Imbalanced regression with applications in space physics

Enrico Camporeale

Astronomy Unit, School of Physical and Chemical Sciences, Queen Mary University of London (UK)
and
University of Colorado (USA)

Acknowledgment: NASA grants 80NSSC20K1580, 80NSSC21K1555, 80NSSC20K1275, 80NSSC23M0192

University of Colorado Boulder

Queen Mary University of London

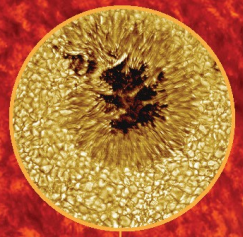
Astronomy Unit

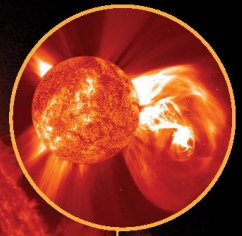
Space Weather

Space weather refers to the variable conditions on the Sun and in the space environment that can influence the performance and reliability of space-based and ground-based technological systems, as well as endanger life or health. Just like weather on Earth, space weather has its seasons, with solar activity rising and falling over an approximate 11 year cycle.

Sunspots

Sunspots are comparatively cool areas at up to 7,700° F and show the location of strong magnetic fields protruding through what we would see as the Sun's surface. Large, complex sunspot groups are generally the source of significant space weather.





Coronal Mass Ejections (CMEs)

Large portions of the corona, or outer atmosphere of the Sun, can be explosively blown into space, sending billions of tons of plasma, or superheated gas, Earth's direction. These CMEs have their own magnetic field and can slam into and interact with Earth's magnetic field, resulting in geomagnetic storms. The fastest of these CMEs can reach Earth in under a day, with the slowest taking 4 or 5 days to reach Earth.

Solar Wind

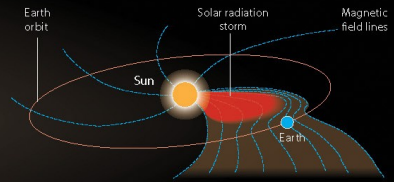
The solar wind is a constant outflow of electrons and protons from the Sun, always present and buffeting Earth's magnetic field. The background solar wind flows at approximately one million miles per hour!

Sun's Magnetic Field

Strong and ever-changing magnetic fields drive the life of the Sun and underlie sunspots. These strong magnetic fields are the energy source for space weather and their twisting, shearing, and reconnection lead to solar flares.

Solar Radiation Storms

Charged particles, including electrons and protons, can be accelerated by coronal mass ejections and solar flares. These particles bounce and gyrate their way through space, roughly following the magnetic field lines and ultimately bombarding Earth from every direction. The fastest of these particles can affect Earth tens of minutes after a solar flare.

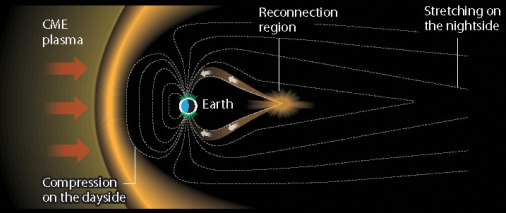


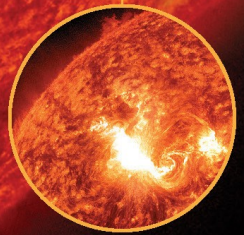
Earth

Corona

Geomagnetic Storms

A geomagnetic storm is a temporary disturbance of Earth's magnetic field typically associated with enhancements in the solar wind. These storms are created when the solar wind and its magnetic field interacts with Earth's magnetic field. The primary source of geomagnetic storms is CMEs which stretch the magnetosphere on the nightside causing it to release energy through magnetic reconnection. Disturbances in the ionosphere (a region of Earth's upper atomosphere) are usually associated with geomagnetic storms.





