Ensemble Learning for Accurate and Reliable Uncertainty Quantification

Enrico Camporeale

CIRES, CU Boulder & NOAA Space Weather Prediction Center

This project is supported by NASA under grant 80NSSC20K1580

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

CIRES

SWx TREC

UCLA

PI: Enrico Camporeale (CIRES, CU Boulder)

Co-PIs: Jacob Bortnik (UCLA)
Rebecca Morrison (CS, CU Boulder)

Senior personnel: Thomas Berger (SWx TREC, CU Boulder)
Claire Monteleoni (CS, CU Boulder)

Postdocs: Andong Hu (CIRES, CU Boulder)
Man Hua (UCLA)
One position still open...

Grad students: Rileigh Bandy (CS, CU Boulder)

NOAA/SWPC liason: Howard Singer, Vic Pizzo, Terry Onsager, Eric Adamson

This project is supported by NASA under grant 80NSSC20K1580

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

CIRES

SWx TREC

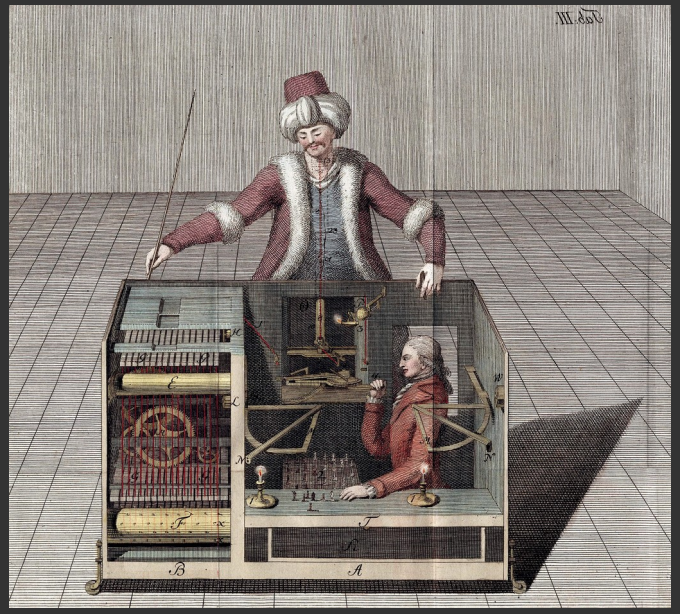
UCLA

Goals

- To create the algorithmic prototype that will combine a small number of high-fidelity (low error, but computationally expensive) runs from physics-based models with a large number of (possibly) less accurate but faster runs from machine learning models
- The goal is to obtain a more accurate overall prediction than any individual model, and an estimation of the associated uncertainties.

The engines under the hood

- ACCRUE (ACCurate and Reliable Uncertainty Estimate*)
  - A-posteriori UQ estimate of a deterministic model
  - Ref: Camporeale, E., & Carè, A. (2021). ACCRUE: Accurate and reliable uncertainty estimate in deterministic models. *International Journal for Uncertainty Quantification*, 11(4).
- ProBoost (Probabilistic Boosting)
  - A new multi-fidelity boosting method based on ACCRUE that builds a hierarchy of models to be combined in an ensemble
  - Ref: in preparation!



https://en.wikipedia.org/wiki/Mechanical_Turk
(Artificial Intelligence in the 18th century)

* Acronym graciously suggested by Gabor Toth (as of now one of the best outcomes of a SWQU teams collaboration)

But first… let's agree on some basics

- A forecast is not a forecast if it's not probabilistic
- But what does a forecast probability mean?
  - That is, X% chance of what??

974 WEATHER AND FORECASTING VOLUME 23

Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public

REBECCA E. MORSS, JULIE L. DEMUTH, AND JEFFREY K. LAZO
National Center for Atmospheric Research, Boulder, Colorado*

Risk Analysis, Vol. 25, No. 3, 2005 DOI: 10.1111/j.1539-6924.2005.00608.x

"A 30% Chance of Rain Tomorrow": How Does the Public Understand Probabilistic Weather Forecasts?

Gerd Gigerenzer,[1]* Ralph Hertwig,[2] Eva van den Broek,[1] Barbara Fasolo,[1] and Konstantinos V. Katsikopoulos[1]

Diversity in Interpretations of Probability: Implications for Weather Forecasting

RAMÓN DE ELÍA AND RENÉ LAPRISE
Département des Sciences de la Terre et de l'Atmosphère, Université du Québec à Montréal, Montréal, Québec, Canada

(Manuscript received 11 February 2004, in final form 5 October 2004)

Active poll

0

Join at
slido.com
#2694 454

There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?

I have run an ensemble of models and 70% of them predict Dst=-300
0%

70% of all the hours with the exact same magnetospheric/solar wind conditions observed now have a Dst = -300 (one hour after)
0%

70% of all hours with Dst=-300 have the same magnetospheric/solar wind conditions as the one observed now (one hour before)
0%

70% of the time that a 70% chance was predicted, the forecast event occurred.
0%

Active poll 0

Join at
slido.com
#2694 454

There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?

I have run an ensemble of models and 70% of them predict Dst=-300
0%

Although this might be a way of estimating uncertainties,
this answer might be correct only
if your model is perfectly calibrated.

Active poll 0

Join at
slido.com
#2694 454

There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?

I have run an ensemble of models and 70% of them predict Dst=-300
0%

70% of all the hours with the exact same magnetospheric/solar wind conditions observed now have a Dst = -300 (one hour after)
0%

This answer is theoretically correct.
However it makes the model reliability
not verifiable/computable
and it is effectively dependent on
the features one chooses to define
the conditions of magnetosphere/solar wind

Active poll

0

Join at
slido.com
#2694 454

There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?

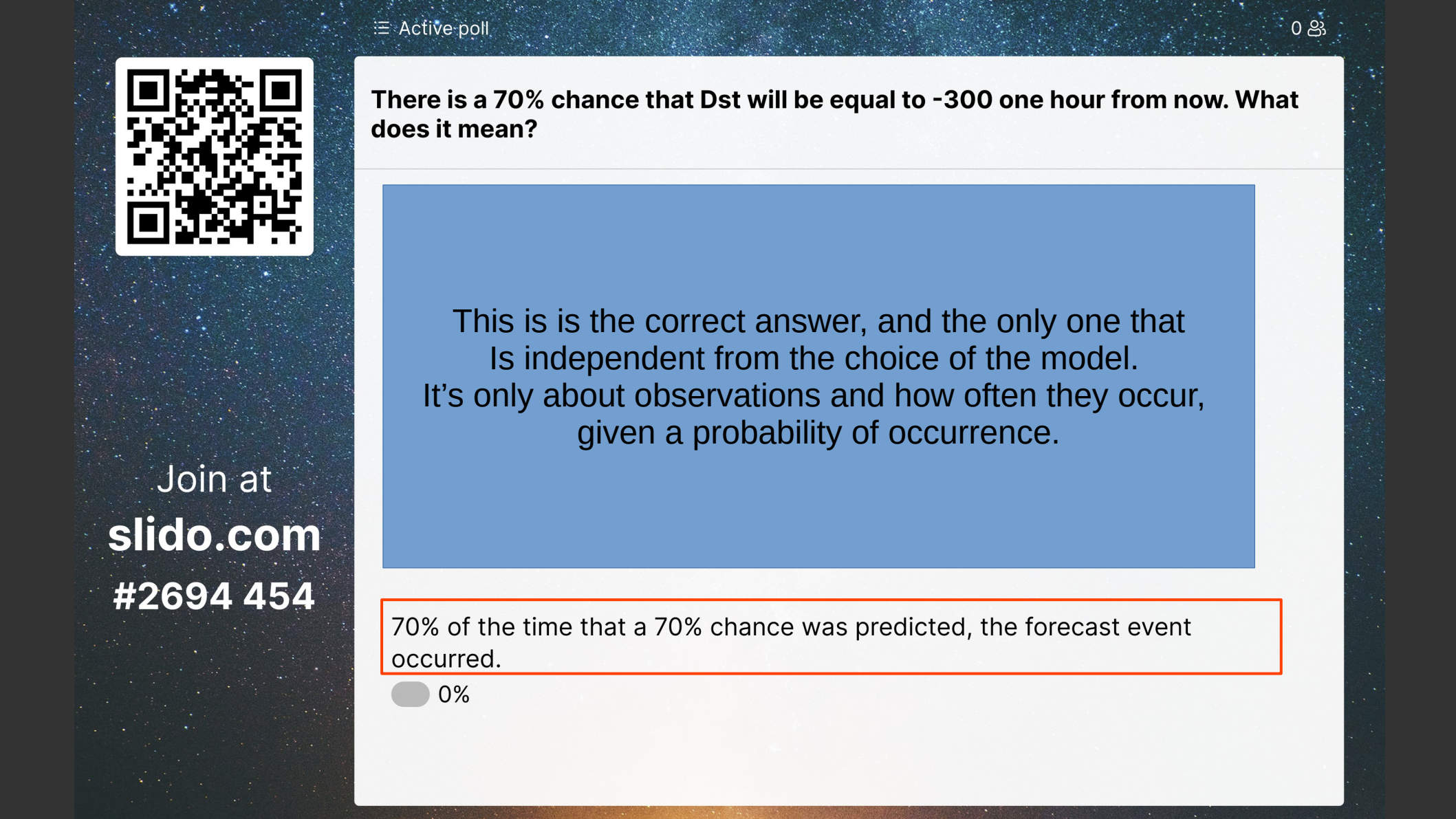
This is just wrong!

70% of all hours with Dst=-300 have the same magnetospheric/solar wind conditions as the one observed now (one hour before)

0%

70% of the time that a 70% chance was predicted, the forecast event occurred.

0%


Active poll
0
There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?
This is is the correct answer, and the only one that
Is independent from the choice of the model.
It's only about observations and how often they occur,
given a probability of occurrence.
70% of the time that a 70% chance was predicted, the forecast event occurred.
0%
Join at
slido.com
#2694 454

Active poll

0

Join at
slido.com
#2694 454

There is a 70% chance that Dst will be equal to -300 one hour from now. What does it mean?

