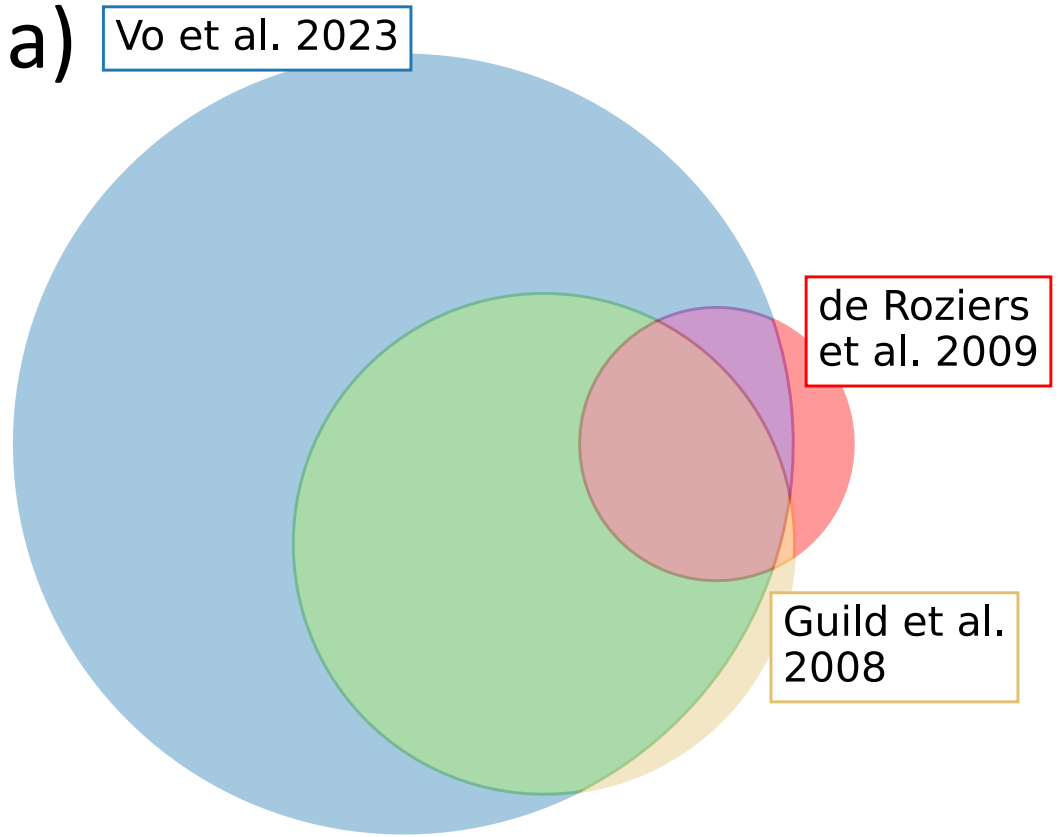


X_test_storm



X_test_total

Classifying plasma sheet is not trivial



Plasma Sheet Criteria			Number
Vo et al. 2023	de Roziers et al. 2009	Guild et al. 2008	
Yes	No	No	1,259,896
No	Yes	No	39,451
No	No	Yes	28,828
Yes	Yes	No	46,399
Yes	No	Yes	686,527
No	Yes	Yes	10,483
Yes	Yes	Yes	170,467