Solar Flares

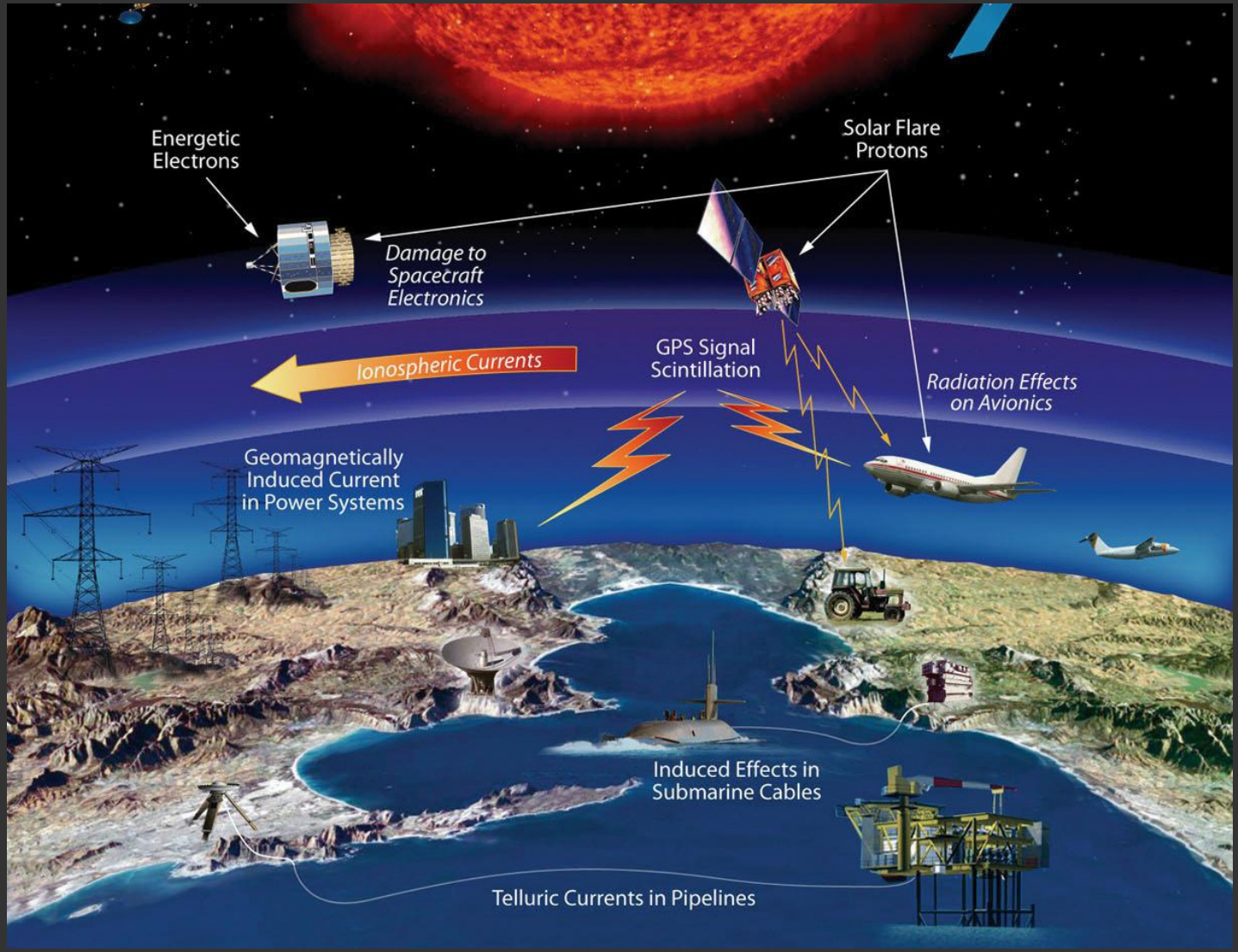
Reconnection of the magnetic fields on the surface of the Sun drive the biggest explosions in our solar system. These solar flares release immense amounts of energy and result in electromagnetic emissions spanning the spectrum from gamma rays to radio waves. Traveling at the speed of light, these emissions make the 93 million mile trip to Earth in just 8 minutes.

Earth's Magnetic Field

Earth's magnetic field, largely like that of a bar magnet, gives the Earth some protection from the effects of the Sun. Earth's magnetic field is constantly compressed on the day side and stretched on the night side by the ever-present solar wind. During geomagnetic storms, the disturbances to Earth's magnetic field can become extreme. In addition to some buffering by the atmosphere, this field also offers some shielding from the charged particles of a radiation storm.

NOAA

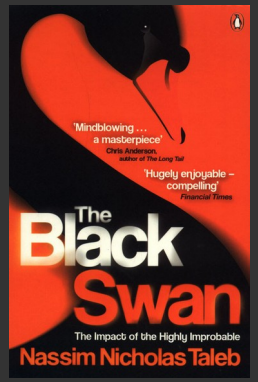
Source images: NASA, NOAA.

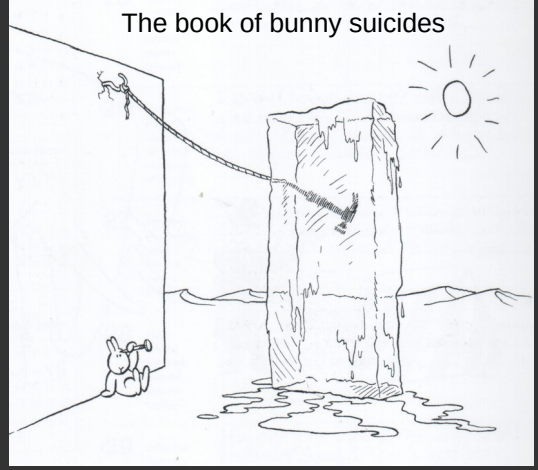


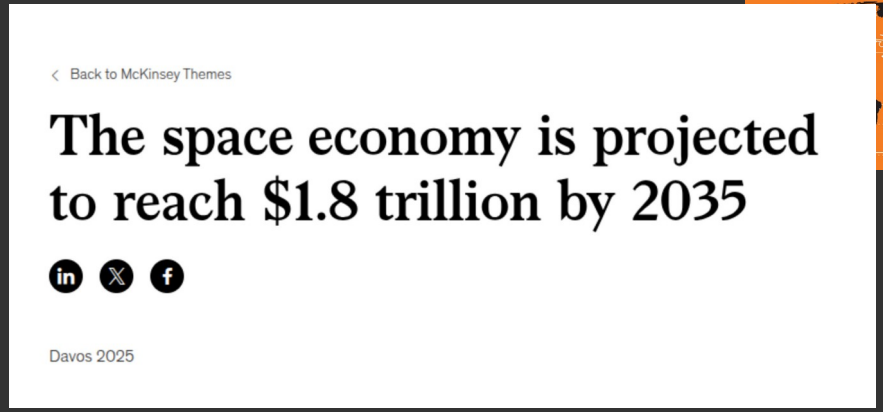
Credits:
NASA

A disaster waiting to happen…

- If a space weather event causes a power outage, we estimate costs to U.S. electricity consumers that may be ~$400 million to ~$10 billion for a moderate event and ~$1 billion to ~$20 billion for a more extreme event.
- The adverse impact of space weather is estimated to cost **$200–$400 million** per year;
- Losses to satellite companies range from thousands of dollars for temporary data outages up to **$200 million** to replace a satellite