I have run an ensemble of models and 70% of them predict Dst=-300
0%

70% of all the hours with the exact same magnetospheric/solar wind conditions observed now have a Dst = -300 (one hour after)
0%

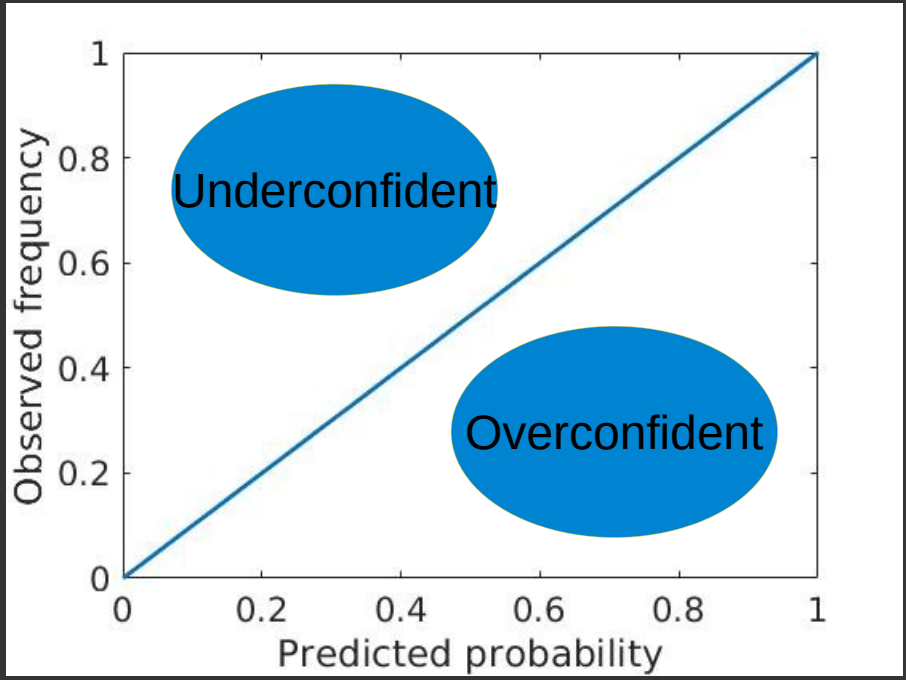
70% of all hours with Dst=-300 have the same magnetospheric/solar wind conditions as the one observed now (one hour before)
0%

70% of the time that a 70% chance was predicted, the forecast event occurred.
0%

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*.

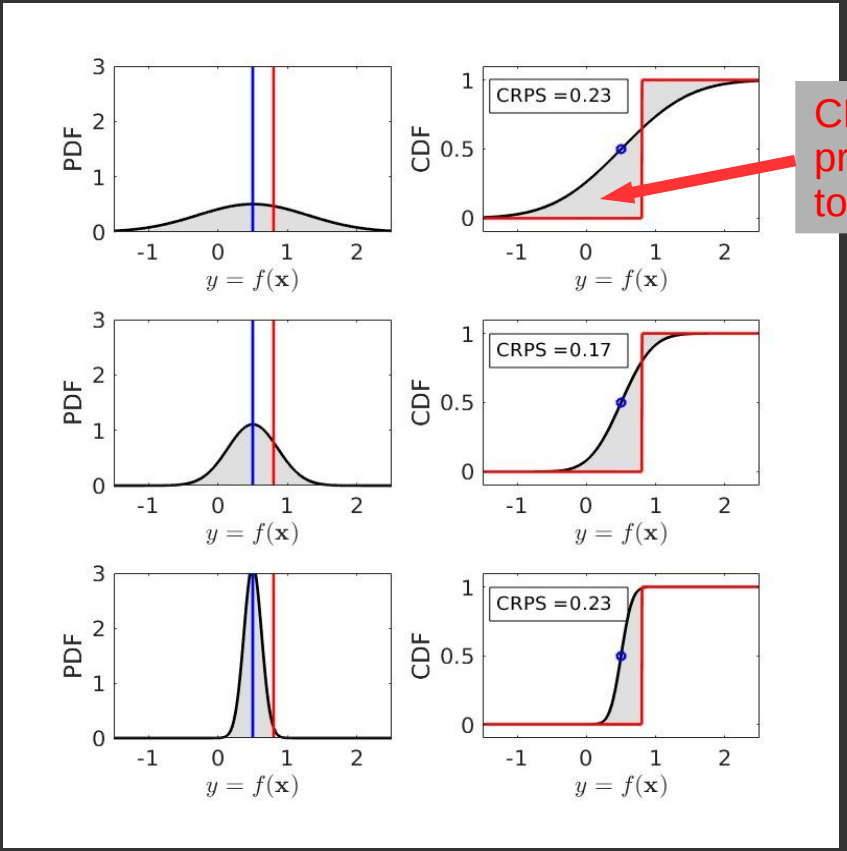


Accuracy of a probabilistic forecast (continuous output)

- **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 =** $\int \left(C\left(y\right)-H\left(\hat{y}\right)\right)^2 dy$

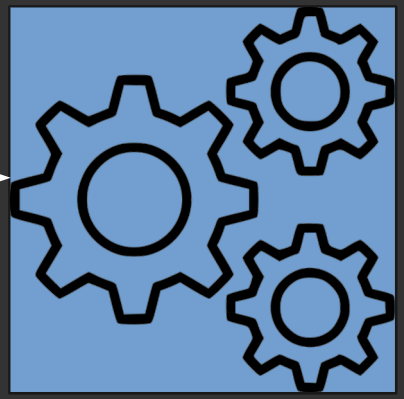
Empirical CDF

Step function (Heaviside)

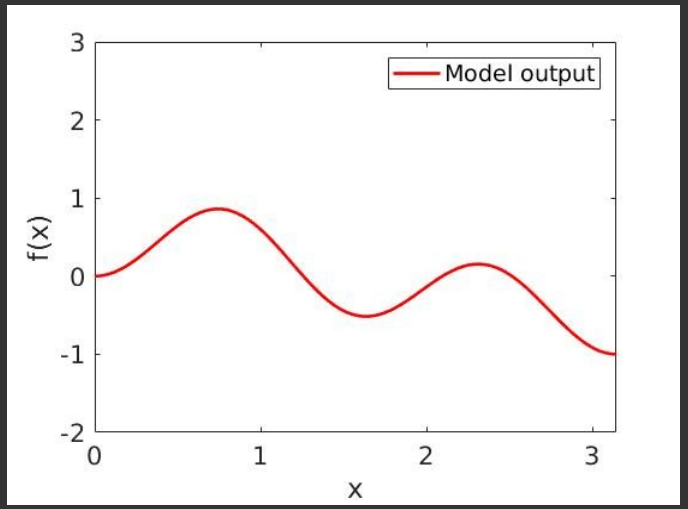


ACCRUE: Problem Statement

Set of Inputs, Parameters, etc.

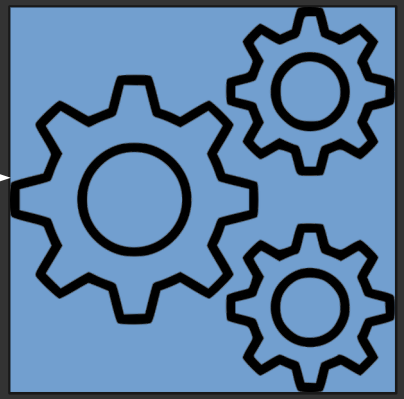


Deterministic model

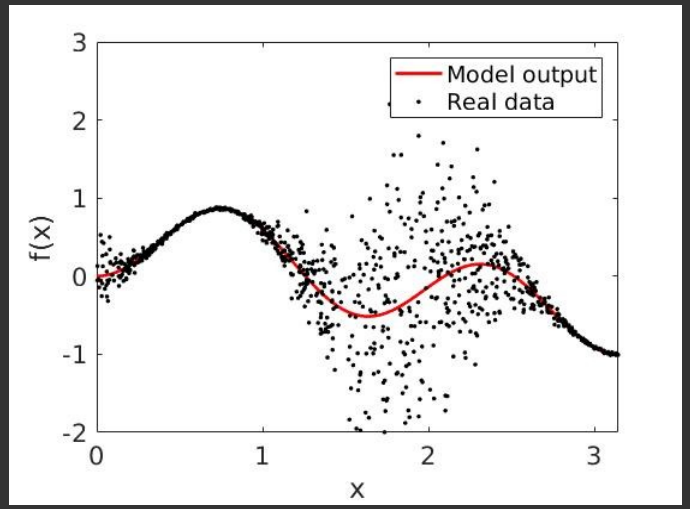


ACCRUE: Problem Statement

Set of Inputs, Parameters, etc.



Deterministic model + Real Data

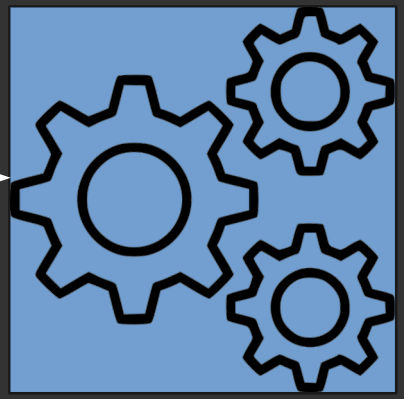


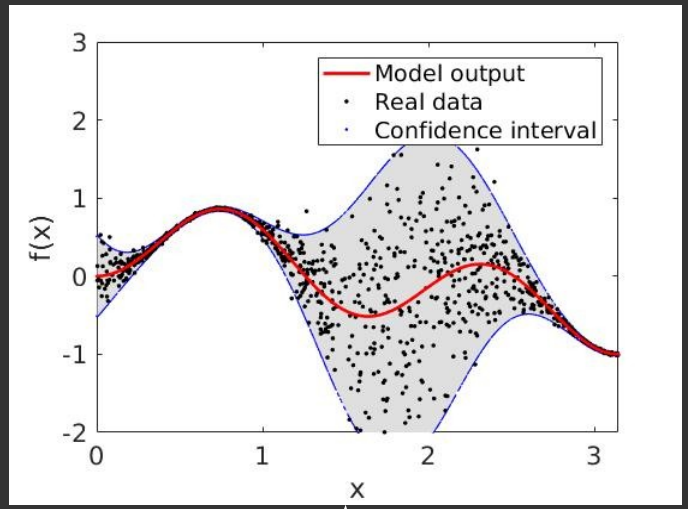
Uncertainties:

- Epistemic (do not know the parameters exactly)
- Aleatoric (physics that is not in the model)
- Algorithmic (numerical errors)

ACCRUE: Problem Statement

Set of Inputs, Parameters, etc.





Deterministic model + Real Data

Uncertainties:

- Epistemic (do not know the parameters exactly)
- Aleatoric (physics that is not in the model)
- Algorithmic (numerical errors)

?

