SPACE WEATHER PREDICTION: PRELIMINARY RESULTS USING MACHINE LEARNING TECHNIQUES

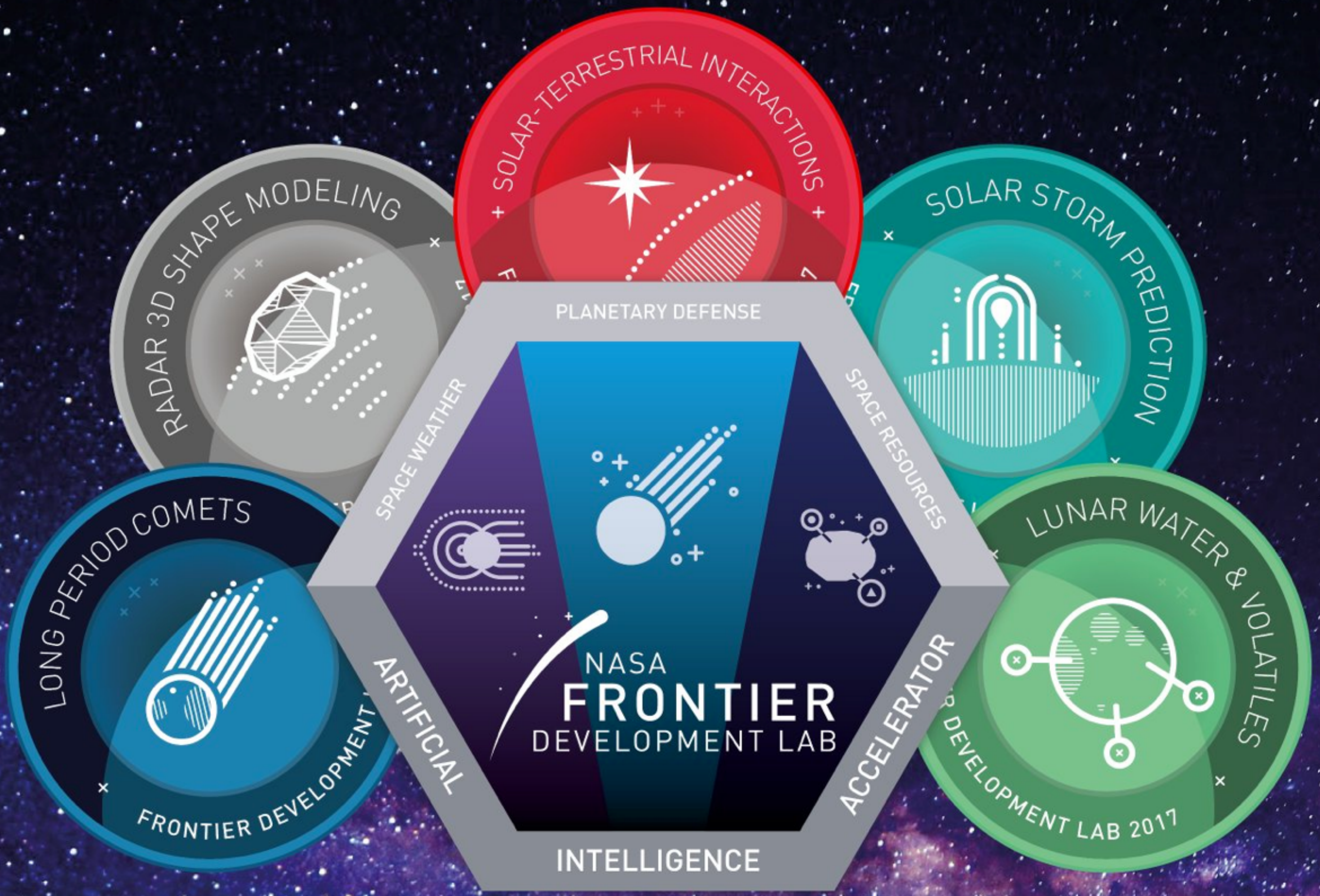
Bala Poduval, Space Science Institute, Boulder, CO

Presented at NOAA/SWPC, Boulder, CO

18 January 2018

NASA FDL 2017

An artificial intelligence (AI) Research and Development accelerator dedicated to bridging knowledge gaps in space programs important to NASA and the future of humanity.



IBM kx SPACE RESOURCES.LU LOCKHEED MARTIN intel SETI INSTITUTE NVIDIA AUTODESK XPRIZE MISO

PROBLEM DEFINITION

socio economic impact

Risk Analysis
Post-facto Analysis
Forecast
Actuary
??

Short-term

Understand "Cause" & "Effect"

Long-term

Analyse Data

Traditional

Non-traditional

(Breakthrough?!)

Solar

Terrestrial

PROBLEM DEFINITION

A data-driven, open source tool for space weather forecasting.

predicts a commonly used index Kp, that represents geomagnetic disturbances.

<u>Value Proposition:</u>

Open Access Science Data x AI and ML frameworks

=

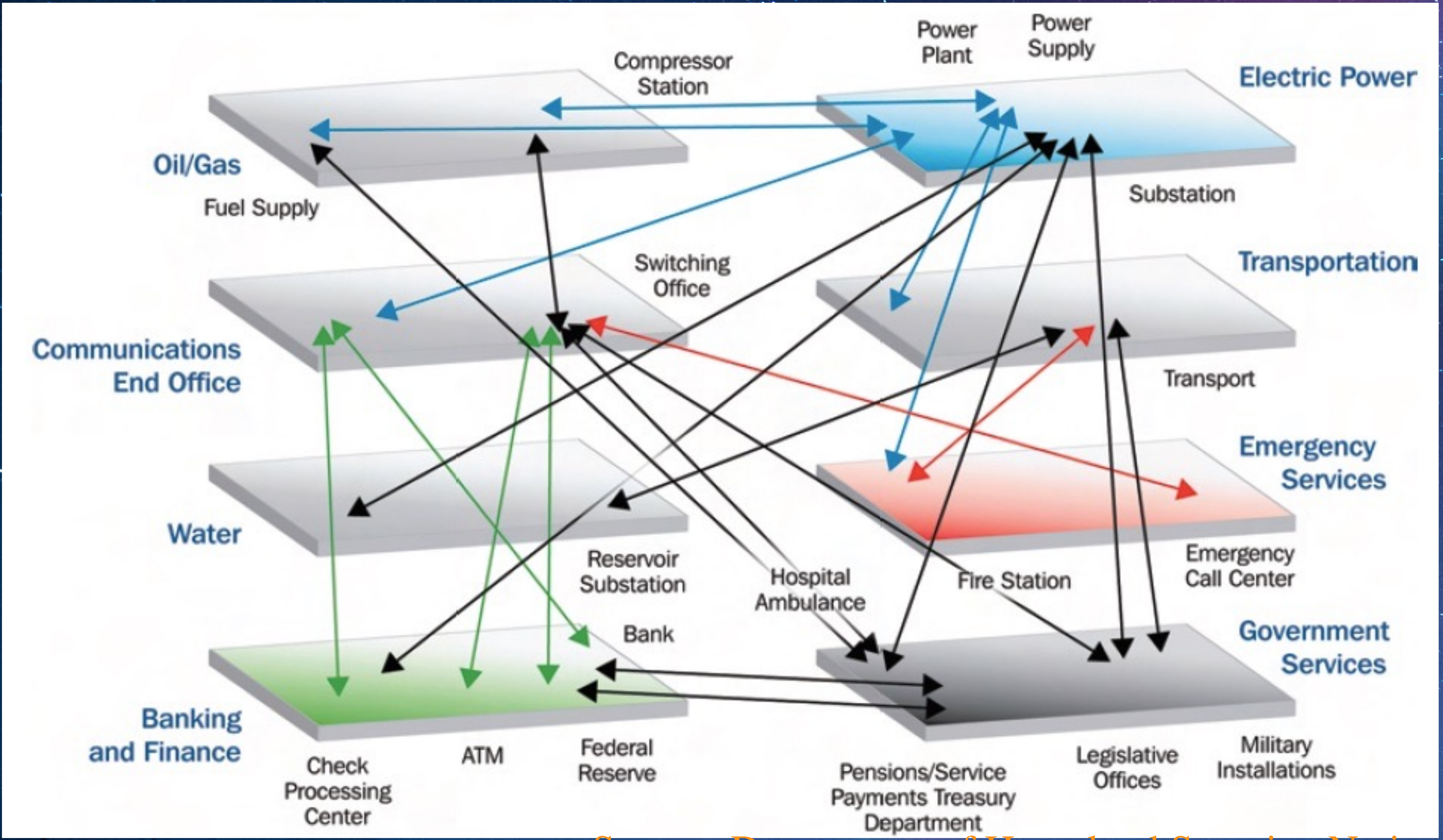
Better Scientific Insights + Better Predictions