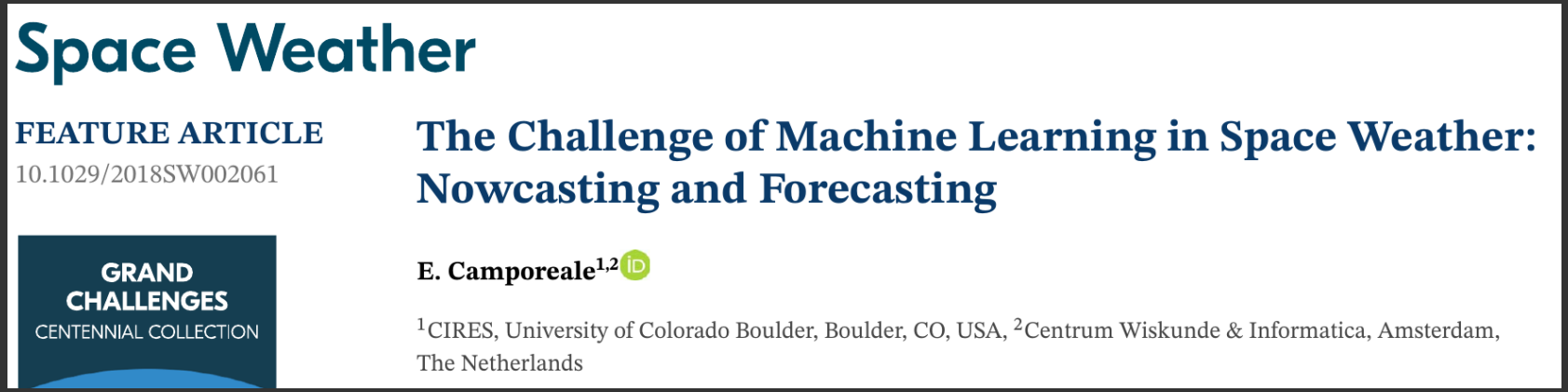


ML in Heliophysics

- In the last 10 years or so, Machine Learning has been applied to ANY task in space weather forecasting.

- Typically, a ML algorithm claims to be outperforming its physics-based equivalent (in terms of accuracy, computational cost, etc)

Open Challenges



Space Weather

FEATURE ARTICLE

10.1029/2018SW002061

GRAND CHALLENGES
CENTENNIAL COLLECTION

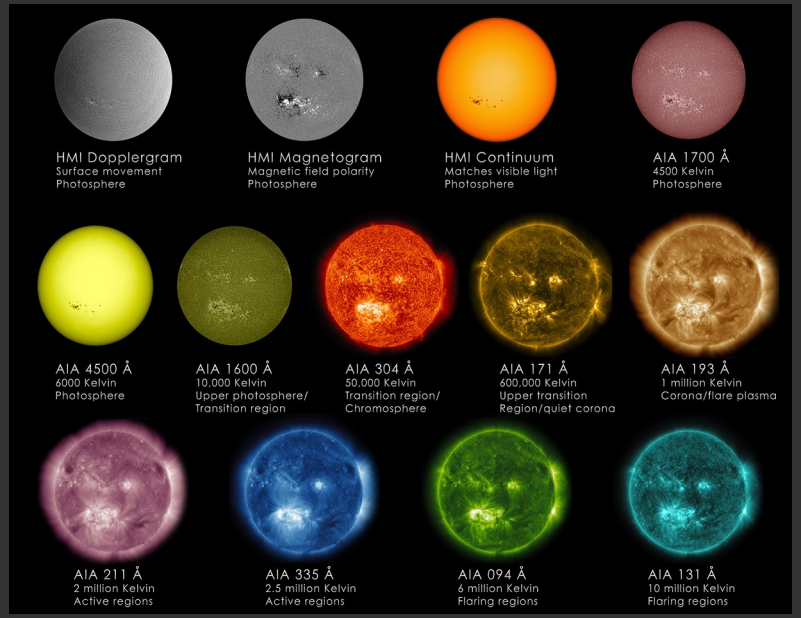
The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting

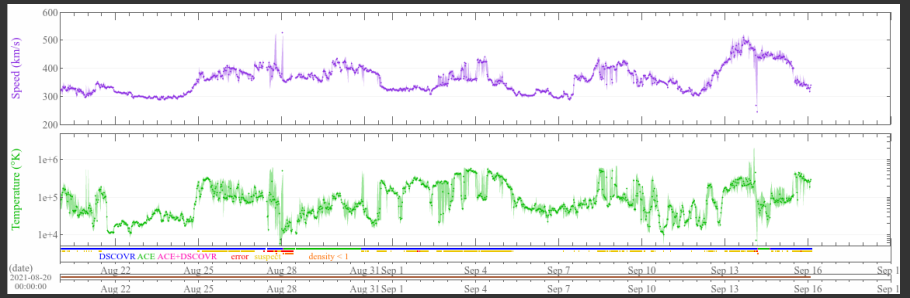
E. Camporeale[1,2]

[1]CIRES, University of Colorado Boulder, Boulder, CO, USA, [2]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Open Challenges

The information problem: What is the minimal physical information required to make a forecast? (and how to extract features...)