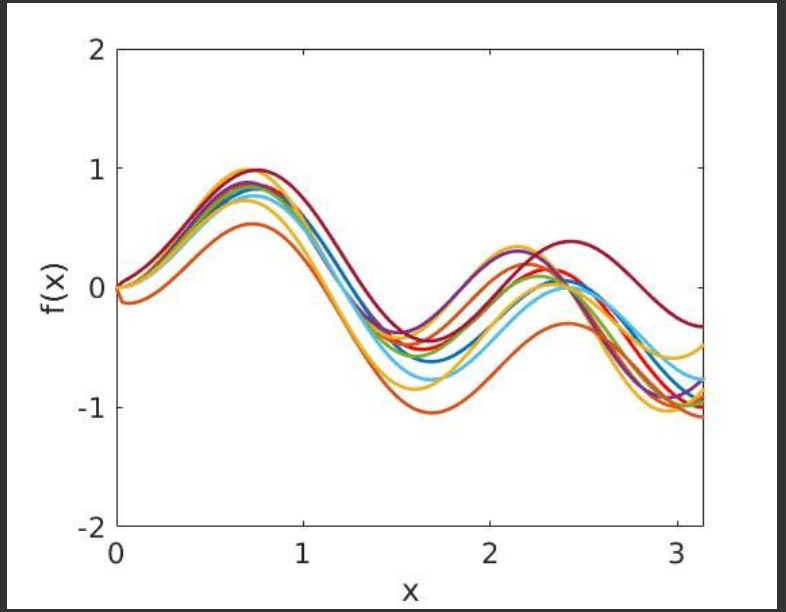
Generate heteroskedastic uncertainty

ACCRUE: Problem Statement

- The golden standard approach to estimate uncertainties based on a deterministic model is by running a

  Monte Carlo ensemble
  (e.g. by small perturbations of initial conditions)

- This has two problems:
  –
  –

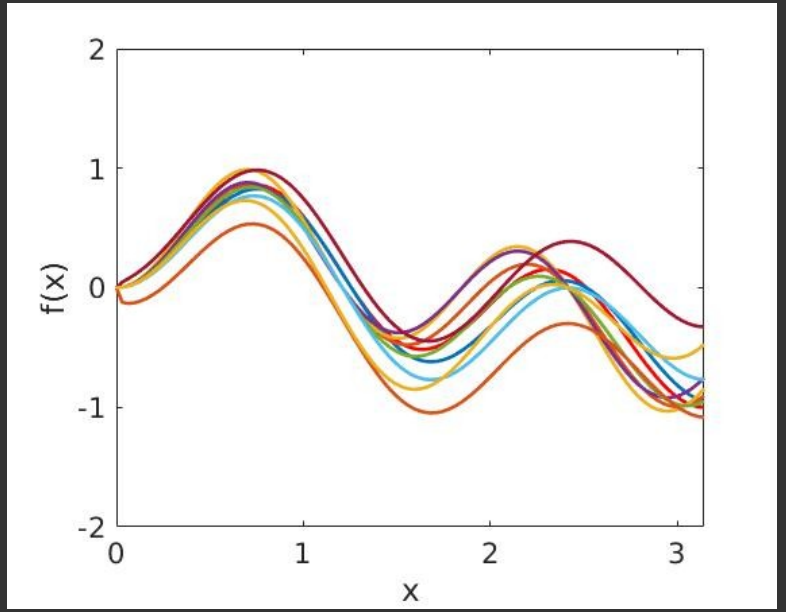


ACCRUE: Problem Statement

- The golden standard approach to estimate uncertainties based on a deterministic model is by running a

  Monte Carlo ensemble
  (e.g. by small perturbations of initial conditions)

- This has two problems:
  - It's expensive: it requires many runs
  -

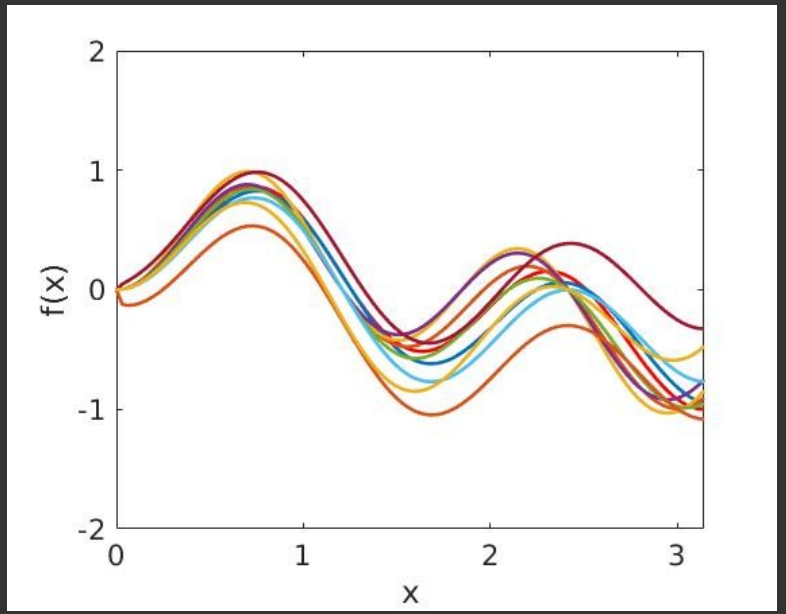


ACCRUE: Problem Statement

- The golden standard approach to estimate uncertainties based on a deterministic model is by running a

  Monte Carlo ensemble
  (e.g. by small perturbations of initial conditions)

- This has two problems:
  - It's expensive: it requires many runs
  - It's very expensive: it requires to know what is the probability distribution of non-observed inputs (aka Calibration)

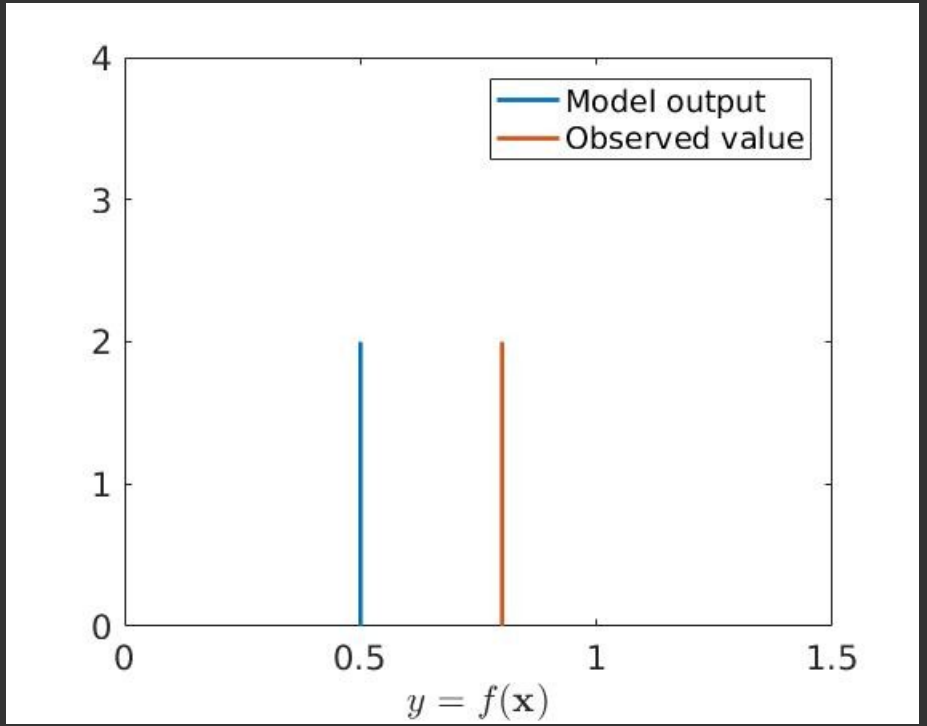


ACCRUE Approach

Let us assume that for a single (multidimensional) input **x**, our model predicts an output y = f(x).

Blue line → Model output
Red line → Real (observed value)

<u>Working hypothesis:</u>
We want to use the model output as the mean of a Gaussian distribution that is interpreted as a probabilistic forecast.



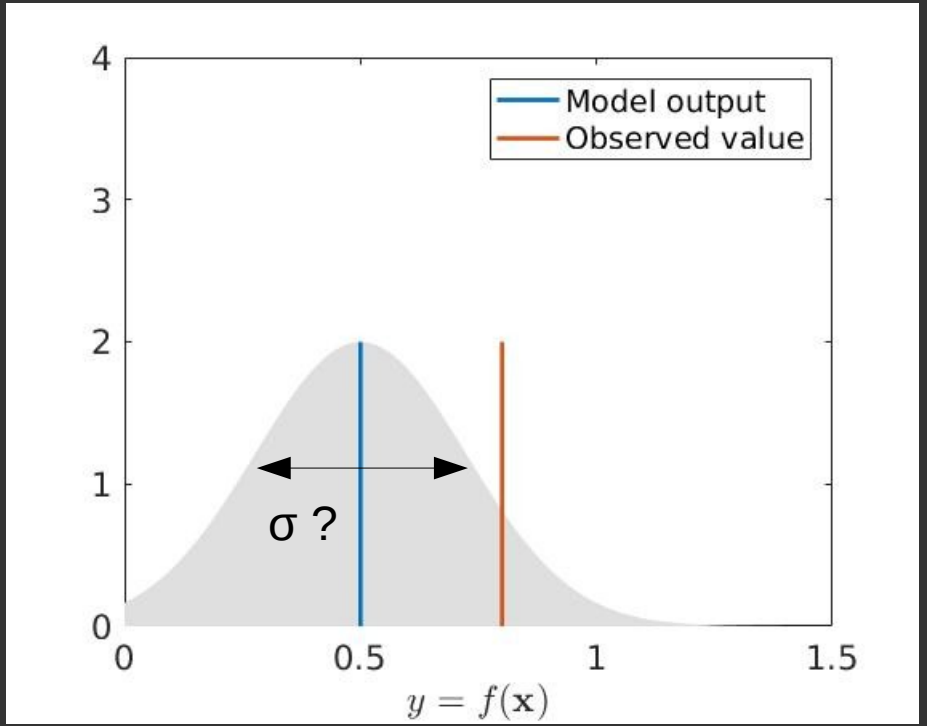
ACCRUE Approach

Let us assume that for a single (multidimensional) input **x**, our model predicts an output y = f(x).

Blue line → Model output
Red line → Real (observed value)

<u>Working hypothesis:</u>
We want to use the model output as the mean of a Gaussian distribution that is interpreted as a probabilistic forecast.

What is the optimal width of a Gaussian forecast?