Socioeconomic Impact

Impact Area	Customer (examples)	Action (examples)	Cost (examples)
Spacecraft (Individual systems to complete spacecraft failure; communications and radiation effects)	• Lockheed Martin • Orbital • Boeing • Space Systems Loral • NASA, DoD	• Postpone launch • In orbit - Reboot systems • Turn off/safe instruments and/or spacecraft	• Loss of spacecraft ~$500M • Commercial loss exceeds $1B • Worst case storm - $100B
Electric Power (Equipment damage to electrical grid failure and blackout conditions)	• U.S. Nuclear Regulatory Commission • N. America Electric Reliability Corp. • Allegheny Power • New York Power Authority	• Adjust/reduce system load • Disconnect components • Postpone maintenance	• Estimated loss ~$400M from unexpected geomagnetic storms • $3-6B loss in GDP (blackout)
Airlines (Communications) (Loss of flight HF radio communications) (Radiation dose to crew and passengers)	• United Airlines • Lufthansa • Continental Airlines • Korean Airlines • NavCanada (Air Traffic Control)	• Divert polar flights • Change flight plans • Change altitude • Select alternate communications	• Cost ~ $100k per diverted flight • $10-50k for re-routes • Health risks
Surveying and Navigation (Use of magnetic field or GPS could be impacted)	• FAA-WAAS • Dept. of Transportation • BP Alaska and Schlumberger	• Postpone activities • Redo survey • Use backup systems	• From $50k to $1M daily for single company

Severe Space Weather Events: Understanding Societal and Economic Impacts: A Workshop Report — The National Academies Press (22 May 2008)

Socioeconomic Impact



interconnected infrastructures
&
their qualitative dependencies and interdependencies

Source: Department of Homeland Security, National Infrastructure Protection Plan (http://www.dhs.gov/xprevprot/programs/editorial_0827.shtm).

Socioeconomic Impact

Power grids — most vulnerable to space weather during periods of light load with heavy electricity flows from generating plants to load centers — a common practice in the middle of the night during spring & fall

space weather forecasts is valuable

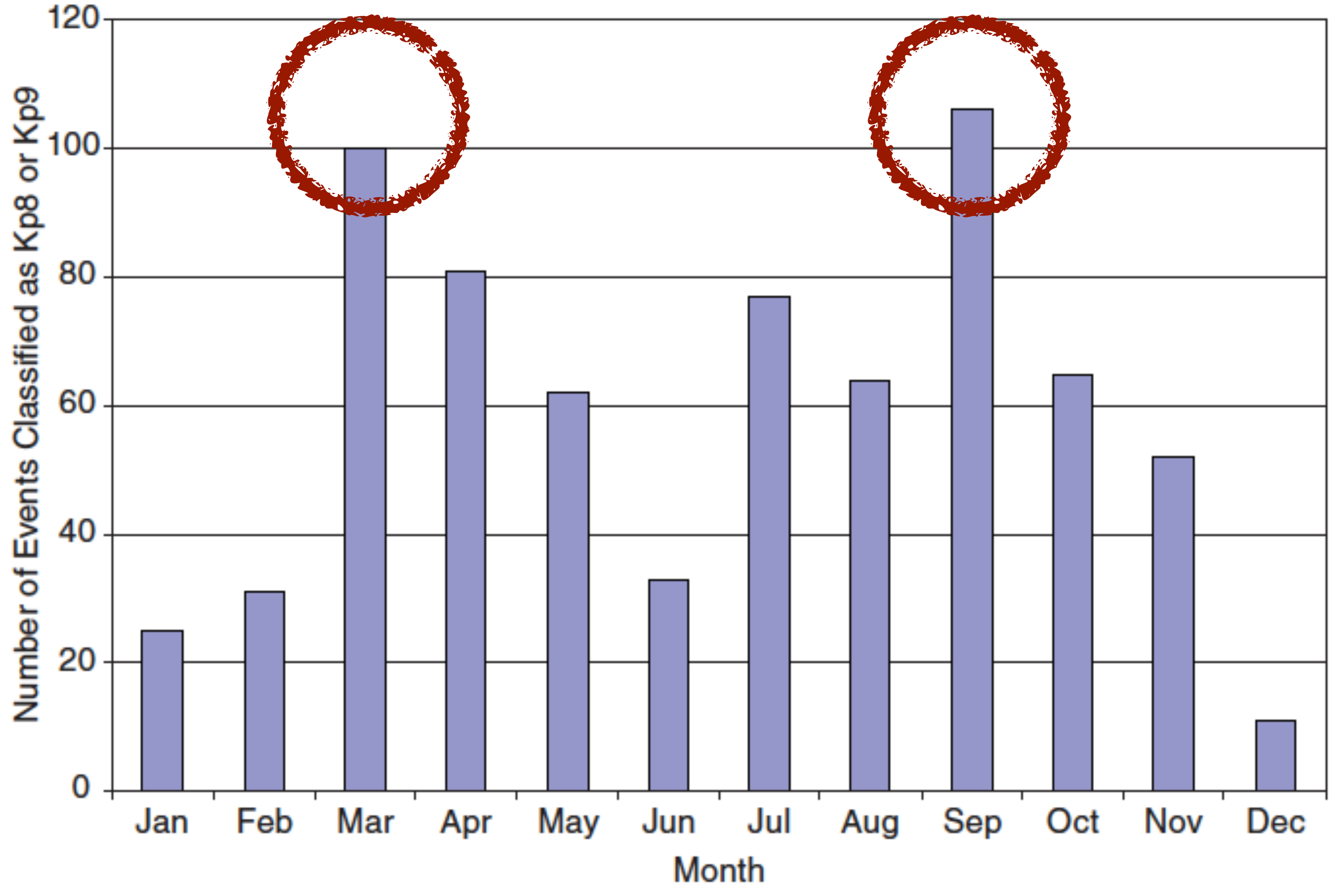


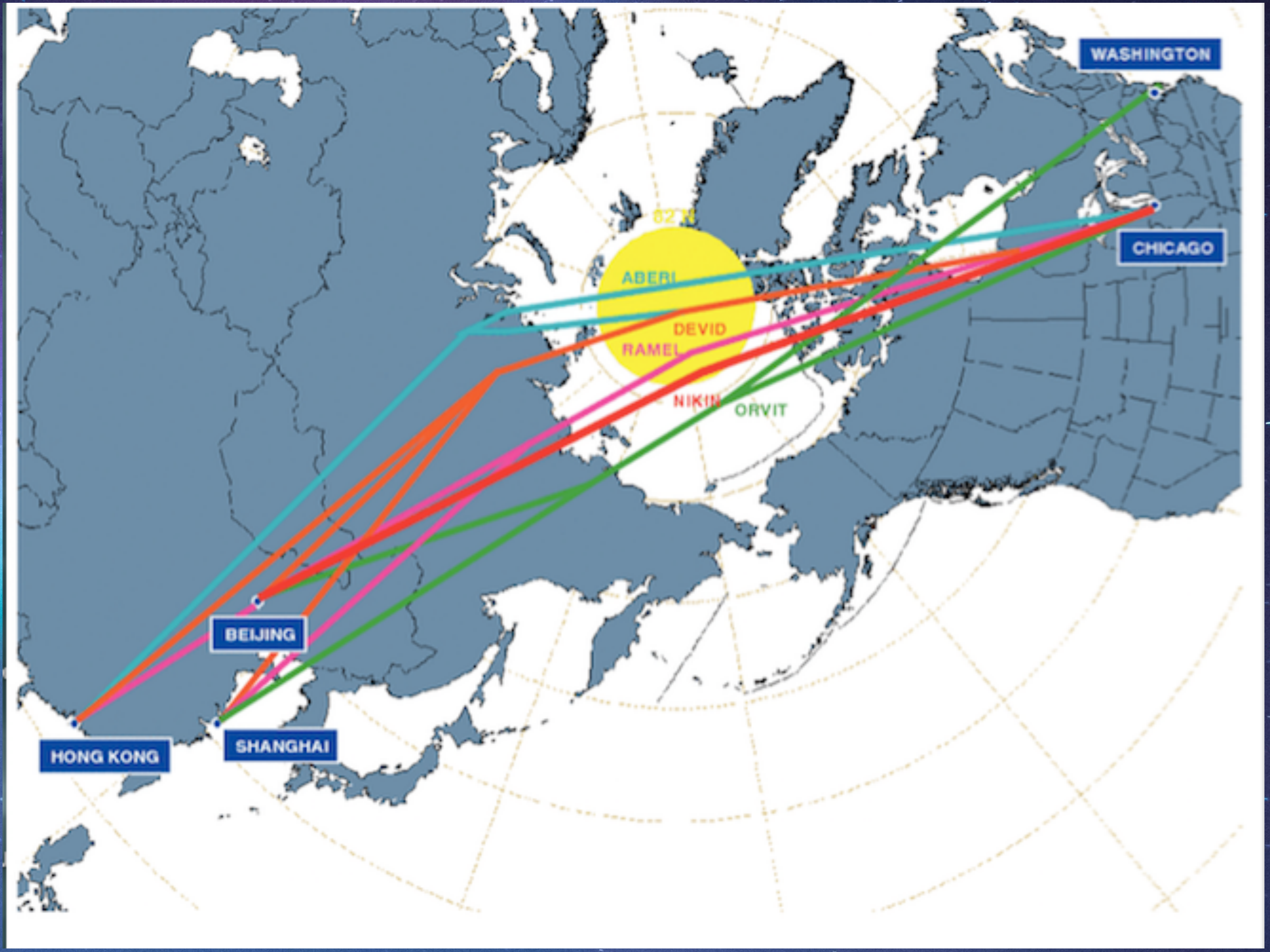
FIGURE 2.3 Incidence of Kp8/Kp9 events by month, 1932-2007, based on an analysis of 222,072 observations. SOURCE: Data from World Data Center for Geomagnetism.