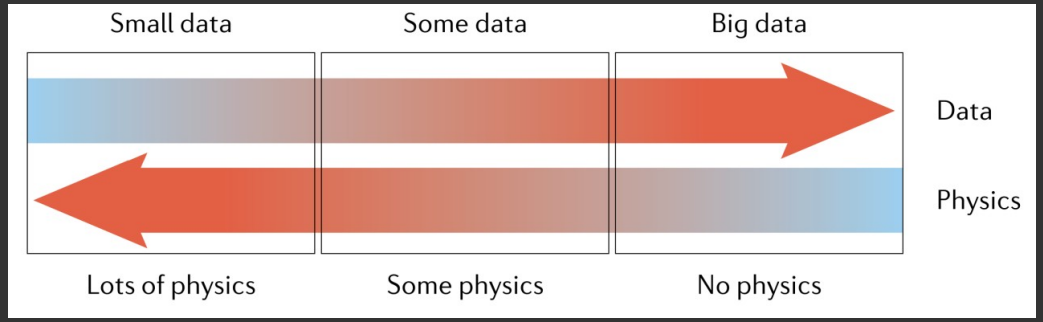




200M pixels → 1 scalar value

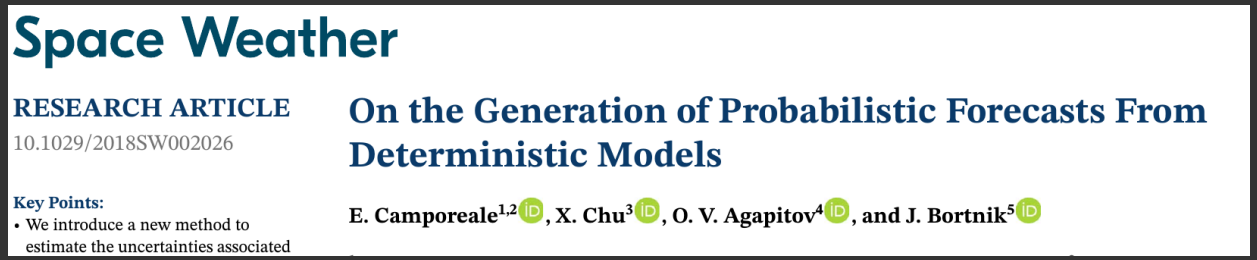
Open Challenges

The gray-box problem: What is the best way to make an optimal use of both our physical understanding and our large amount of data in the Sun-Earth system?



Open Challenges

The uncertainty problem: How to estimate the uncertainty of a deterministic model? And how to propagate those uncertainties?



Space Weather

RESEARCH ARTICLE

10.1029/2018SW002026

Key Points:

- We introduce a new method to estimate the uncertainties associated

On the Generation of Probabilistic Forecasts From Deterministic Models

E. Camporeale[1,2], X. Chu[3], O. V. Agapitov[4], and J. Bortnik[5]

International Journal for Uncertainty Quantification, 11(4):81–94 (2021)

ACCRUE: ACCURATE AND RELIABLE UNCERTAINTY ESTIMATE IN DETERMINISTIC MODELS

Enrico Camporeale[1,] & Algo Carè[2]*

Open Challenges

- *The too often too quiet problem:* Space weather data sets are typically imbalanced: many days of quiet conditions and a few hours of storms. This poses a serious problem for any machine learning algorithm. It is also problematic for defining meaningful metrics that actually assess the ability of a model to predict interesting but rare events.

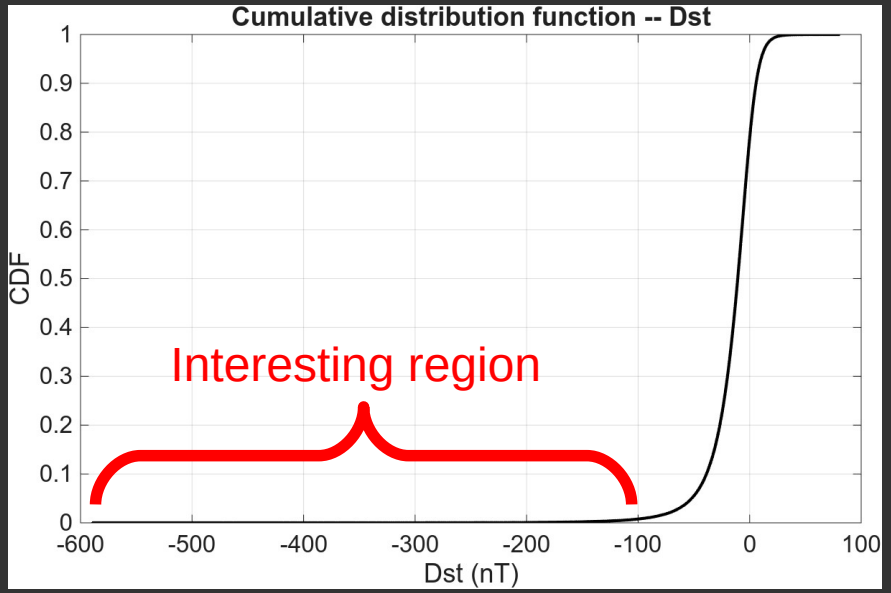
JOURNAL ARTICLE

More Data or Better Data? A Statistical Decision Problem Get access ›

Jeff Dominitz, Charles F. Manski

The Review of Economic Studies, Volume 84, Issue 4, October 2017, Pages 1583–1605, https://doi.org/10.1093/restud/rdx005

Published: 20 February 2017 **Article history** ▾



Three engines under the hood
(Take home messages…)

- **ACCRUE** (ACCurate and Reliable Uncertainty Estimate):
  - Takes a deterministic model and makes it into a probabilistic one

International Journal for Uncertainty Quantification, 11(4):81–94 (2021)

ACCRUE: ACCURATE AND RELIABLE UNCERTAINTY ESTIMATE IN DETERMINISTIC MODELS

Enrico Camporeale[1,*] *& Algo Carè*[2]

[1]*University of Colorado, Boulder, Colorado, USA*

iversity of Brescia, Brescia, Italy

dress all correspondence to: Enrico Camporeale, University of Colorado, Boulder, Colorado, USA,
ail: enrico.camporeale@colorado.gov