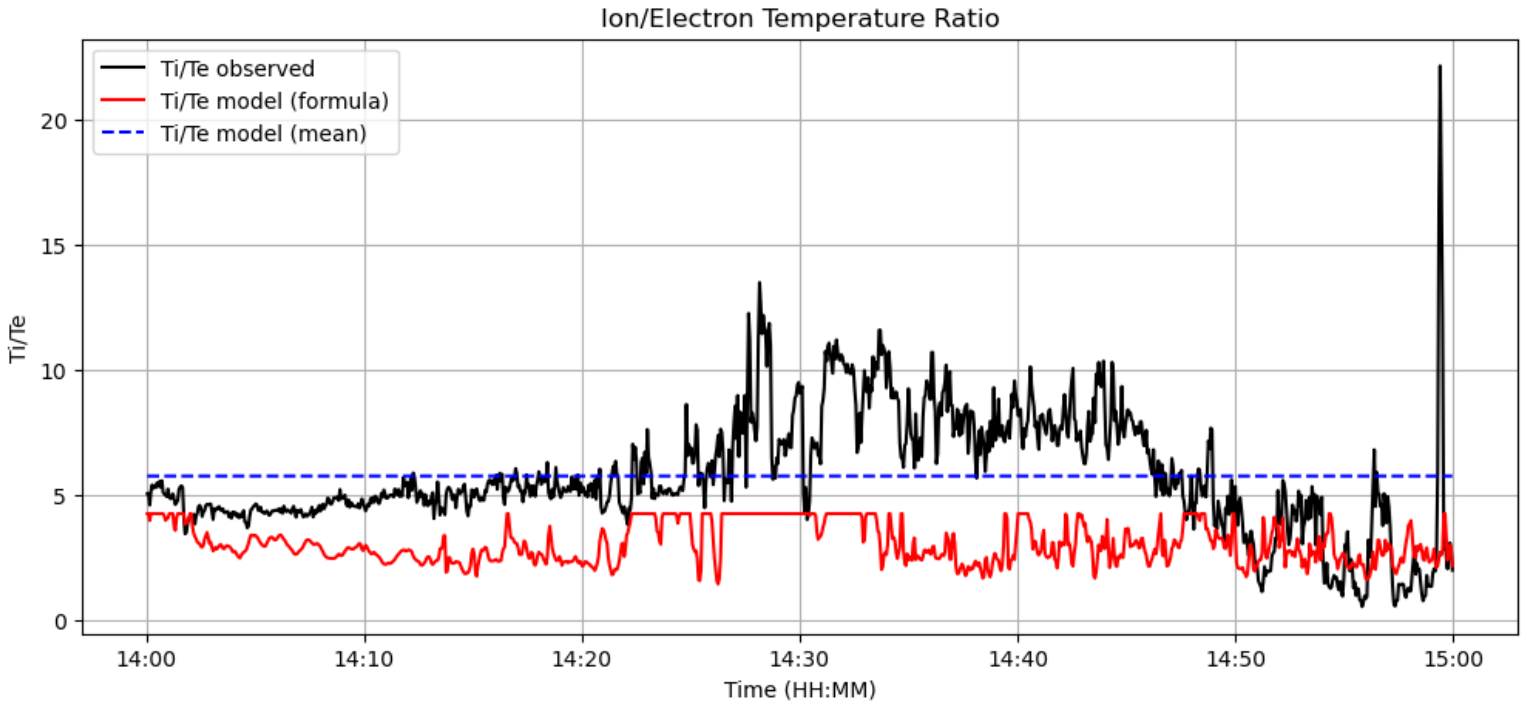
Note Vo+2023, had a multi-step process based on interval, this is just using the point-by-point classification

How is Ti/Te doing there?

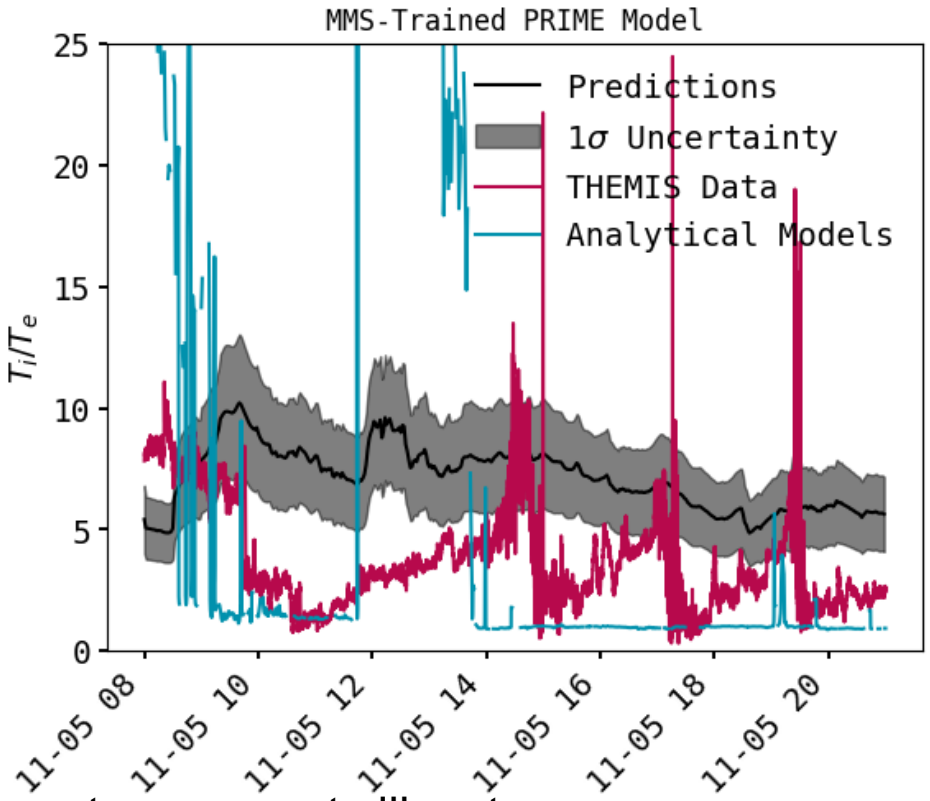
Mean Model: MAE = 1.67

Formula Model (a polynomial fit from another period): MAE = 3.21

Formula vs Mean: MAE = -92.64%



Old model shown, new is actually better



Key Message: Modeling (in-situ or from SW) struggle during extreme events like storms