major space weather events more frequent during spring & fall

Socioeconomic Impact

transpolar flights rely on HF radio communications

magnetic storm/polar cap absorption (PCA) — cause ionospheric density disturbances interfere with HF, VHF, UHF radio communications & navigation signals from GPS satellites



Severe Space Weather Events: Understanding Societal and Economic Impacts: A Workshop Report — The National Academies Press (22 May 2008)

Kp Index

G-Scale	Kp	Activity Level	Occurrence Frequency
G0	4 & lower	Below Storm	
G1	5	Minor Storm	1700 per cycle (900 days per cycle)
G2	6	Moderate Storm	600 per cycle (360 days per cycle)
G3	7	Strong Storm	200 per cycle (130 days per cycle)
G4	8	Severe Storm	100 per cycle (60 days per cycle)
G5	9	Extreme Storm	4 per cycle (4 days per cycle)

Kp Index & Space Weather Events

LEVEL	EVENT	DATE
Extreme	Carrington Event widespread disruption of telegraph	1 September 1859
Extreme	Bastille Day Event	14 July 2000
Extreme	Halloween Event Affected airlines, caused power outage, damaged transformers, led astronauts on ISS to take shelter	31 October 2003
Severe	Hydro-Quebec 9 hour blackout	13 March 1989
Moderate	Anik-E1 & Anik-E2 failed Disrupted TV & Computer transmission	20/21 January 1994

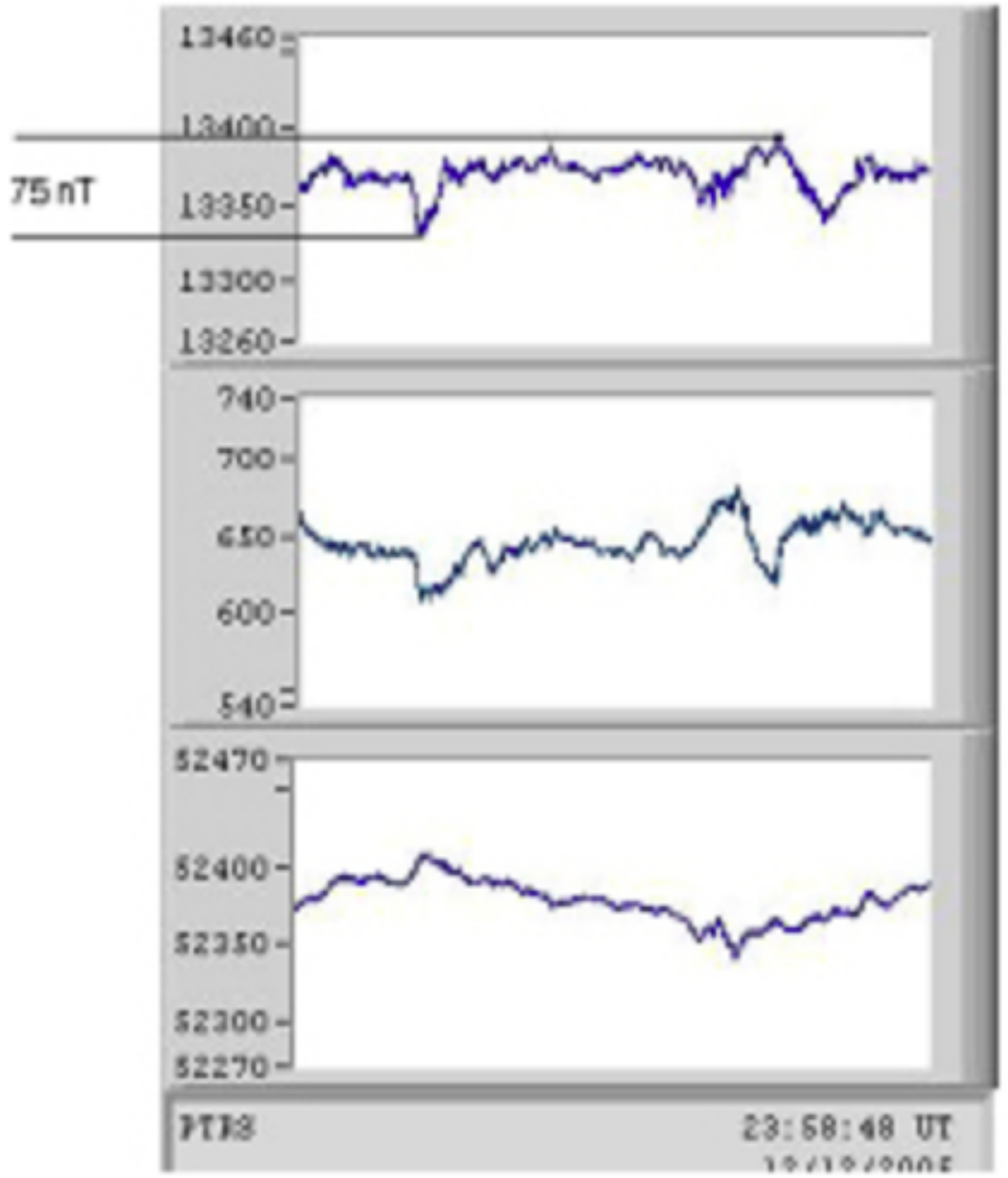
Scale	Description		Physical measure	Average Frequency (1 cycle = 11 years)
G 5	Extreme		Kp = 9	4 per cycle (4 days per cycle)
G 4	Severe		Kp = 8, including a 9-	100 per cycle (60 days per cycle)
G 3	Strong		Kp = 7	200 per cycle (130 days per cycle)
G 2	Moderate		Kp = 6	600 per cycle (360 days per cycle)
G 1	Minor	**Power systems:** Weak power grid fluctuations can occur. **Spacecraft operations:** Minor impact on satellite operations possible. **Other systems:** Migratory animals are affected at this and higher levels; aurora is commonly visible at high latitudes (northern Michigan and Maine).	Kp = 5	1700 per cycle (900 days per cycle)

Kp Index

Planetary Kp Index (Bartels, 1938)

— refers to a range of geomagnetic activity levels within 3-hr intervals each day (UT)