Space Weather

RESEARCH ARTICLE
10.1029/2018SW002026

On the Generation of Probabilistic Forecasts From Deterministic Models

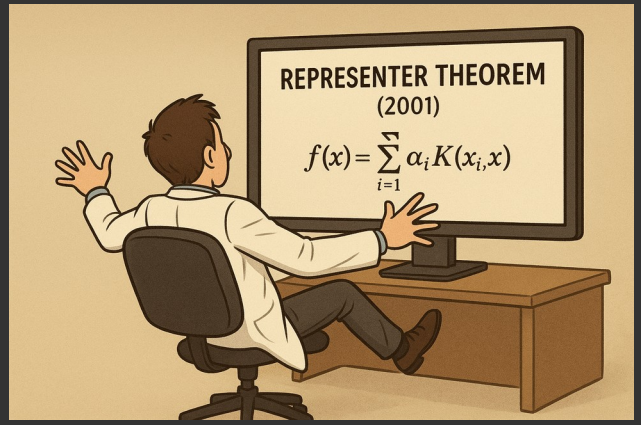
E. Camporeale[1,2], X. Chu[3], O. V. Agapitov[4], and J. Bortnik[5]

[1]Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, [2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, [3]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA, [4]Space Sciences Laboratory, University of California Berkeley, Berkeley, CA, USA, [5]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA

Key Points:
- We introduce a new method to estimate the uncertainties associated with single-point outputs generated by a deterministic model
- The method ensures a trade-off between accuracy and reliability of the generated probabilistic forecasts
- Computationally cheap model:

Three engines under the hood
(Take home messages…)

- **ACCRUE** (ACCurate and Reliable Uncertainty Estimate):
  - Takes a deterministic model and makes it into a probabilistic one
- **PARIS** (Pruning via Analysis of Representer-based Influence Scores)
  - Prunes a dateset to remove uninformative/redundant samples and improve performance for an imbalanced regression problem



(generate an image of a researcher just learning about the representer theorem and literally falling out his chair)

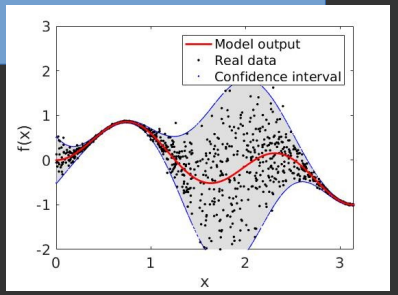
Three engines under the hood
(Take home messages…)

- **ACCRUE** (ACCurate and Reliable Uncertainty Estimate):
  - Takes a deterministic model and makes it into a probabilistic one
- **PARIS** (Pruning via Analysis of Representer-based Influence Scores)
  - Prunes a dateset to remove uninformative/redundant samples and improve performance for an imbalanced regression problem
- **ProBoost** (Probabilistic Boosting)
  - Create ensemble of uncertainty-aware sub-models, each one specialized on a subset of feature space

    (more or less probabilistic extension of Gradient Boost)

ACCRUE model

Deterministic model → **ACCRUE** → Probabilistic model



A-posteriori UQ estimate of a deterministic model (enforcing accuracy and reliability)

Space Weather

RESEARCH ARTICLE
10.1029/2018SW002026

On the Generation of Probabilistic Forecasts From Deterministic Models

E. Camporeale[1,2], X. Chu[3], O. V. Agapitov[4], and J. Bortnik[5]

[1]Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, [2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, [3]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA, [4]Space Sciences Laboratory, University of California Berkeley, Berkeley, CA, USA, [5]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA

Key Points:
- We introduce a new method to estimate the uncertainties associated with single-point outputs generated by a deterministic model
- The method ensures a trade-off between accuracy and reliability of the generated probabilistic forecasts
- Computationally cheap model:

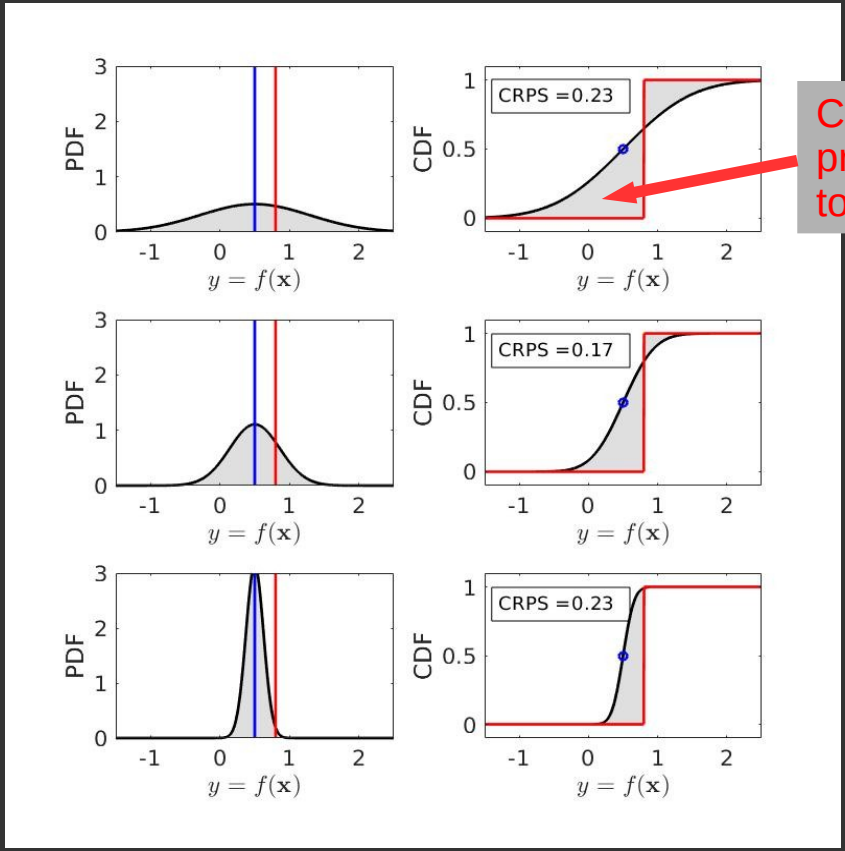
- No need to run ensembles
- Plug-and-play
- Code available on https://zenodo.org/record/1485608