ACCRUE Recipe

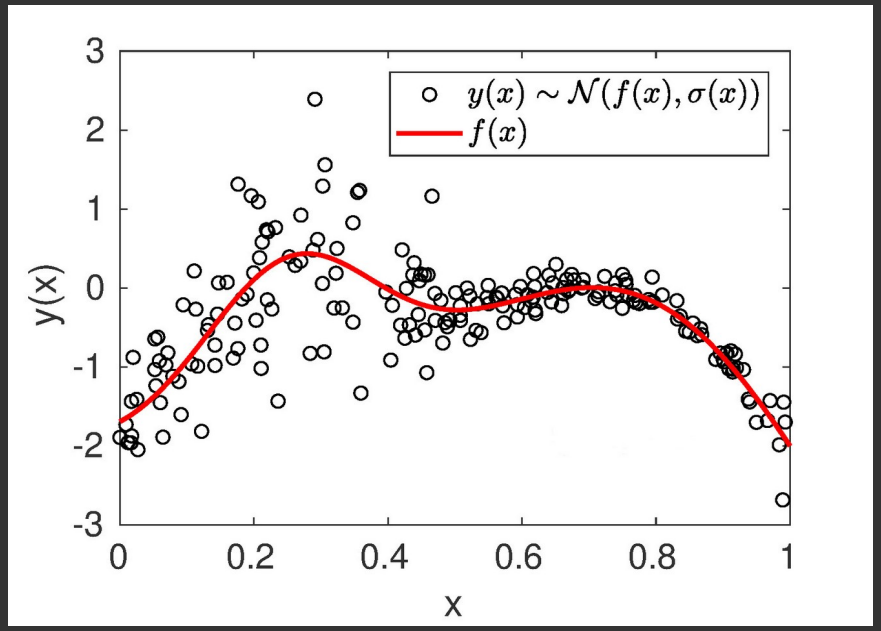
- The optimal Gaussian width (std) is the one that optimizes both accuracy and reliability
- This is a two-objective optimization problem, because reliability and accuracy are competing objectives.
- We define the Accuracy-Reliability (AR) cost function:

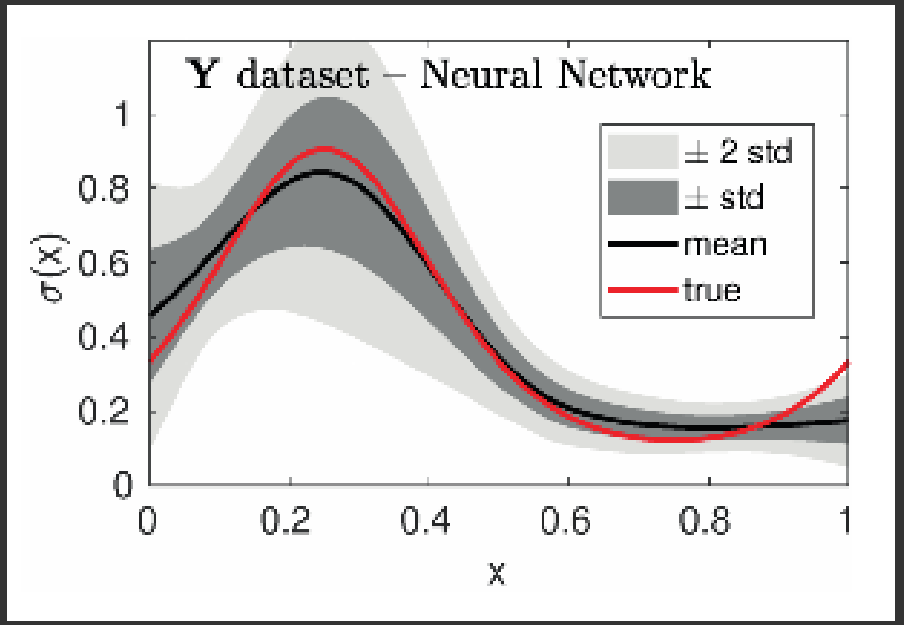
$$AR = CRPS + RS$$

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
  - We have used a polynomial function for std (in 1D) and a neural network (in multiple-D)

Synthetic Example

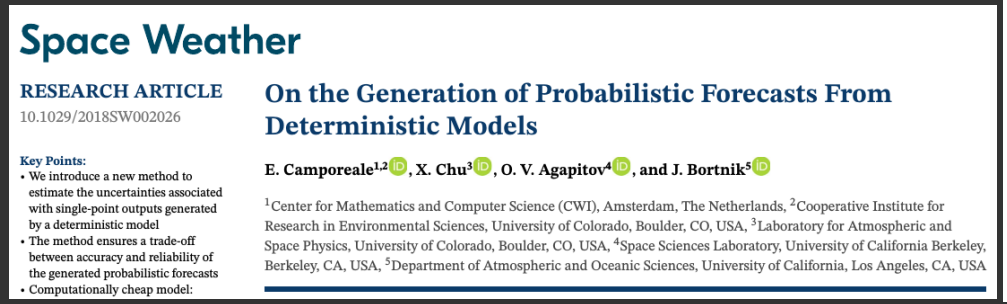




ACCRUE Take Home Message

The ACCRUE method:

- Estimates the uncertainties associated with single-point outputs generated by a deterministic model, in terms of Gaussian distributions;
- Ensures the optimal trade-off between accuracy and reliability;
- Does not need to run ensembles. It costs as much as training a neural network
- Code available: zenodo.1485608



Space Weather

RESEARCH ARTICLE
10.1029/2018SW002026

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:

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

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

ACCRUE: ACCURATE AND RELIABLE UNCERTAINTY ESTIMATE IN DETERMINISTIC MODELS

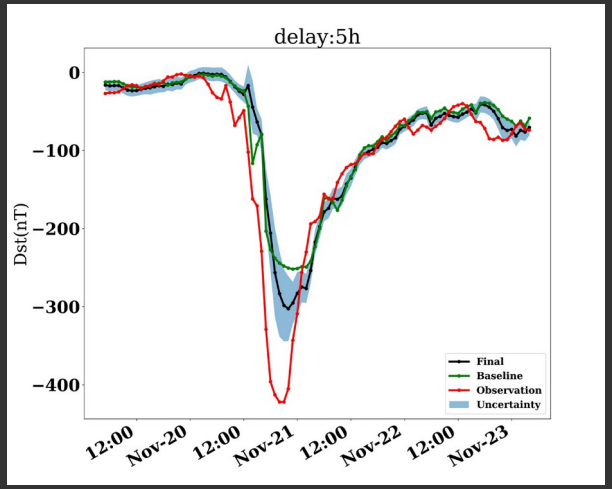
Enrico Camporeale[1,*] *& Algo Carè*[2]

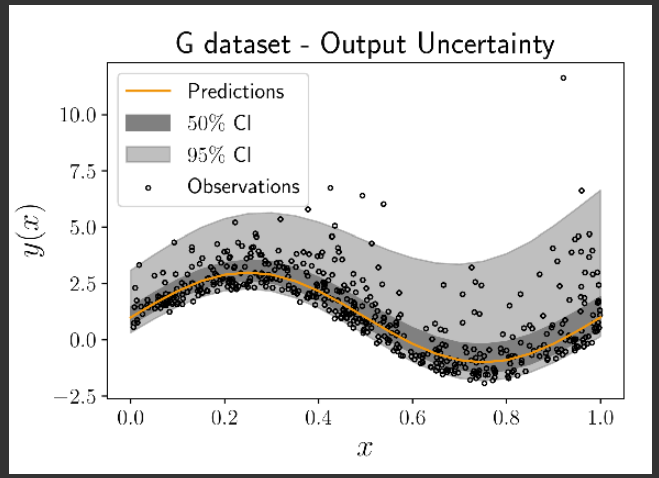
ACCRUE: next steps

- Generalize the method to non-Gaussian distributions (errors tend to be skewed)

Skewed Uncertainty Estimates for Deterministic Predictions

Rileigh Bandy [1] Andong Hu [1,2] Rebecca Morrison [1] Enrico Camporeale [1,2]

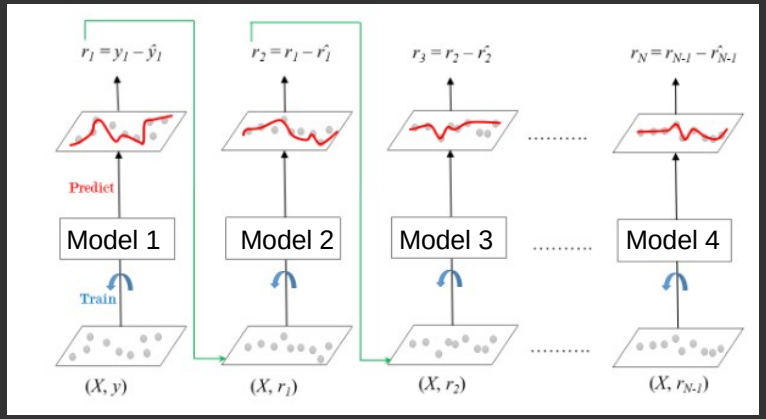


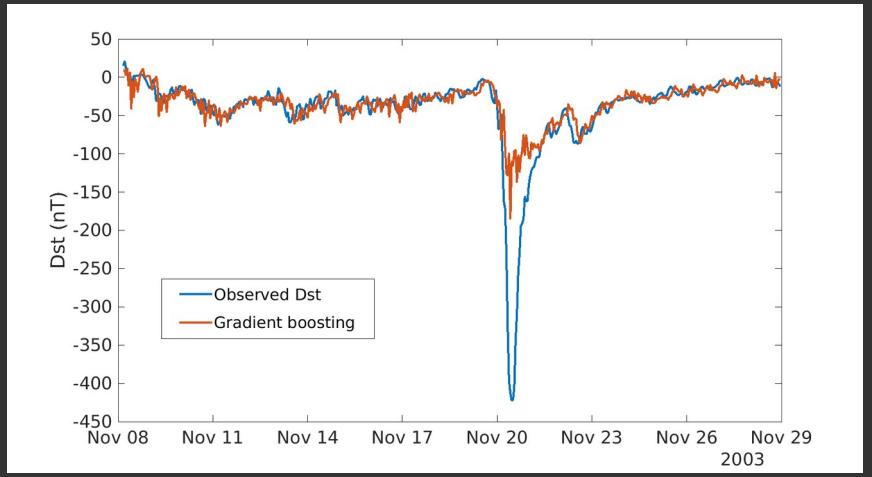


Poster by
Rileigh Bandy

Gradient Boosting in one slide

- A hierarchy of models is built
- Each one is trained on the residuals (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

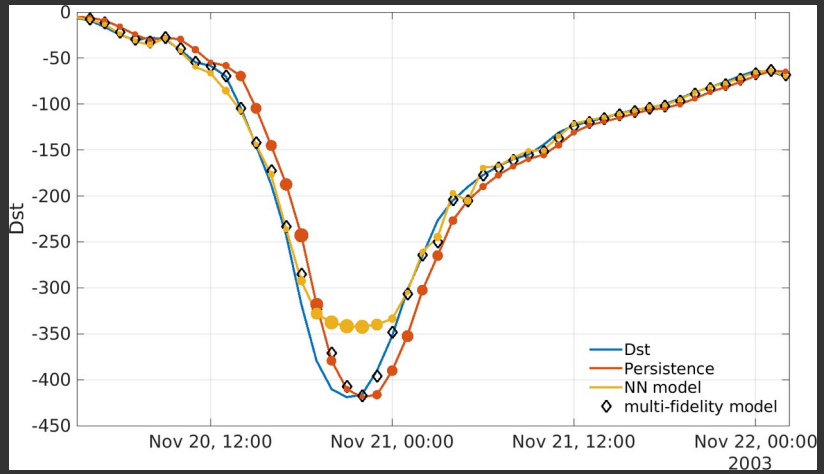
- We can use the power of ACCRUE to identify regions (in feature space) where a predictor works well and where it doesn’t