Kp varies from 0 to 9, quasi-logarithmically

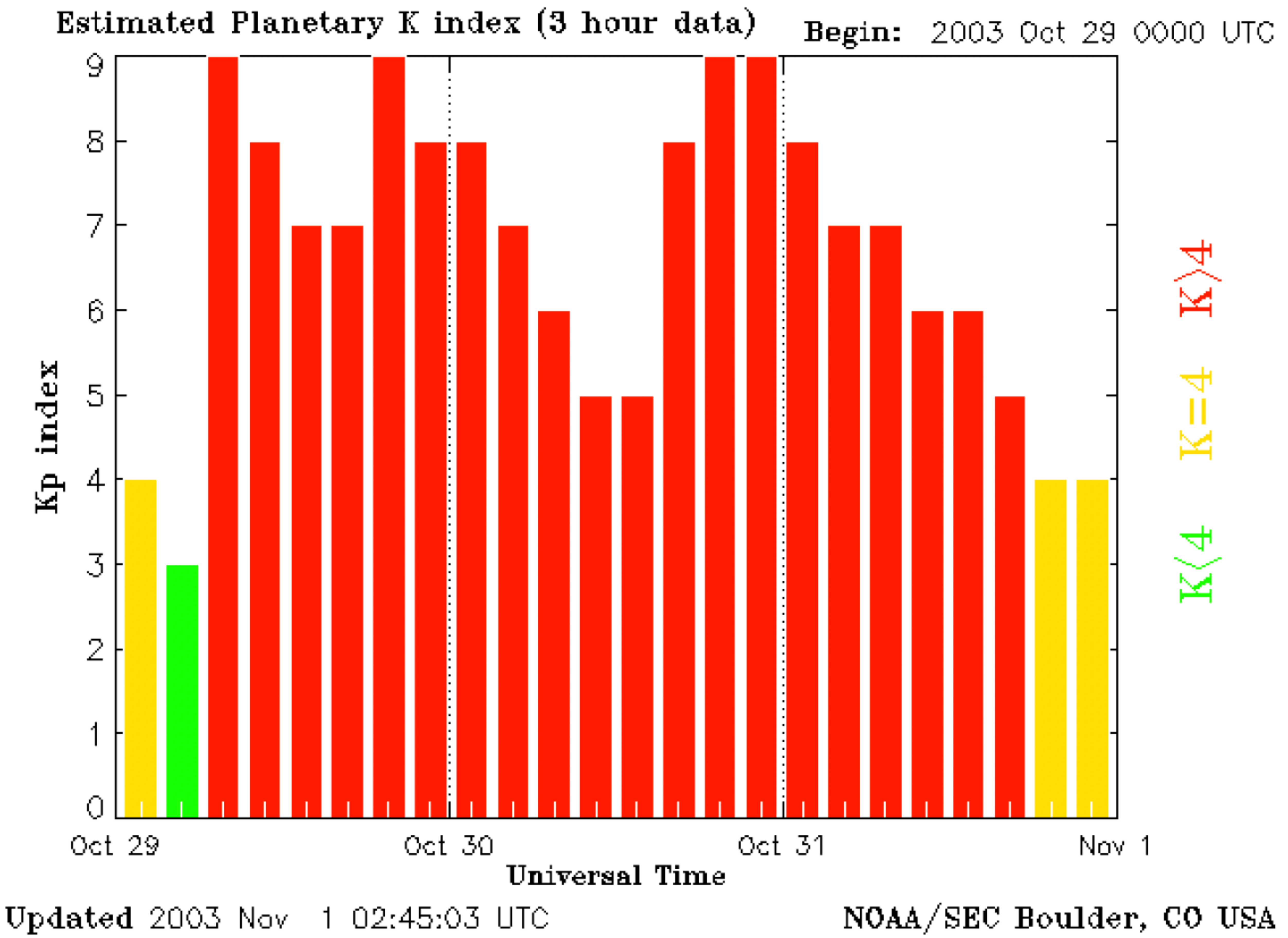


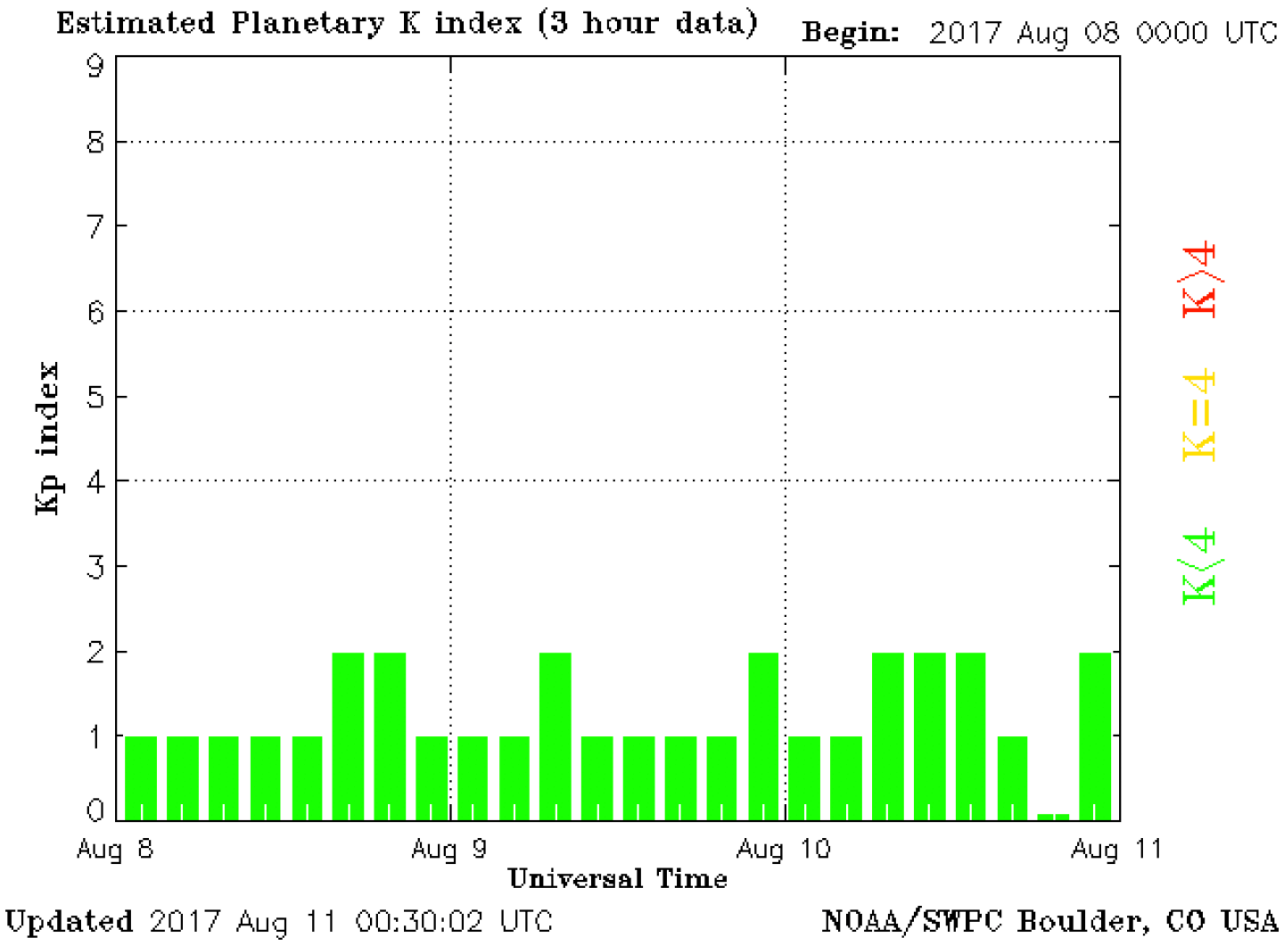
Petersburg, AK magnetometer data with a 75 nT change in the X-direction (Magnetic North)

Use this table on the right to convert the difference in the maximum and minimum x-values for today to a K index. The larger the K index, the stormier it is in Earth's magnetic field.

K index	nT diff.
0	0-5
1	5-10
2	10-20
3	20-40
4	40-70
5	70-120
6	120-200
7	200-330
8	330-500
9	>500