Accuracy of a probabilistic forecast (continuous output)

- **Several metrics to define ‘accuracy’. We use Continuous Rank Probability Score (CRPS)**
- **CRPS is a generalization of Brier score**
- **It has a simple graphical interpretation**
- **CRPS = 0 for perfect forecast**
- **CRPS =** $\int \left(C\left(y\right)-H\left(\hat{y}\right)\right)^2 dy$

Empirical CDF

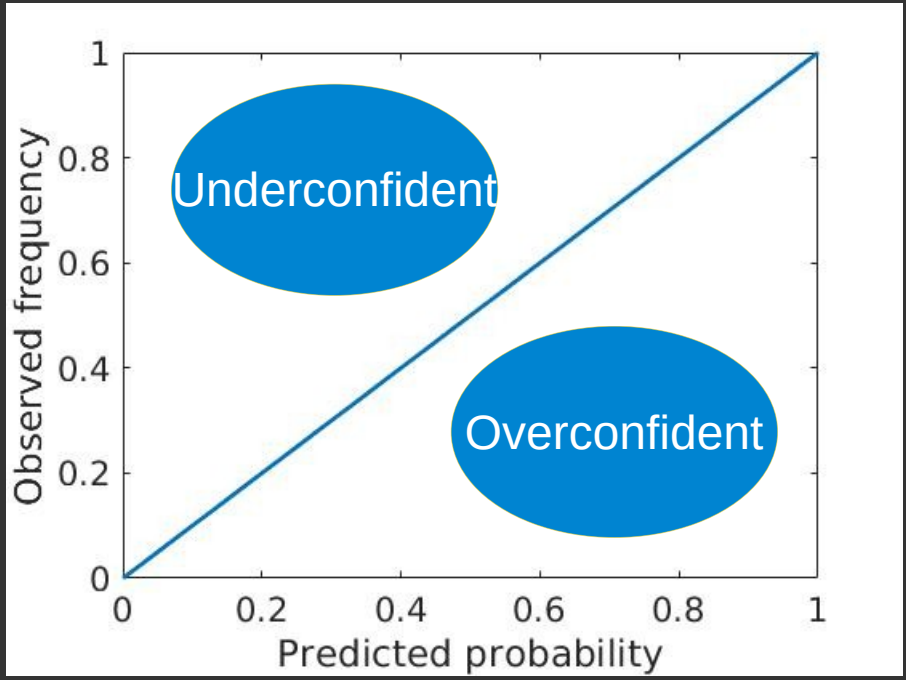
Step function (Heaviside)



Reliability diagram

Reliability is the property of a probabilistic model that measures its **statistical consistency with observations.**

For example, for forecasts of discrete events, the reliability measures if an event occurs on average with frequency *p*, when it has been predicted to occur with probability *p*.



ACCRUE Recipe

- The optimal Gaussian width (std) is the one that optimizes both accuracy and reliability
- This is a two-objective optimization problem, because reliability and accuracy are competing objectives.
- We define the Accuracy-Reliability (AR) cost function:

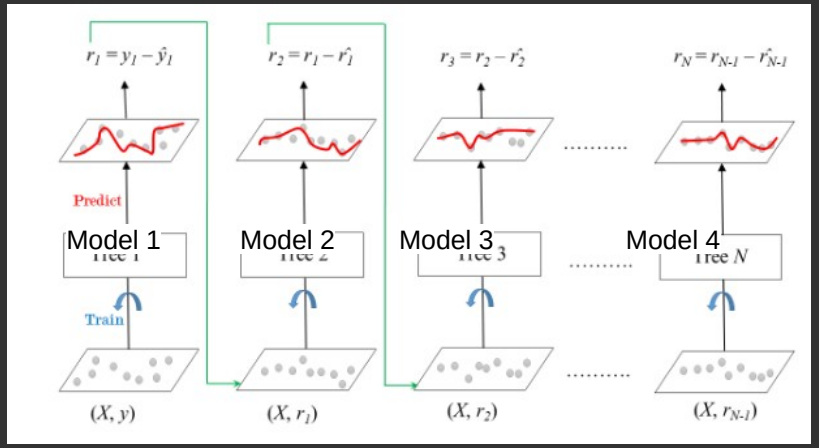
$$AR = CRPS + RS$$

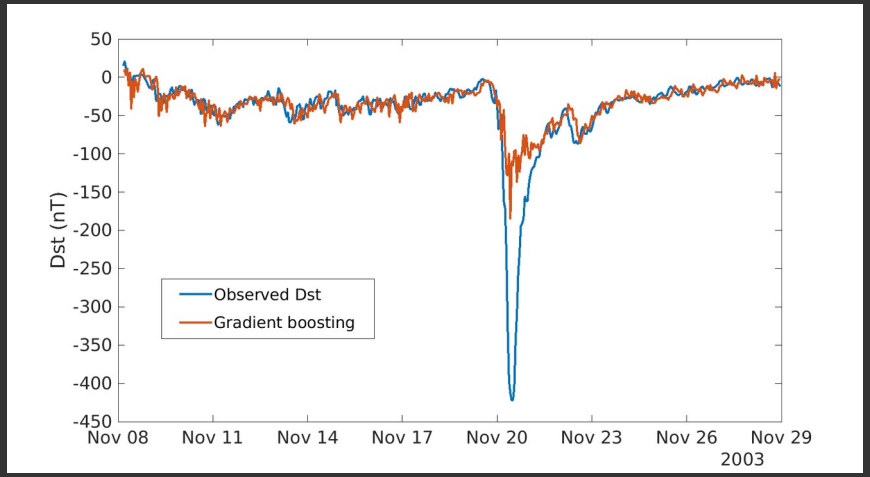
Accuracy Reliability

- Accuracy and Reliability cannot both be minimized simultaneously and we have to find the best trade-off
- The minimization problem can be solved in a number of ways
  - In multiple dimensions, a deep neural network works well to minimize the cost function