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 = $1/\sigma^2$

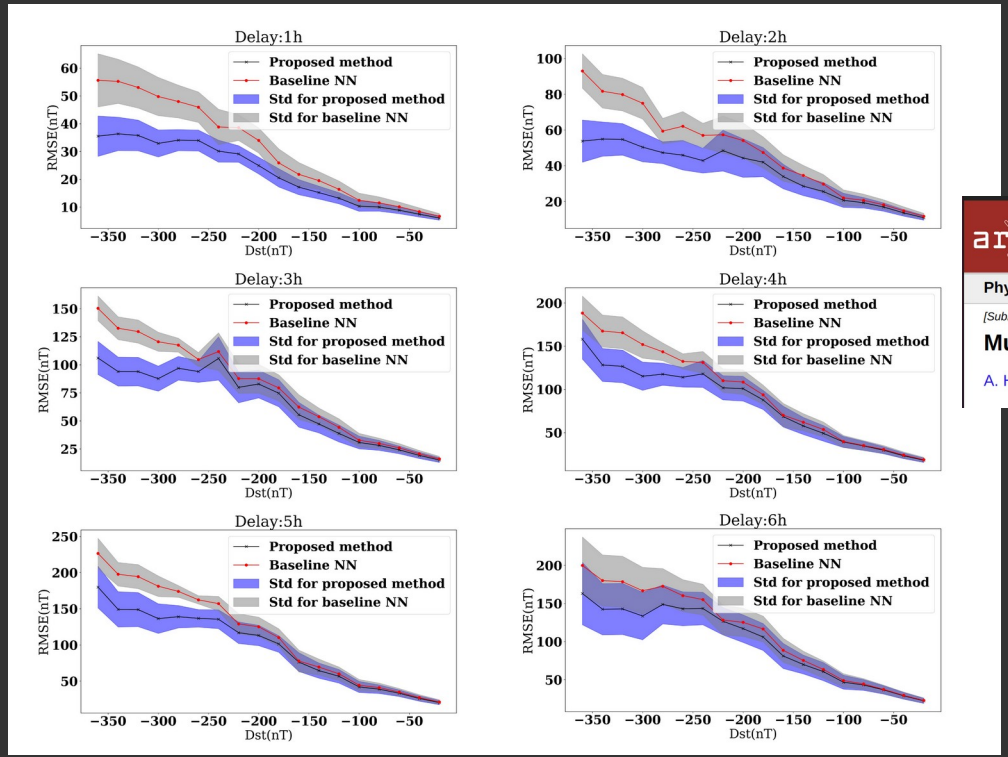


Areas of application

- Magnetosphere-solar wind coupling
  - Geomagnetic indices
  - Ground magnetic and electric field
- Solar wind forecasting
- Earth’s radiation belts

Dst forecasting

- Dst is a geomagnetic index that is a proxy for the ring current. It is measured by averaging the magnetic perturbations observed on four equatorial stations



to appear in *Space Weather*

arXiv > physics > arXiv:2209.12571

Physics > Space Physics

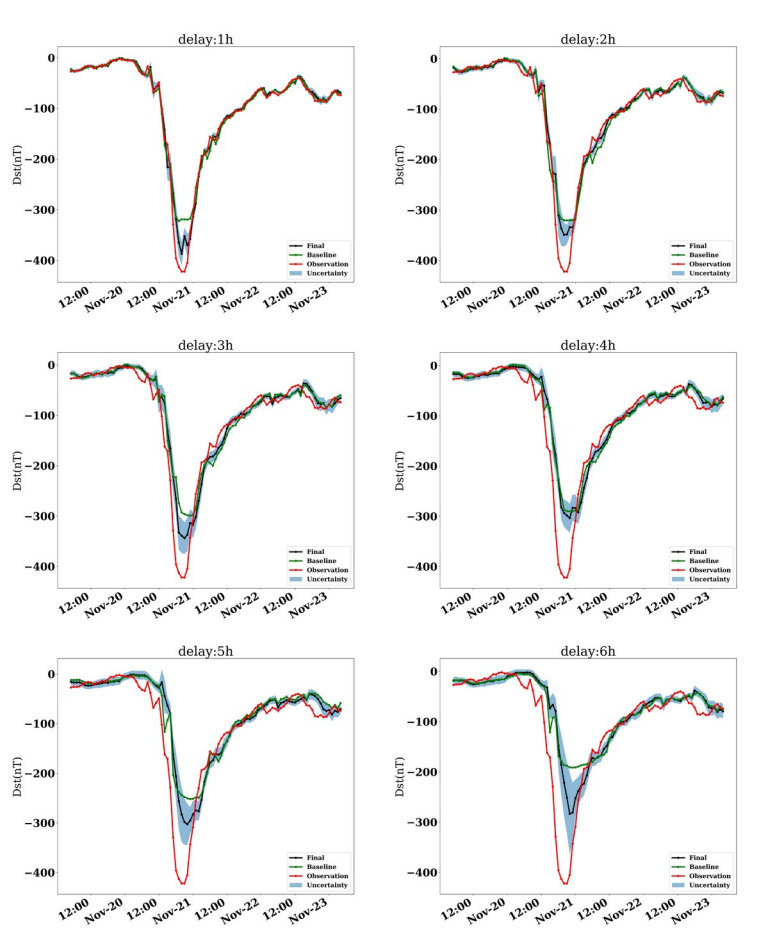
[Submitted on 26 Sep 2022]

Multi-Hour Ahead Dst Index Prediction Using Multi-Fidelity Boosted Neural Networks

A. Hu, E. Camporeale, B. Swiger

Dst forecasting

- Dst [illegible] proxy for the ring current. It is measured by a [illegible] ons observed on four equatorial stat[illegible]



to appear in *Space Weather*

arXiv > physics > arXiv:2209.12571

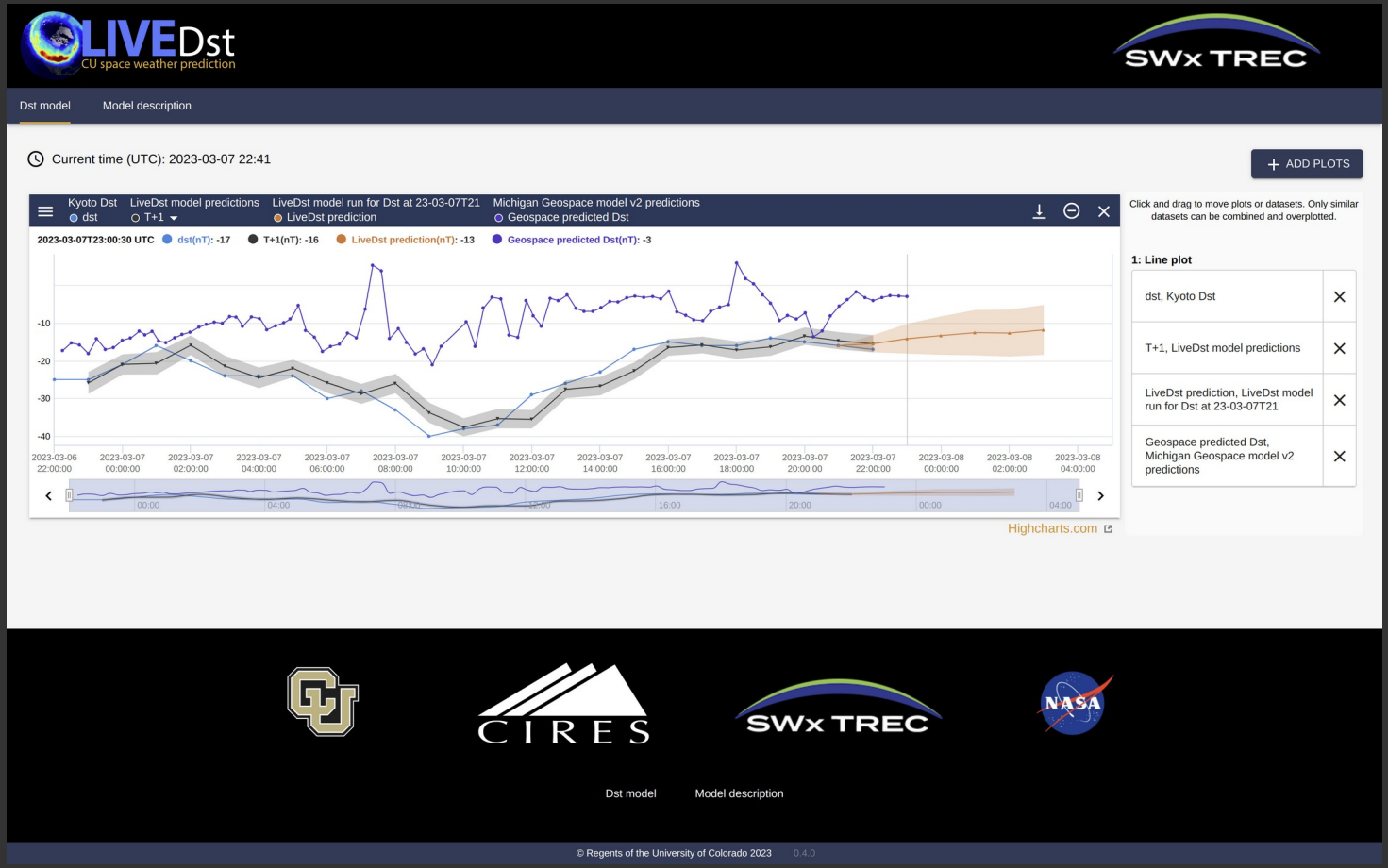
Physics > Space Physics

[Submitted on 26 Sep 2022]

Multi-Hour Ahead Dst Index Prediction Using Multi-Fidelity Boosted Neural Networks