Kp Index





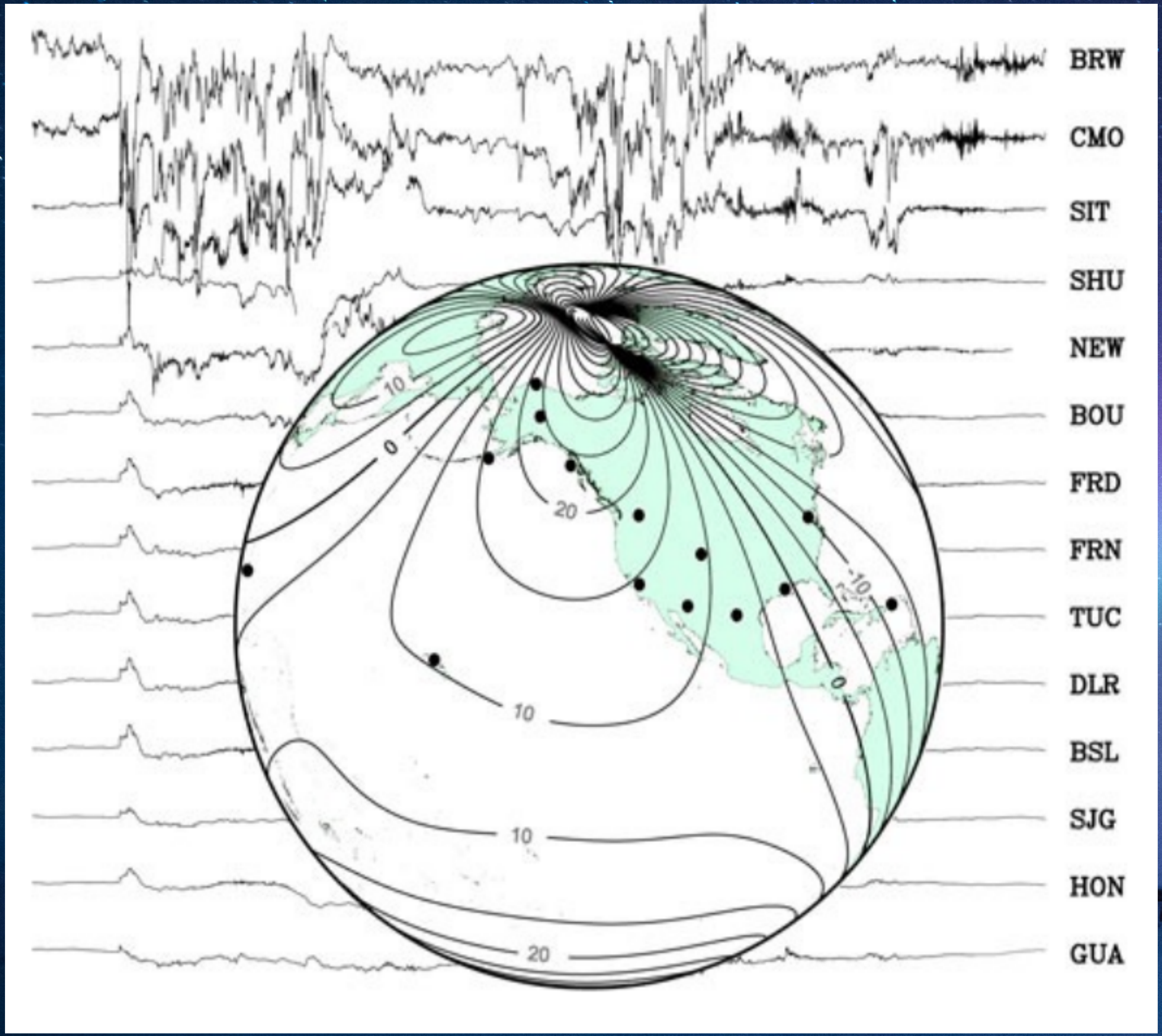
EXTREME

Halloween Event

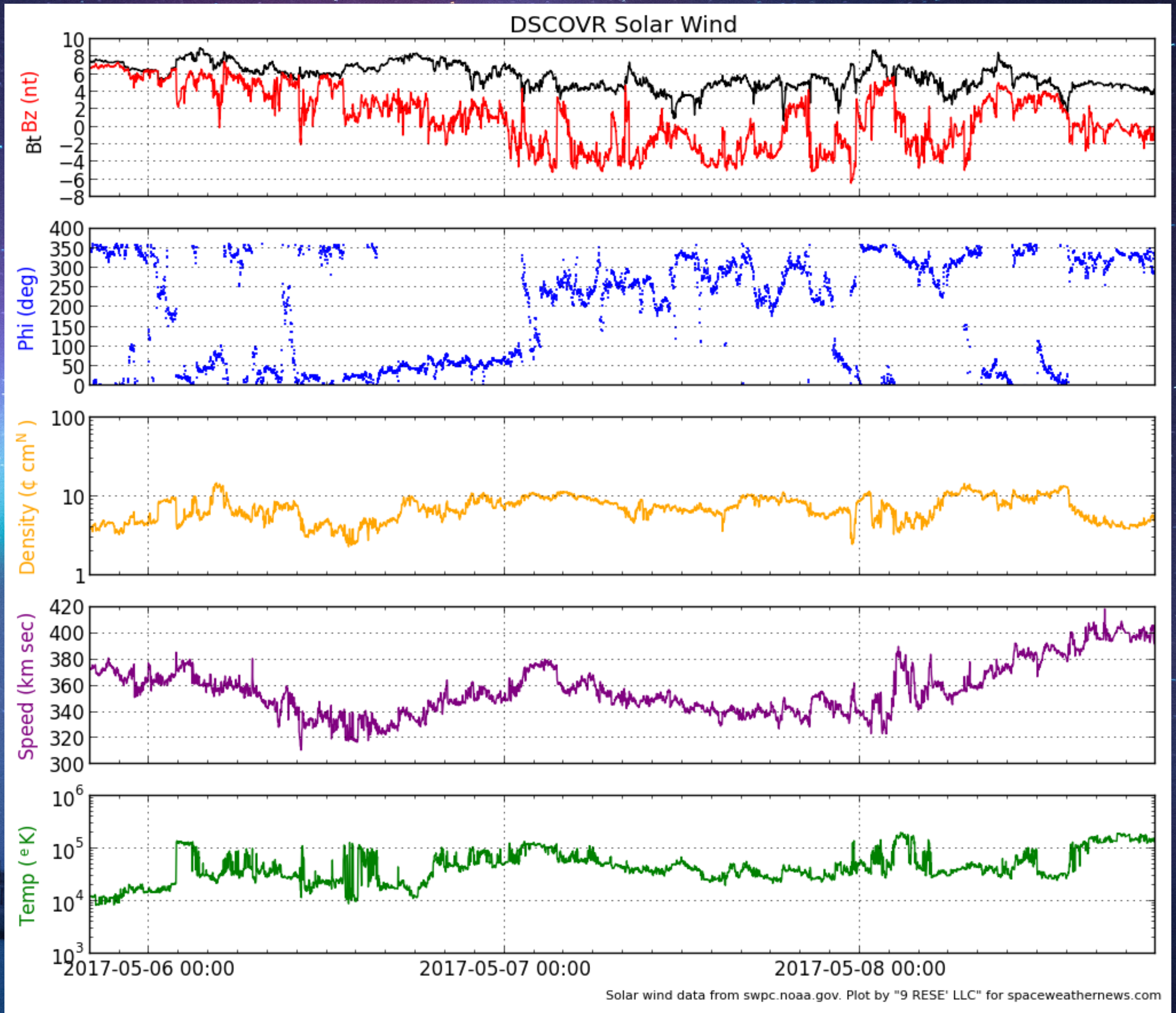
QUIET

Data Used

GEOMAG DATA



SOLAR WIND DATA



Machine Learning

we made use of the industry--standard, open source ML software packages:

- scikit-learn
- Keras
- Tensorflow

Machine Learning

KERAS	Scikit-Learn	TensorFlow
an open source neural network (NN) library written in Python	a free ML library for Python contains various classification, regression & clustering algorithms support vector machines (SVMs) random forests (RF) gradient boosting designed to operate with Python numerical & scientific libraries NumPy & SciPy	an open source software library designed for building & training neural networks to detect and decipher patterns & correlations

Machine Learning

supervised machine learning

function approximation

classification & regression

ensemble model

combines several weak learners (usually, decision trees) that have a slight performance advantage over random guessing, to produce a powerful predictive model

ENSEMBLE MODELS

Data Used

Data Used	Models	Type
Geomag, Solar Wind, Kp index	Random Forest Gradient Boosting AdaptiveBoost Extra Trees	Ensemble
Geomag, Solar Wind	Long Short Term Memory (LSTM)	Recurrent Neural Network

ENSEMBLE MODELS

Data Used

Geomag,
Solar Wind,
Kp index

Gradient Boosting
AdaBoost
Extra Trees
Random Forest

These models rank the input features (parameters) according to their importance in making the decision for the output (prediction).

Ensemble

ENSEMBLE MODELS

main drawback

of decision trees

they tend to overfit the training data

this can be overcome

by combining several trees

each tree different from others

each tree does a good prediction by overfitting on part of the data but different from other trees

the overfitting can be reduced

by combining several of such trees
average their results

the reduction in overfitting while retaining the predictive power of trees can be proved using rigorous mathematics

ENSEMBLE MODELS

Ensemble models considers additive models of the form:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Similar to other boosting algorithms GBRT builds the additive model in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

At each stage the decision tree $h_m(x)$ is chosen to minimize the loss function L given the current model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x))$$

The initial model F_0 is problem specific, for least-squares regression one usually chooses the mean of the target values.

Gradient Boosting attempts to solve this minimization problem numerically via steepest descent: The steepest descent direction is the negative gradient of the loss function evaluated at the current model F_{m-1} which can be calculated for any differentiable loss function:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i))$$

Where the step length γ_m is chosen using line search:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)})$$

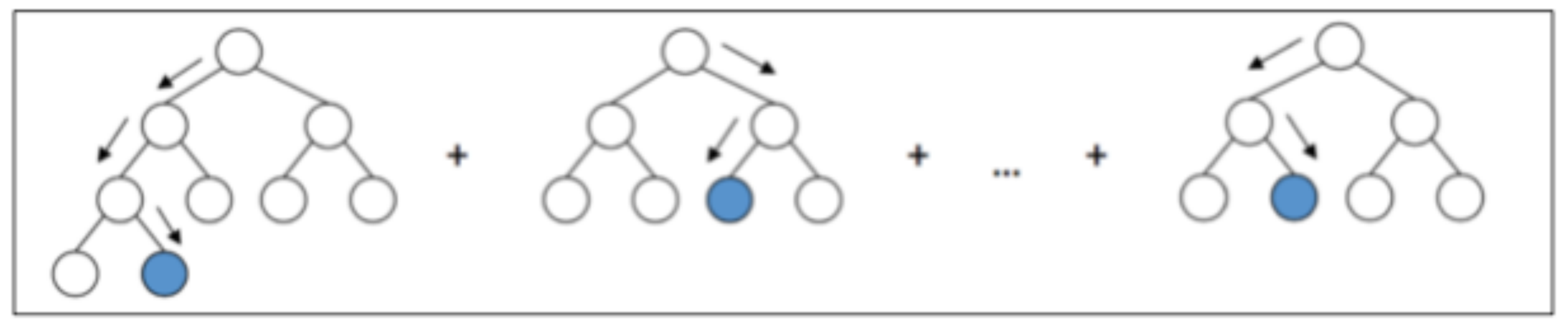
$$y1 = F(x) \ \ [\text{minimizing } (y1 - y)^2]$$

$$F_1(x) = F(x) + h(x)$$

$$= y$$

$$Therefore, \ \ h(x) = y - F(x) \ [\text{the residual}]$$

ENSEMBLE MODELS



ENSEMBLE MODELS

HyperParameters

- n_estimator
- max_features /max_depth
- n_features
- learning_rate

- the number of trees to be built
- the tree-depth
- the number of input features
- controls over-fitting

default values of max_features

- for classification

max_features = sqrt(n_features)

- for regression

max_features = n_features

ENSEMBLE MODELS

HyperParameters

Random Forest

trees are determined randomly

- n_estimator = 10
- max_features = default
- n_features = over 50

default values of max_features

- for classification
  max_features = sqrt(n_features)
- for regression
  max_features = n_features

RF can be parallelized across multiple CPU cores, especially on large data sets
we haven't implemented it in our present study