Gradient Boosting in one slide

- A hierarchy of models is built
- Each model is trained on the errors of the previous one
- The final model is an additive combination of all sub-models
- One of the strongest algorithm
  - But it under-performs on imbalanced datasets





ProBoost in one slide

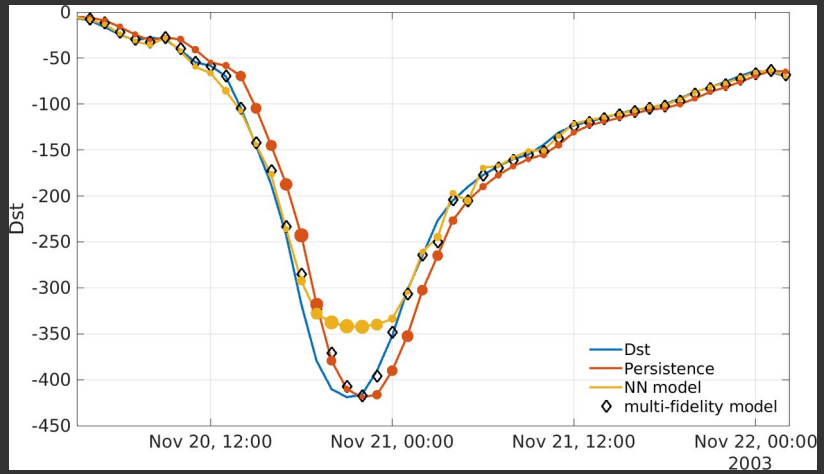
- We can use the power of ACCRUE to identify regions (in feature space) where a predictor works well and where it doesn’t

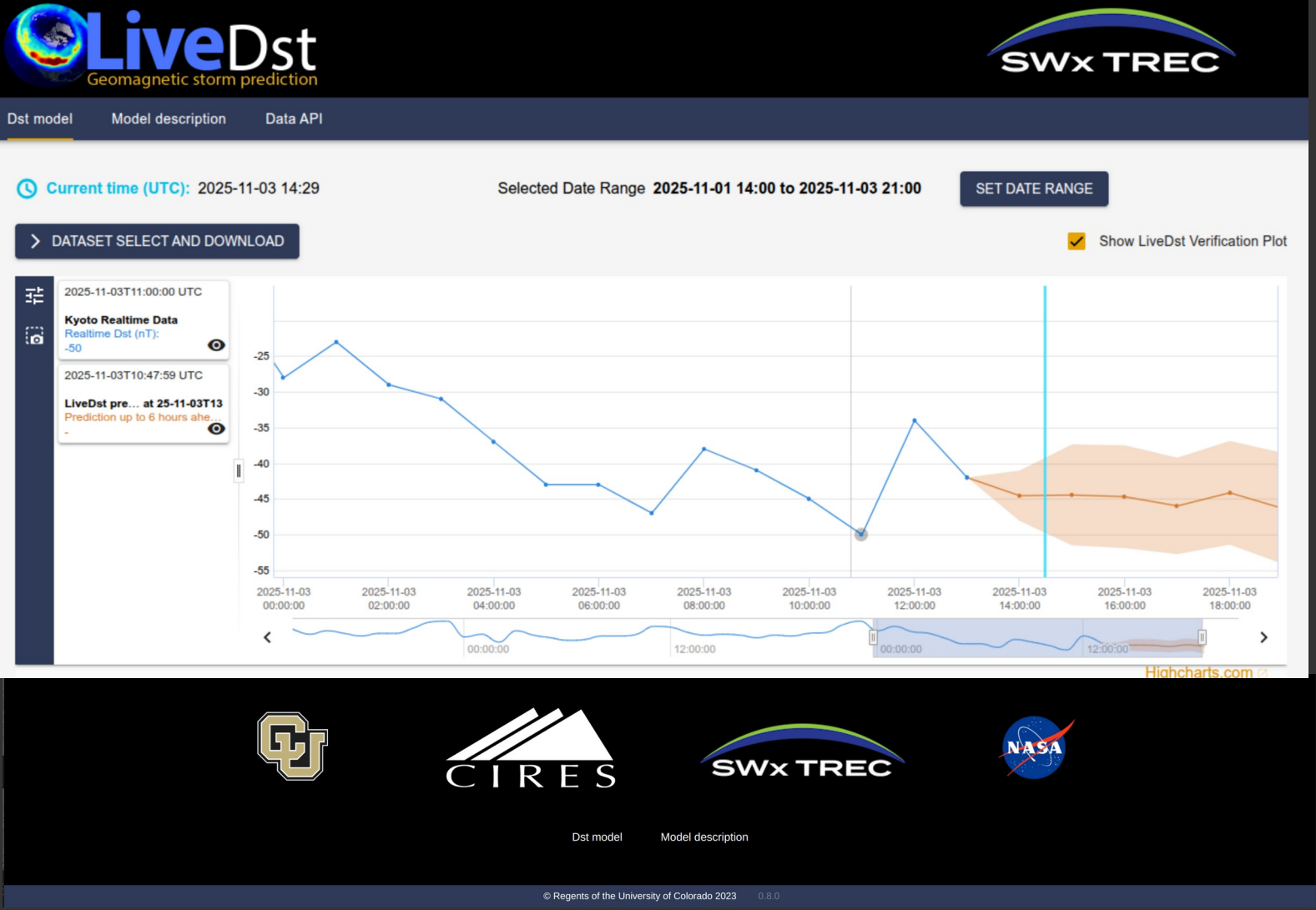
ProBoost in one slide

- We can use the power of ACCRUE to identify regions (in feature space) where a predictor works well and where it doesn’t
- An ensemble of models is built by sub-sampling the training set
  - The sub-sampling criterion is based on the ACCRUE uncertainty