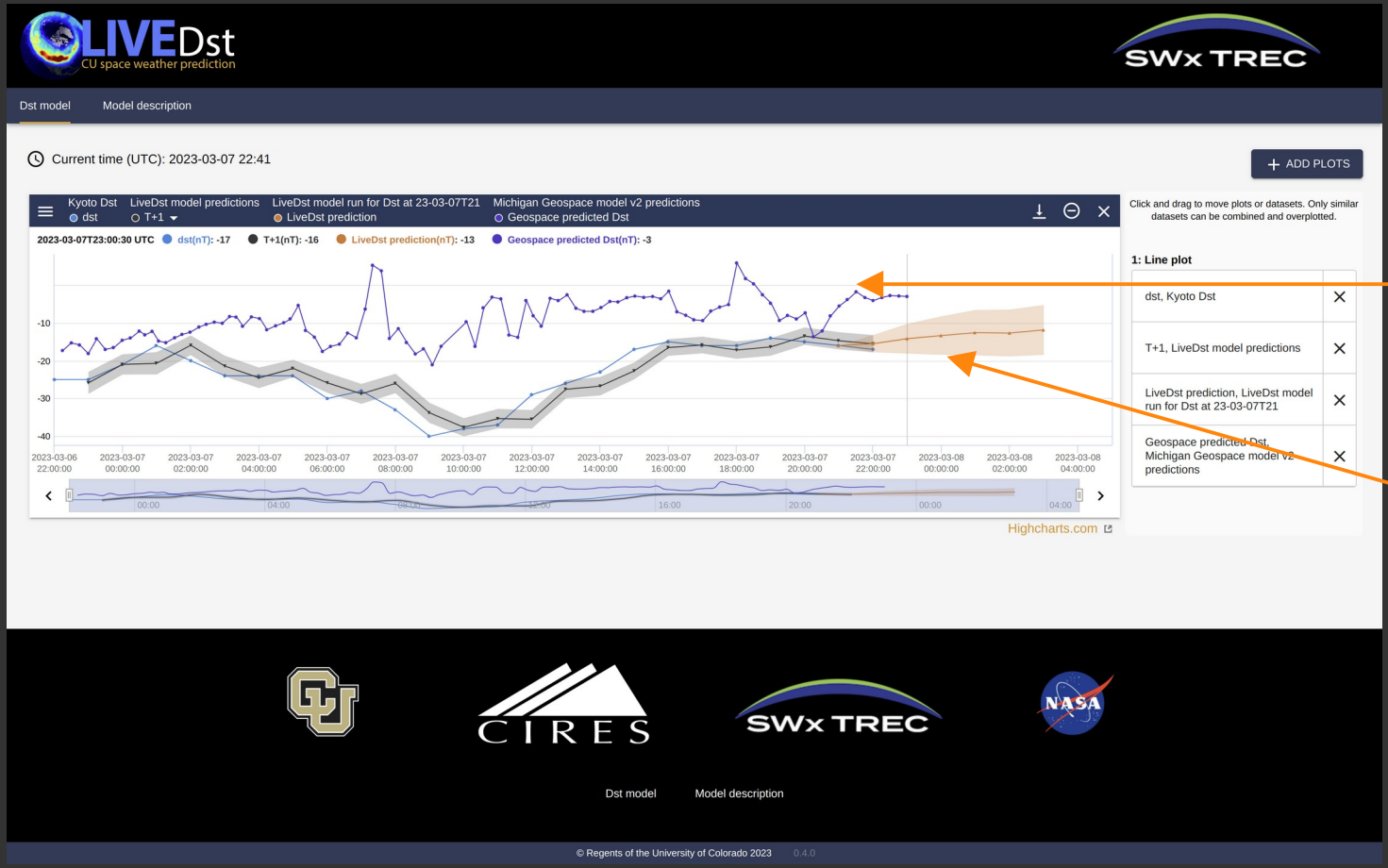
A. Hu, E. Camporeale, B. Swiger

LiveDst in real-time



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

LiveDst in real-time

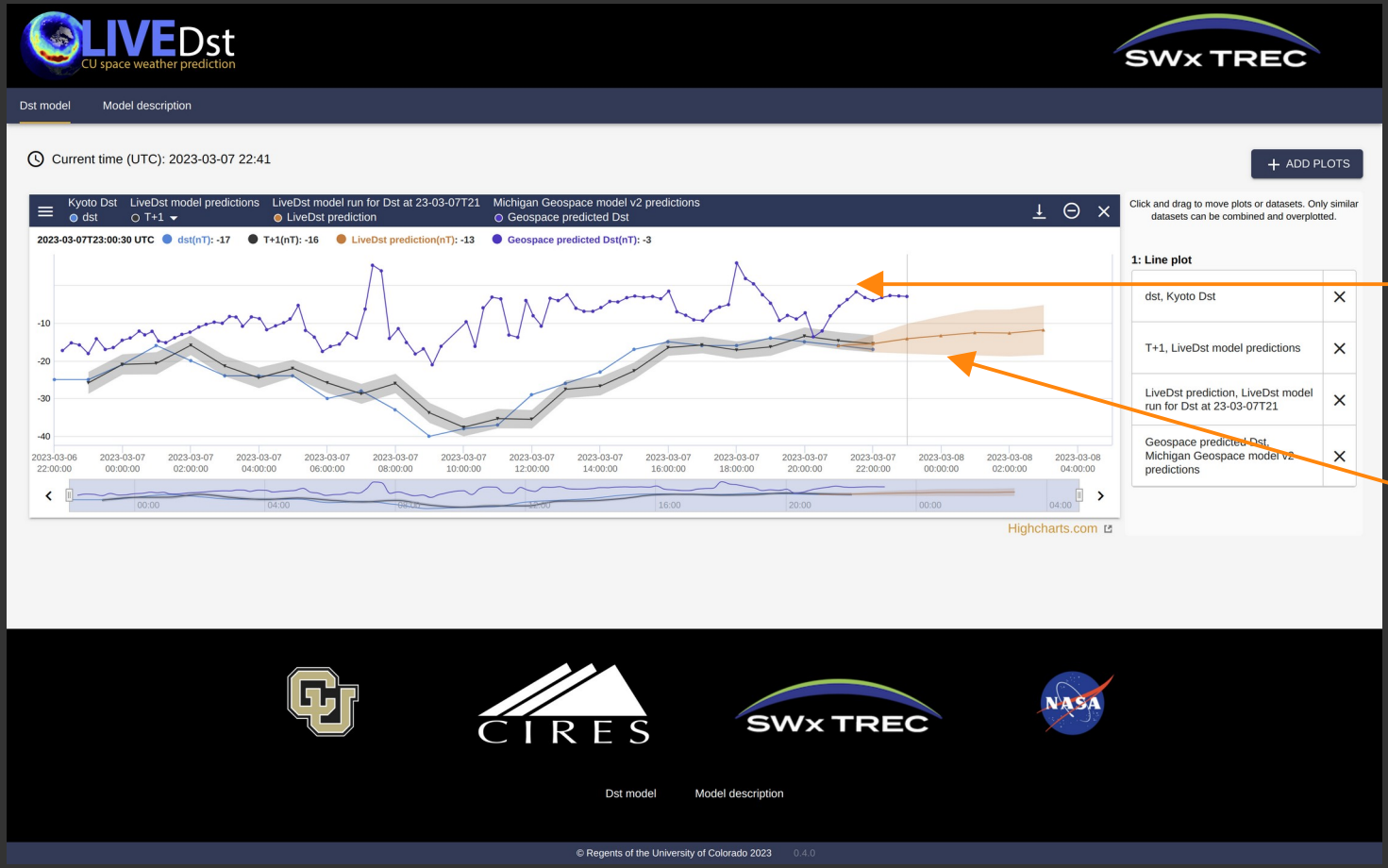


How much does it cost to run?

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

Thanks to:
Greg Lucas
Jenny Knuth
Brandon Stone
(CU-LASP)

LiveDst in real-time



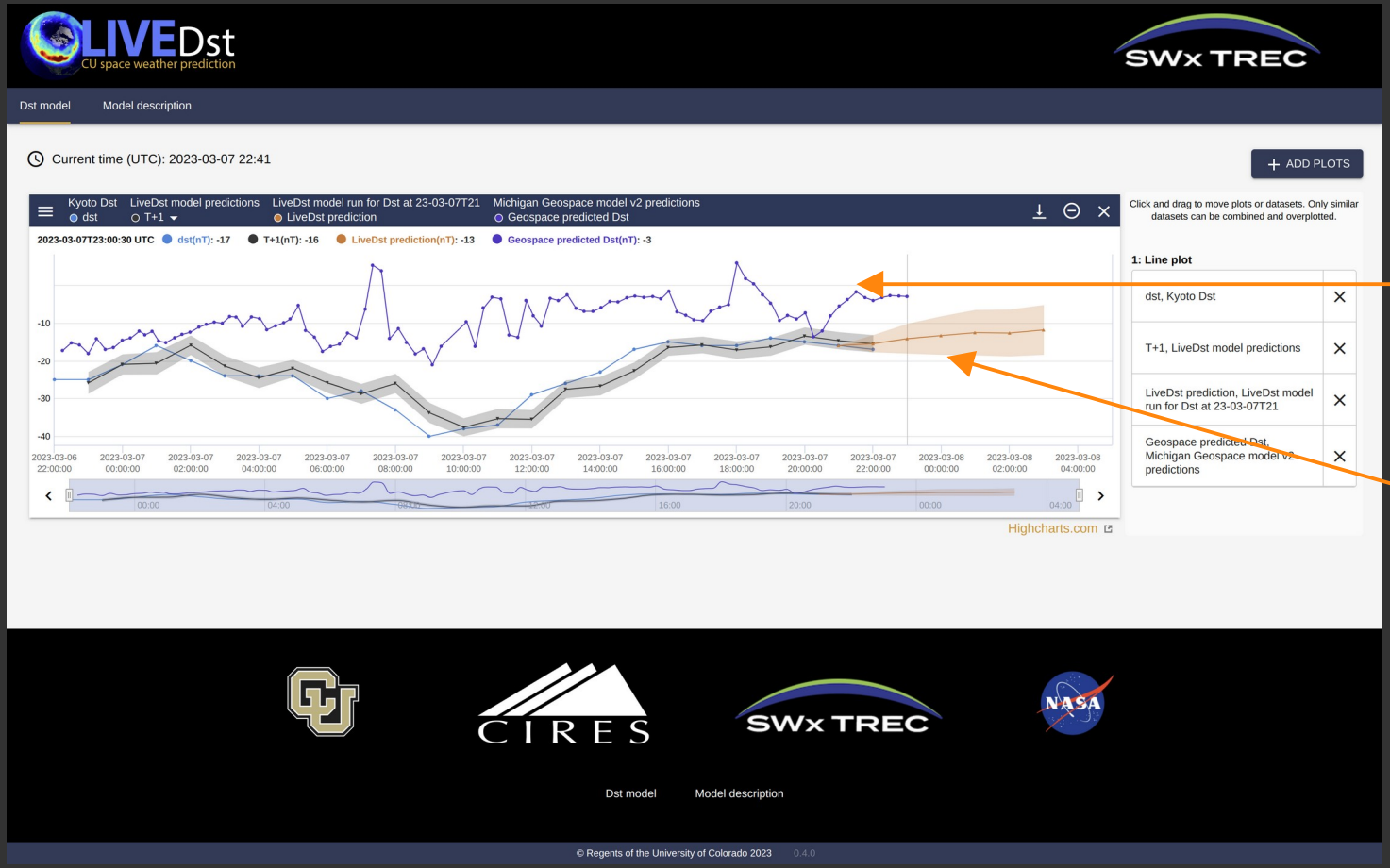
How much does it cost to run?