ENSEMBLE MODELS

HyperParameters

Adaptive Boosting

- n_estimator = 50
- learning_rate = 1
- n_features = over 50

ExtraTree

- n_estimator = 10
- max_features = default
- n_features = over 50

<u>default values of max_features</u>

- for classification — max_features = sqrt(n_features)
- for regression — max_features = n_features

ENSEMBLE MODELS

HyperParameters

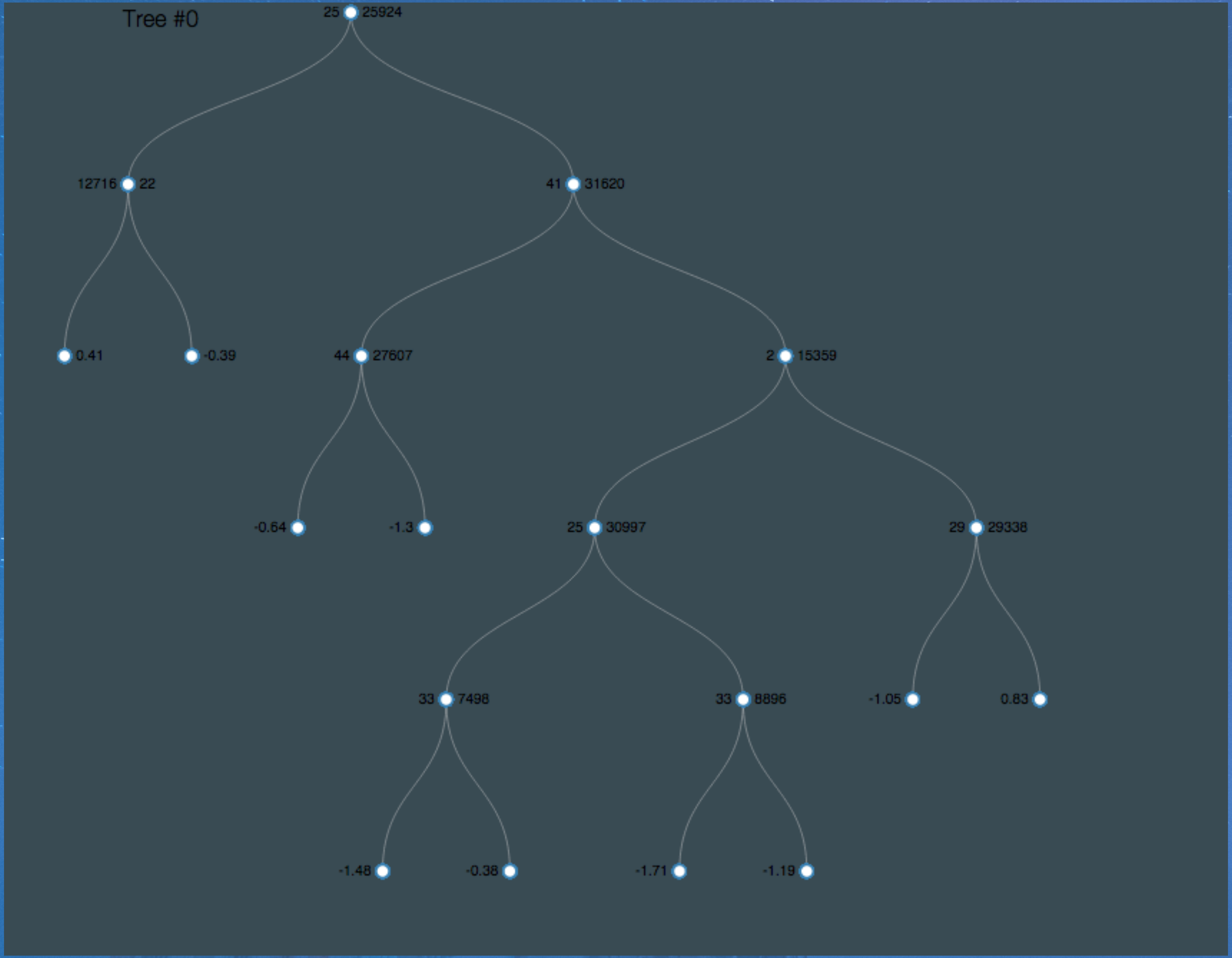
Gradient Boosting

trees are built serially & are shallow

- n_estimator = 100
- max_depth = 3
- learning_rate = 0.1
- n_features = over 50

tuning the parameters of GB accurately (the challenge and, therefore, the drawback), it can provide great accuracy — the most widely used supervised machine learning method

Gradient Boosting



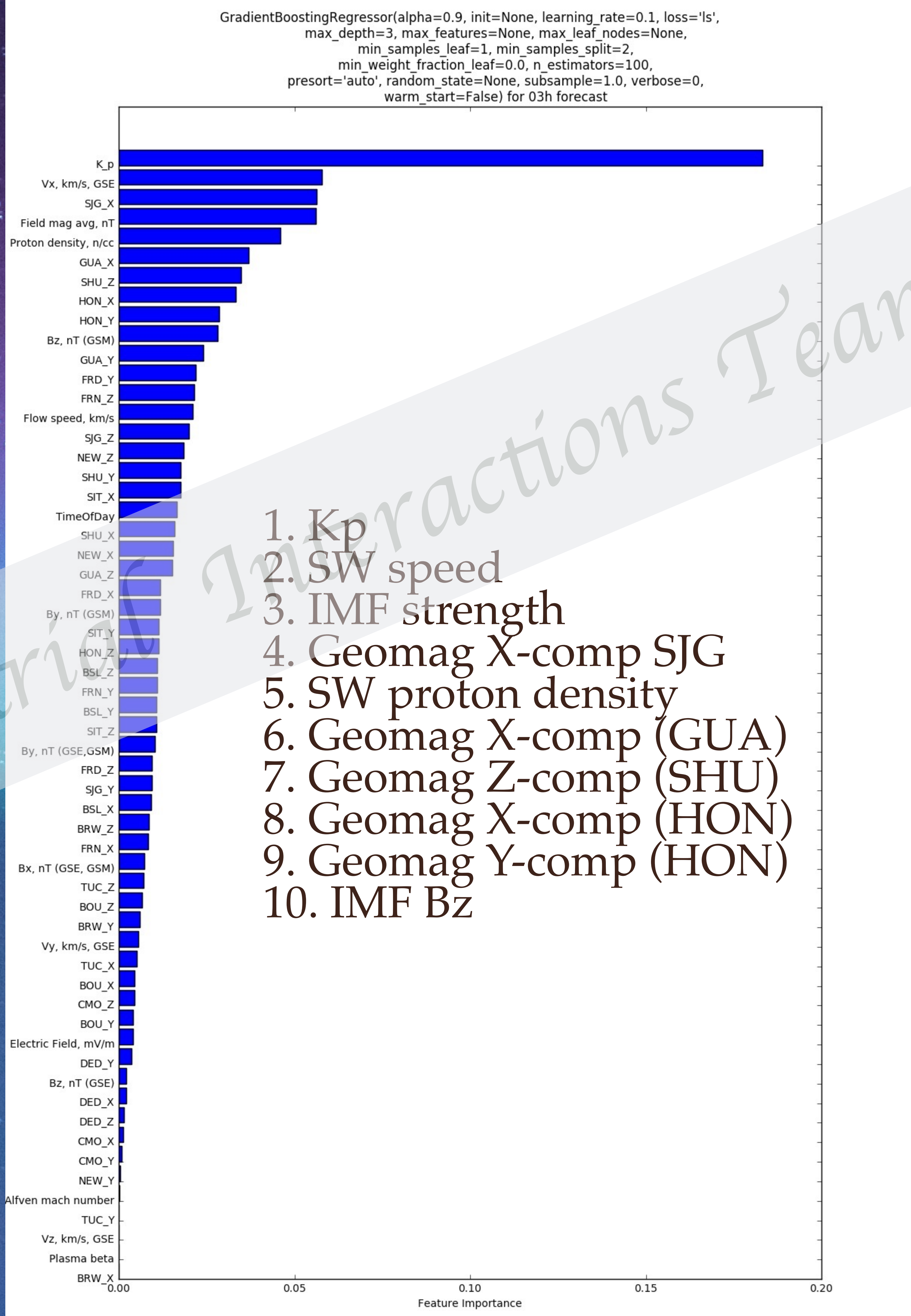
Mean Square Errors

ML method	1h ahead	3h ahead	6h ahead
Persist	0.007	0.020	0.025
Mean	0.046	0.046	0.046
Median	0.048	0.048	0.048
Gradient Boosting	0.007	0.015	0.021
Adaptive Boost	0.012	0.018	0.032
Extra Trees	0.009	0.021	0.027
Random Forest	0.015	0.015	0.026