ProBoost in one slide

- We can use the power of ACCRUE to identify regions (in feature space) where a predictor works well and where it doesn't
- An ensemble of models is built by sub-sampling the training set
  - The sub-sampling criterion is based on the ACCRUE uncertainty
- Sub-models are combined weighted by their precision = $1/\sigma^2$





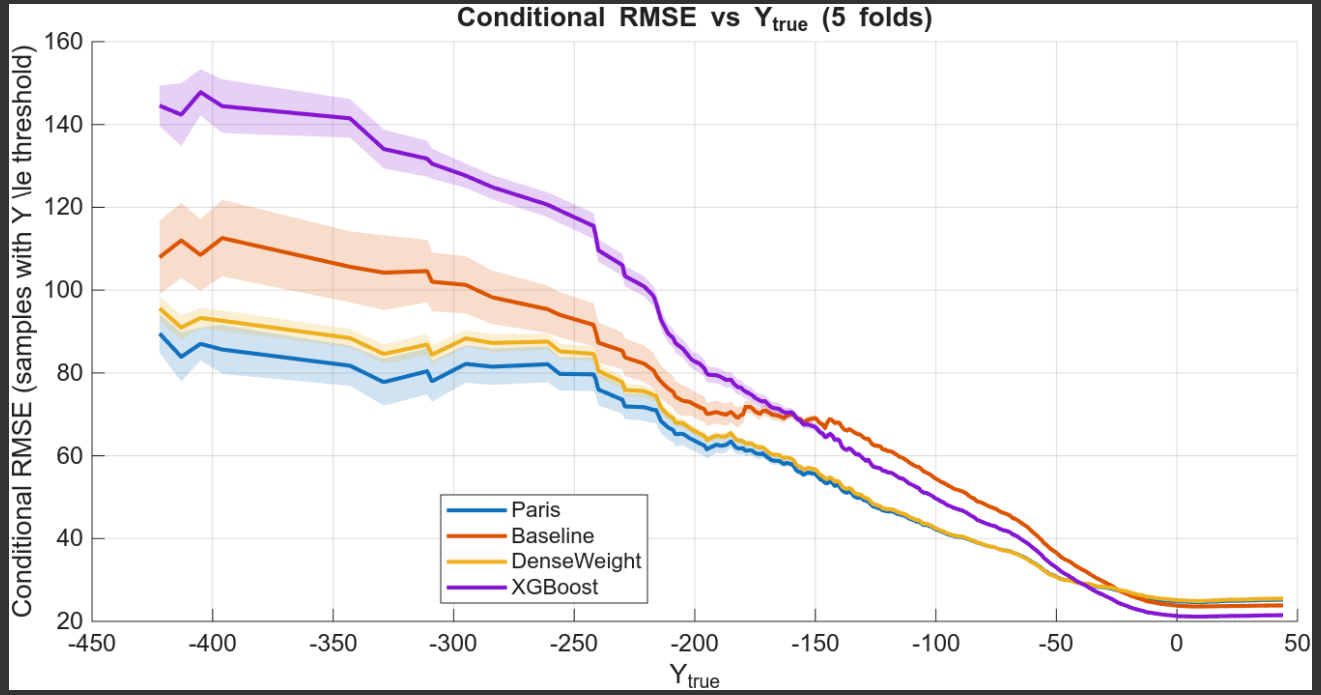
Real-time 1-to-6 hours ahead probabilistic forecast of Dst

University of Colorado portal

https://swx-trec.com/dst/

PARIS

- Simple idea: pruning the training dataset to get rid of uninformative/harmful samples

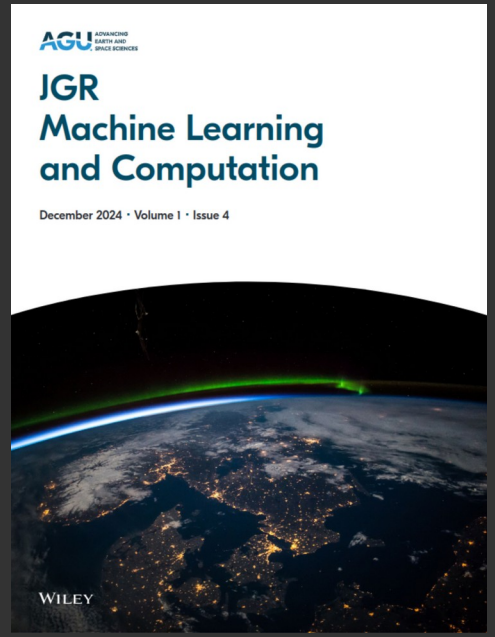


Summary

I apply ML techniques to problems in space physics and space weather.

Some of the open problems I am interested in:

- *Feature extraction (the information problem)*
- *Uncertainty quantification for ML*
- *Extreme and rare events (the too often too quiet problem)*



Enrico.Camporeale@qmul.ac.uk