The operational NOAA Geospace model costs >$15,000/year

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

Thanks to:
Greg Lucas
Jenny Knuth
Brandon Stone
(CU-LASP)

LiveDst in real-time



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

How much does it cost to run?

The operational NOAA Geospace model costs >$15,000/year

LiveDst runs on AWS and costs <$10/year

Thanks to:
Greg Lucas
Jenny Knuth
Brandon Stone
(CU-LASP)

1 day ahead Dst prediction

Space Weather®

RESEARCH ARTICLE
10.1029/2022SW003064

Probabilistic Prediction of *Dst* Storms One-Day-Ahead Using Full-Disk SoHO Images

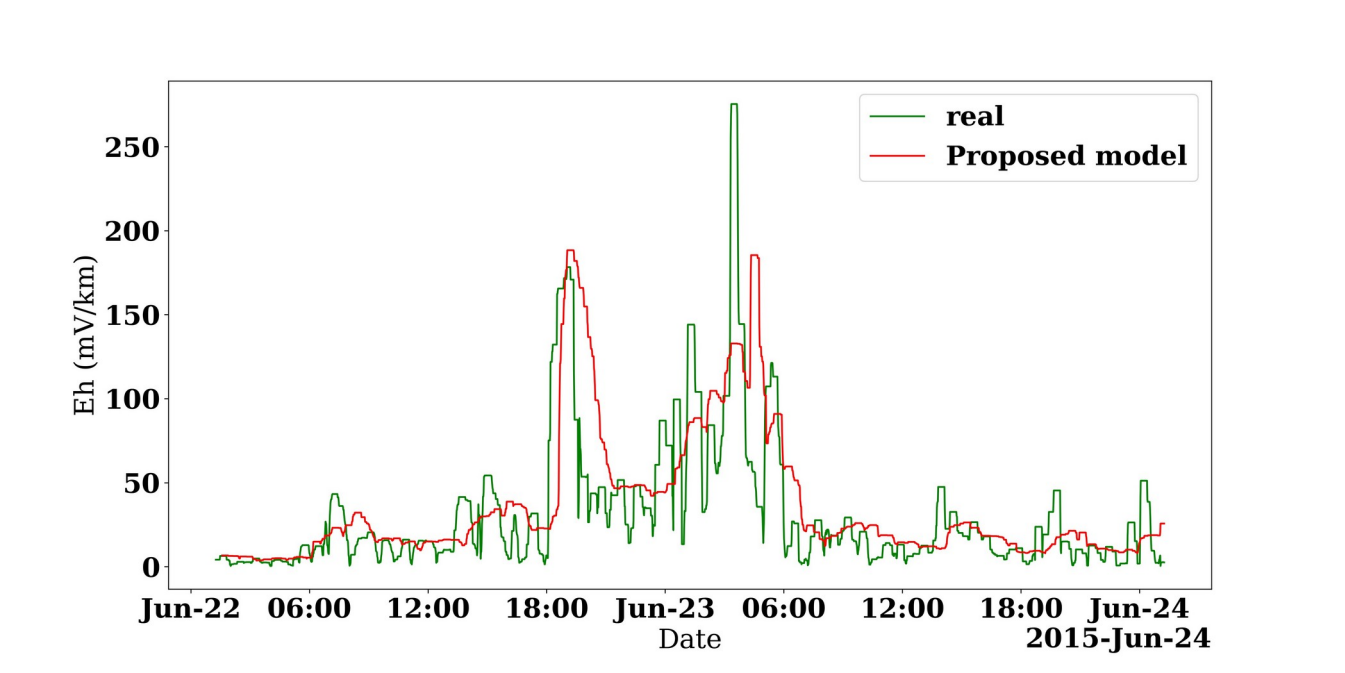
A. Hu[1,2], C. Shneider[1], A. Tiwari[1], and E. Camporeale[2,3]

Key Points:

- A new disturbance storm time (*Dst*)

(video)

Regional GeoElectric field forecast



1-hr ahead forecast of the horizontal Electric field (Eh)

Work in progress by Andong Hu

Radiation belt

Space Weather®

RESEARCH ARTICLE
10.1029/2022SW003051

Key Points:

- We perform ensemble simulations of radiation belt electron flux decay with a dynamically changing environment at a high time resolution
- Uncertainties in wave amplitude dominantly cause simulation errors, compared to density, wave peak

Ensemble Modeling of Radiation Belt Electron Flux Decay Following a Geomagnetic Storm: Dependence on Key Input Parameters

Man Hua[1], Jacob Bortnik[1], Adam C. Kellerman[2], Enrico Camporeale[3], and Qianli Ma[1,4]

[1]Department of Atmospheric and Oceanic Sciences, UCLA, Los Angeles, CA, USA, [2]Department of Earth, Planetary, and Space Sciences, UCLA, Los Angeles, CA, USA, [3]CIRES, University of Colorado, Boulder, CO, USA, [4]Center for Space Physics, Boston University, Boston, MA, USA

Space Weather®

Research Article | Free Access

Ensemble Modeling of Radiation Belt Electron Acceleration by Chorus Waves: Dependence on Key Input Parameters

Man Hua, Jacob Bortnik, Adam C. Kellerman, Enrico Camporeale, Qianli Ma

First published: 17 January 2023 | https://doi.org/10.1029/2022SW003234

Sensitivity analysis of drivers of radiation belt's electron dynamics, by running ensemble simulations

See poster by Man Hua!

Radiation belt

JGR Space Physics

RESEARCH ARTICLE
10.1029/2022JA030377

Special Section:
Machine Learning in Heliophysics

Data-Driven Discovery of Fokker-Planck Equation for the Earth's Radiation Belts Electrons Using Physics-Informed Neural Networks

E. Camporeale[1,2], George J. Wilkie[3], Alexander Y. Drozdov[4], and Jacob Bortnik[4]

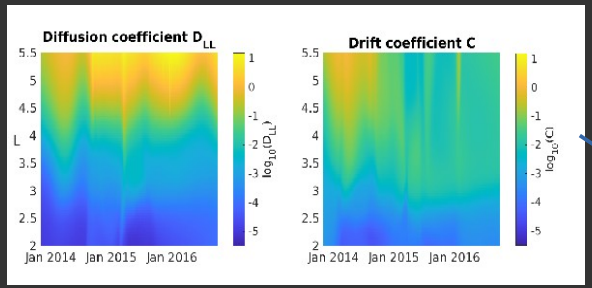
Parameter estimation with Physics-Informed Neural Network

Assumption:
The physics obeys FP equation

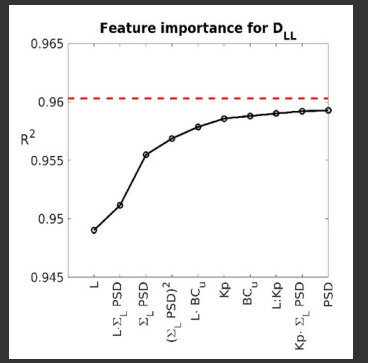
$$\frac{\partial f(L,t)}{\partial t} = L^2 \frac{\partial}{\partial L}\left(\frac{D_{LL}}{L^2}\frac{\partial f(L,t)}{\partial L}\right) - \frac{\partial C f(L,t)}{\partial L},$$