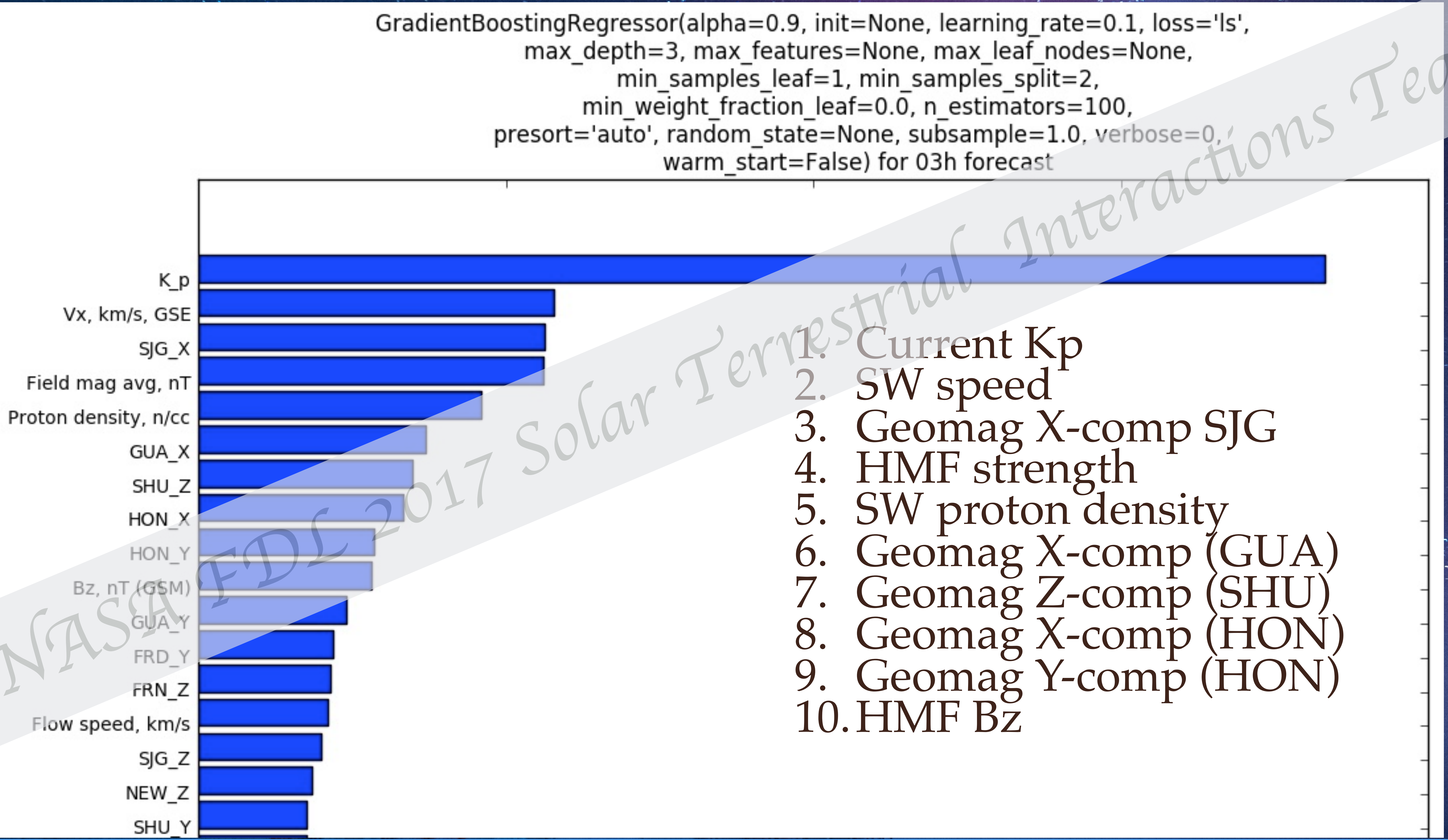
> 95% confidence level

Ranking

1. Kp
2. SW speed
3. IMF strength
4. Geomag X-comp SJG
5. SW proton density
6. Geomag X-comp (GUA)
7. Geomag Z-comp (SHU)
8. Geomag X-comp (HON)
9. Geomag Y-comp (HON)
10. IMF Bz



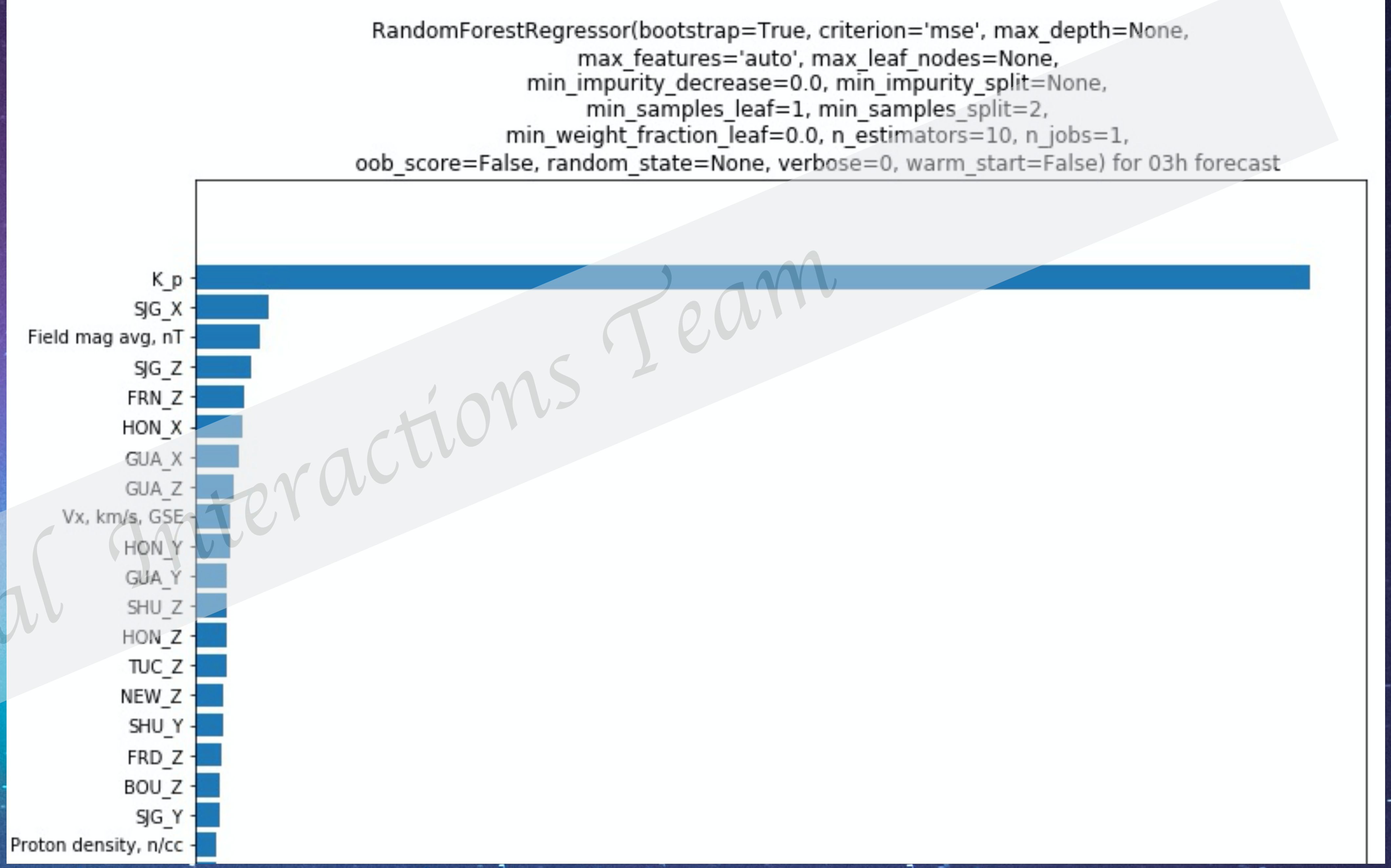
Ranking



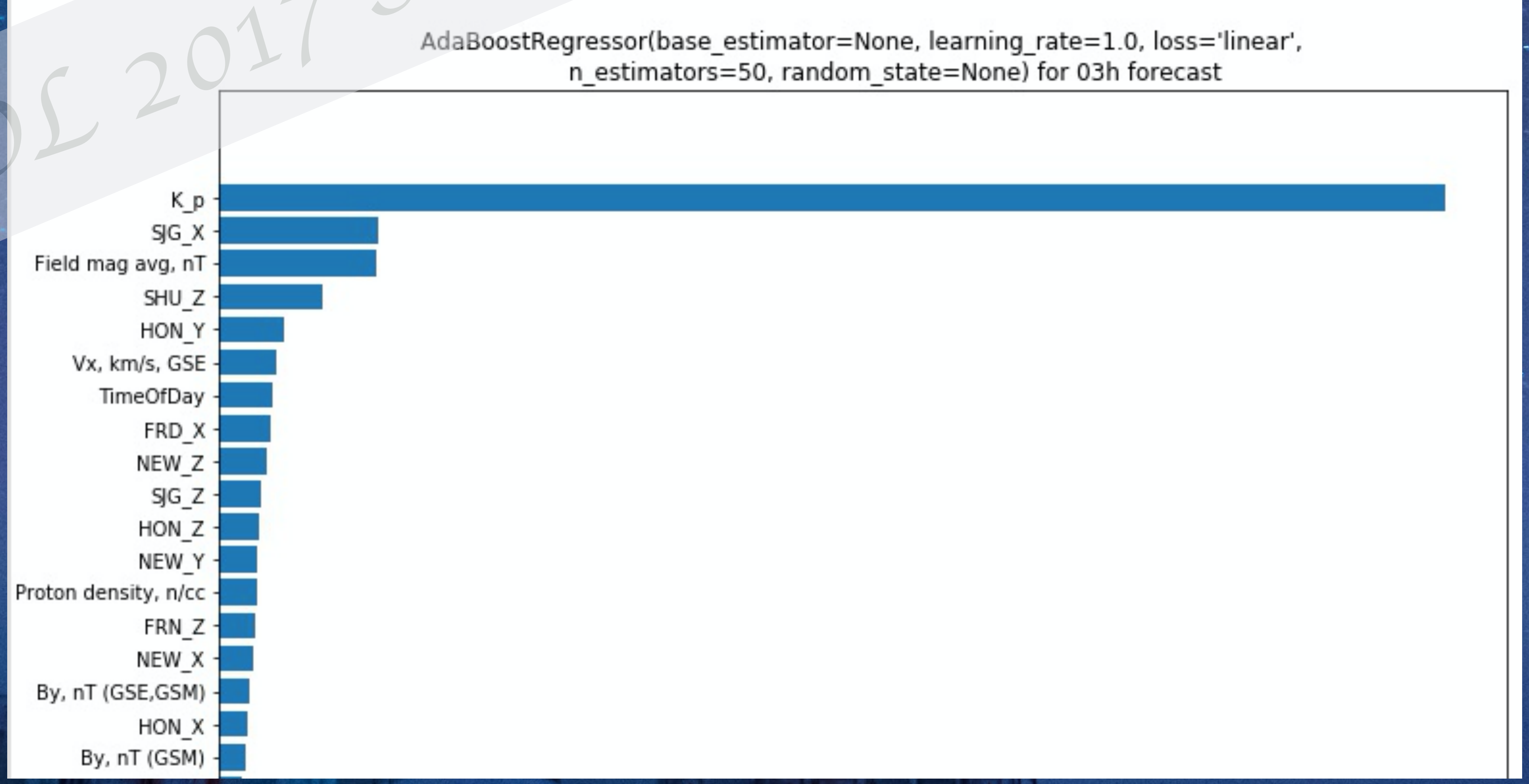
1. Current Kp
2. SW speed
3. Geomag X-comp SJG
4. HMF strength
5. SW proton density
6. Geomag X-comp (GUA)
7. Geomag Z-comp (SHU)
8. Geomag X-comp (HON)
9. Geomag Y-comp (HON)
10. HMF Bz

Ranking

ExtraTreesRegressor(bootstrap=False, criterion='mse', max_depth=None,
max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
oob_score=False, random_state=None, verbose=0, warm_start=False) for 03h forecast
K_p
SJG_X
HON_X
SJG_Z
HON_Y
GUA_X
GUA_Z
SHU_Z
FRD_Z
FRN_Z
GUA_Y
TUC_Z
HON_Z
BOU_Z
NEW_Z
BSL_X
SJG_Y
BSL_Z
TUC_X
SHU_Y
FRN_Y
FRD_X

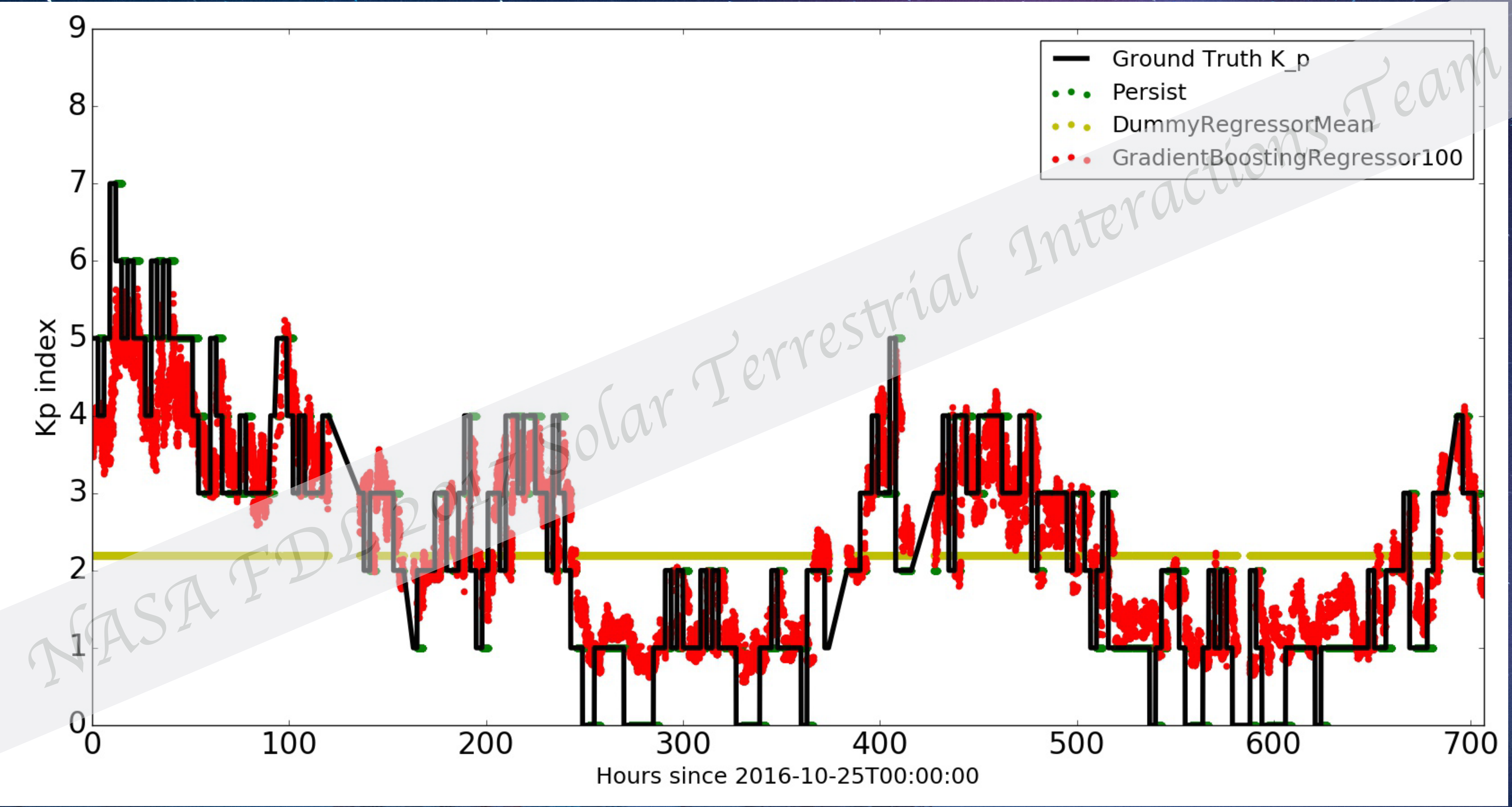


RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
oob_score=False, random_state=None, verbose=0, warm_start=False) for 03h forecast
K_p
SJG_X
Field mag avg, nT
SJG_Z
FRN_Z
HON_X
GUA_X
GUA_Z
Vx, km/s, GSE
HON_Y
GUA_Y
SHU_Z
HON_Z
TUC_Z
NEW_Z
SHU_Y
FRD_Z
BOU_Z
SJG_Y
Proton density, n/cc



AdaBoostRegressor(base_estimator=None, learning_rate=1.0, loss='linear',
n_estimators=50, random_state=None) for 03h forecast
K_p
SJG_X
Field mag avg, nT
SHU_Z
HON_Y
Vx, km/s, GSE
TimeOfDay
FRD_X
NEW_Z
SJG_Z
HON_Z
NEW_Y
Proton density, n/cc
FRN_Z
NEW_X
By, nT (GSE,GSM)
HON_X
By, nT (GSM)

Predicted Kp



Concluding Remarks

input features, according to their order of importance, as raked by GB model:

1. current Kp index
2. solar wind speed,
3. X-comp. geomagnetic field - SJG (San Juan, Puerto Rico, 18^{o} N),
4. HMF total strength,
5. solar wind proton density,
6. X-comp. geomagnetic field - GUA (Guam, 13^{o} N),
7. Z-comp. geomagnetic field - SHU (Shumagin, Alaska, 53^{o} N),
8. X-comp. of geomagnetic field - HON (Honolulu, Hawaii, 21^{o} N),
9. Y-comp. of geomagnetic field - HON,
10. HMF z-component, Bz (GSM).

Concluding Remarks

the N-S component of the geomagnetic field at lower latitude stations

important precursors

Guam (GUA), Hawaii (HON), Puerto Rico (SJG)

influenced largely by ring current

the model points to —> the importance of considering the effects of ring current in the prediction of geomagnetic storm

This result was a total surprise since the ML algorithm was not expected to be capable of learning such heuristics without prior knowledge!

Concluding Remarks

Scope

the method can be applied to address other aspects of the socio-economic impact of space weather by predicting the appropriate variable, provided that sufficient data exist in the public domain

Ultimate goal

To couple
the complex and dynamic
solar--terrestrial system using AI.

The
Solar Terrestrial Interactions Team
with the
Unexpected Discovery Award