PINN

Optimal coefficients



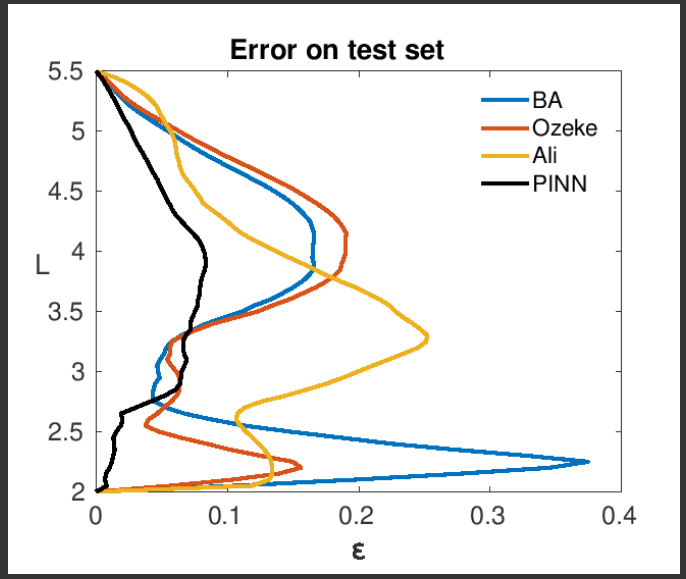
Feature selection

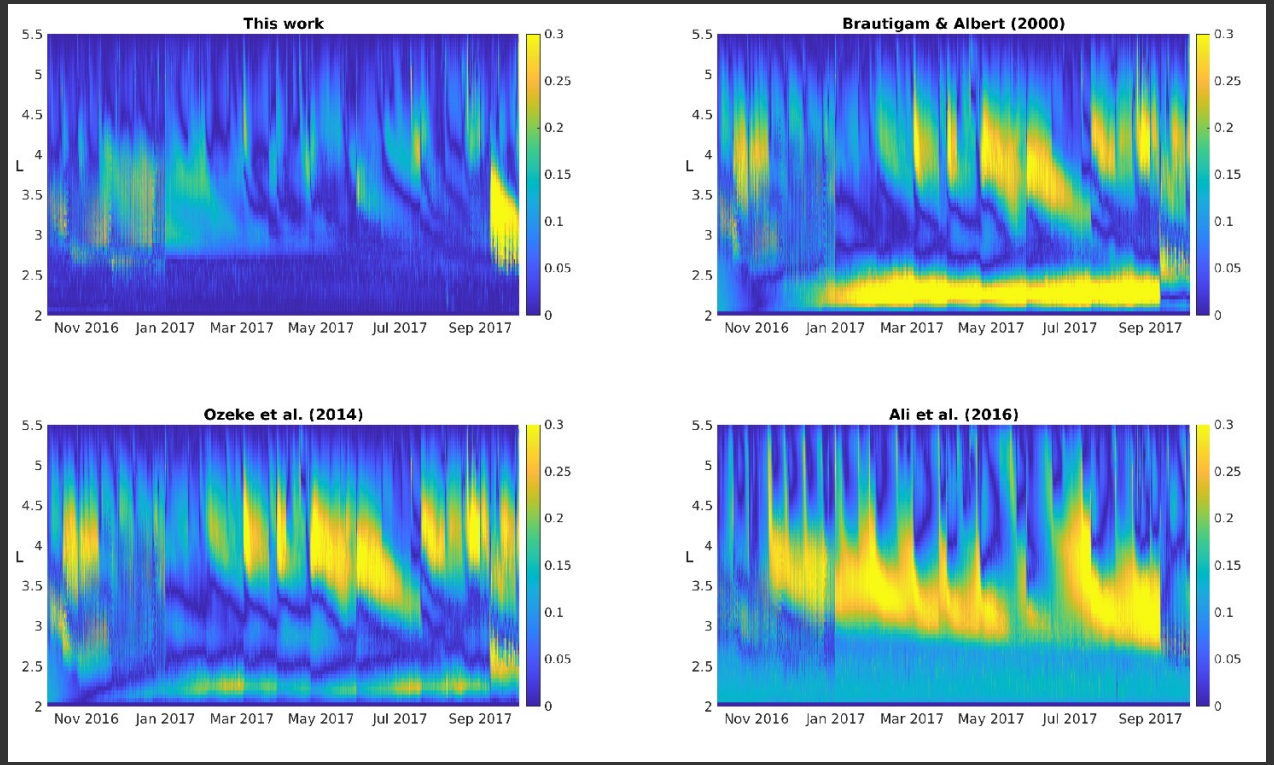


Train a ML model that predicts D_{LL} and C at a given time

Solve the equation with PINN-discovered and ML-learned coefficients!

Results on test set





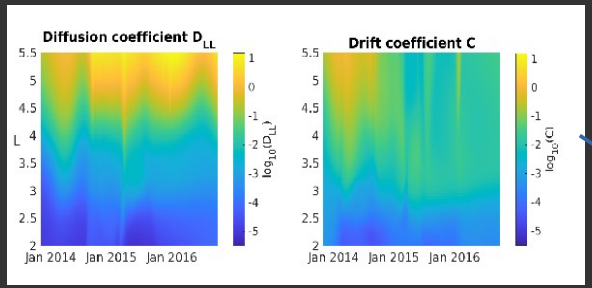
The final frontier: Interpretable AI

Assumption:
The physics obeys FP equation

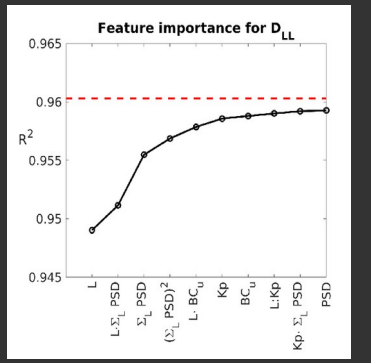
$$\frac{\partial f(L,t)}{\partial t} = L^2 \frac{\partial}{\partial L}\left(\frac{D_{LL}}{L^2}\frac{\partial f(L,t)}{\partial L}\right) - \frac{\partial C f(L,t)}{\partial L},$$

PINN

Optimal coefficients

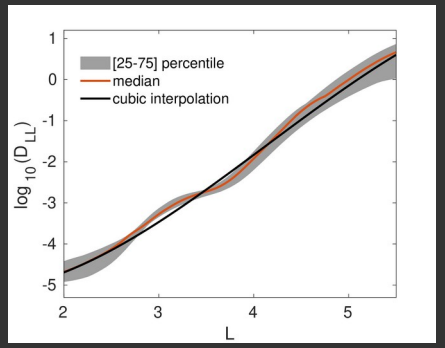


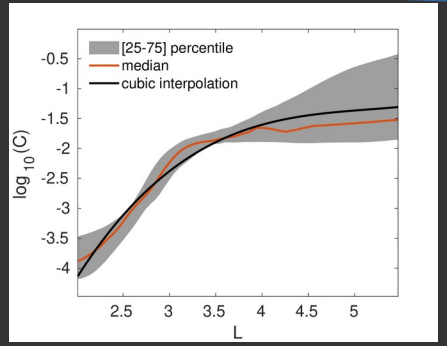
Feature selection



Train a ML model that predicts D_{LL} and C at a given time

Solve the equation with PINN-discovered and ML-learned coefficients!



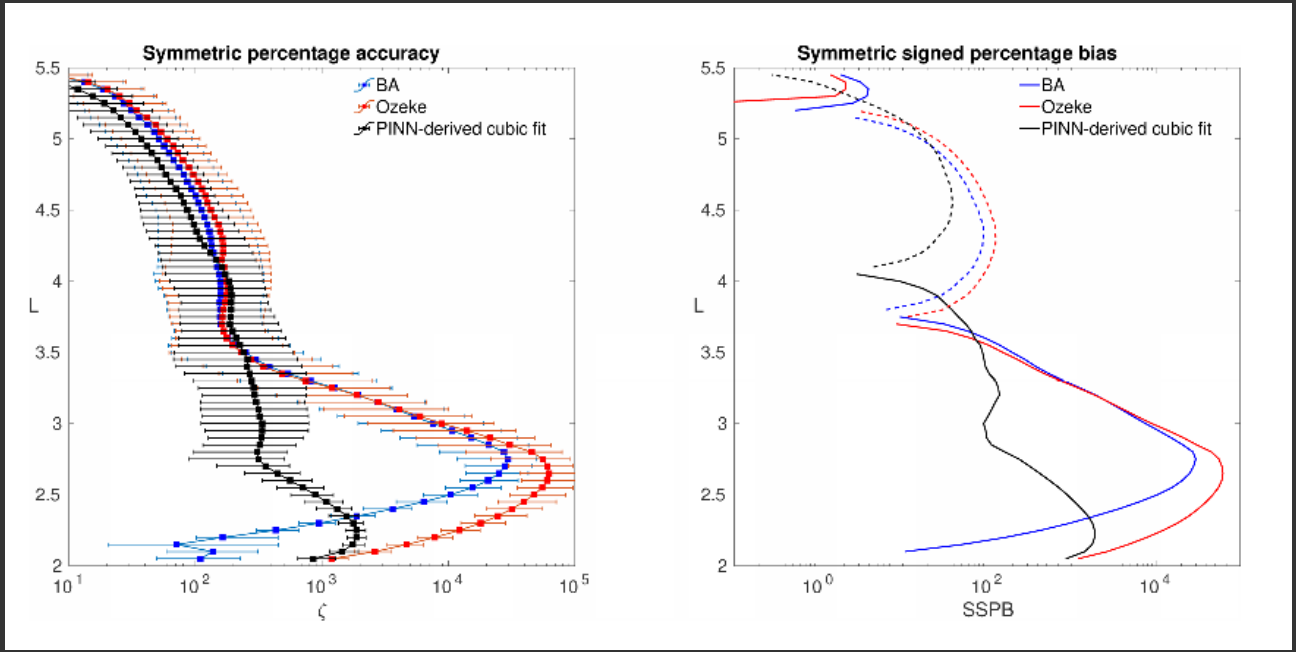


Instead: approximate D_{LL} and C with cubic interpolation

$$\log_{10} D_{LL} = -0.0593L^3 + 0.7368L^2 - 1.33L - 4.505$$

$$\log_{10} C = 0.0777L^3 - 1.2022L^2 + 6.3177L - 12.6115$$

The final frontier: Interpretable AI



$$\frac{\partial f(L,t)}{\partial t} = L^2 \frac{\partial}{\partial L}\left(\frac{D_{LL}}{L^2}\frac{\partial f(L,t)}{\partial L}\right) - \frac{\partial C f(L,t)}{\partial L},$$

D_{LL} and C are functions of L only !!
The FP equation has no free parameters:
Completely determined by BC !!

SynWind: synthetic solar wind

Why do we need synthetic solar wind?

1) To **augment** training data for machine learning algorithms (train on synthetic, test on real)

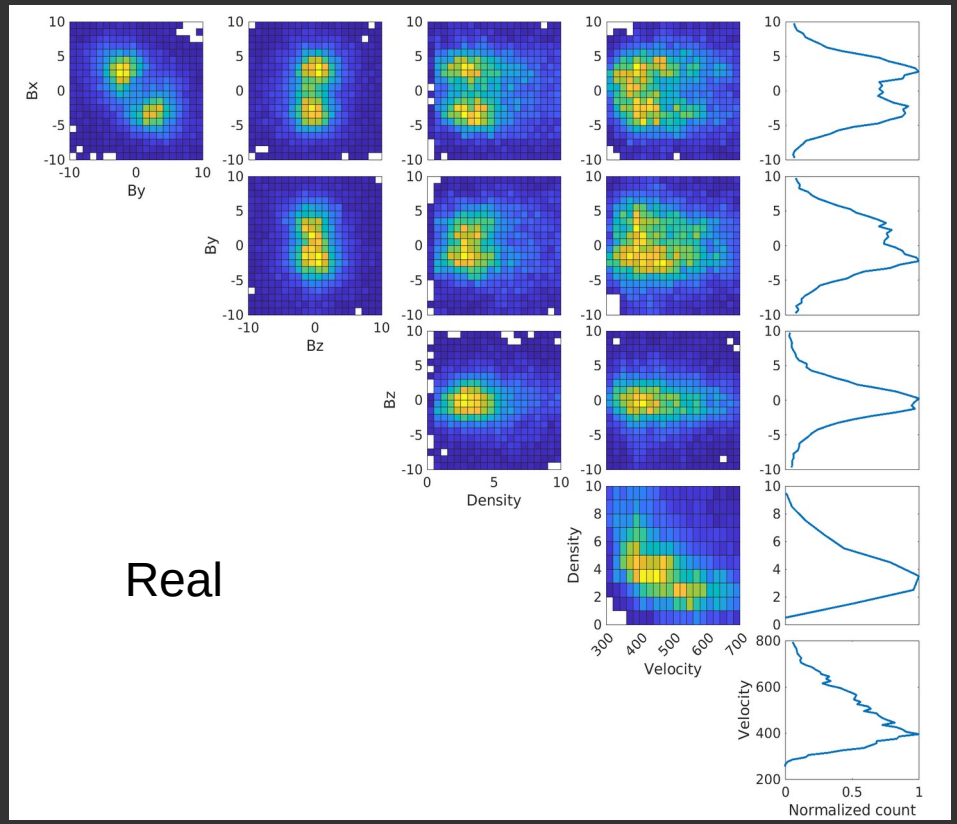
2) To generate SW **ensembles** that are statistically consistent with the real SW to be used in driving physics-based ensemble simulations

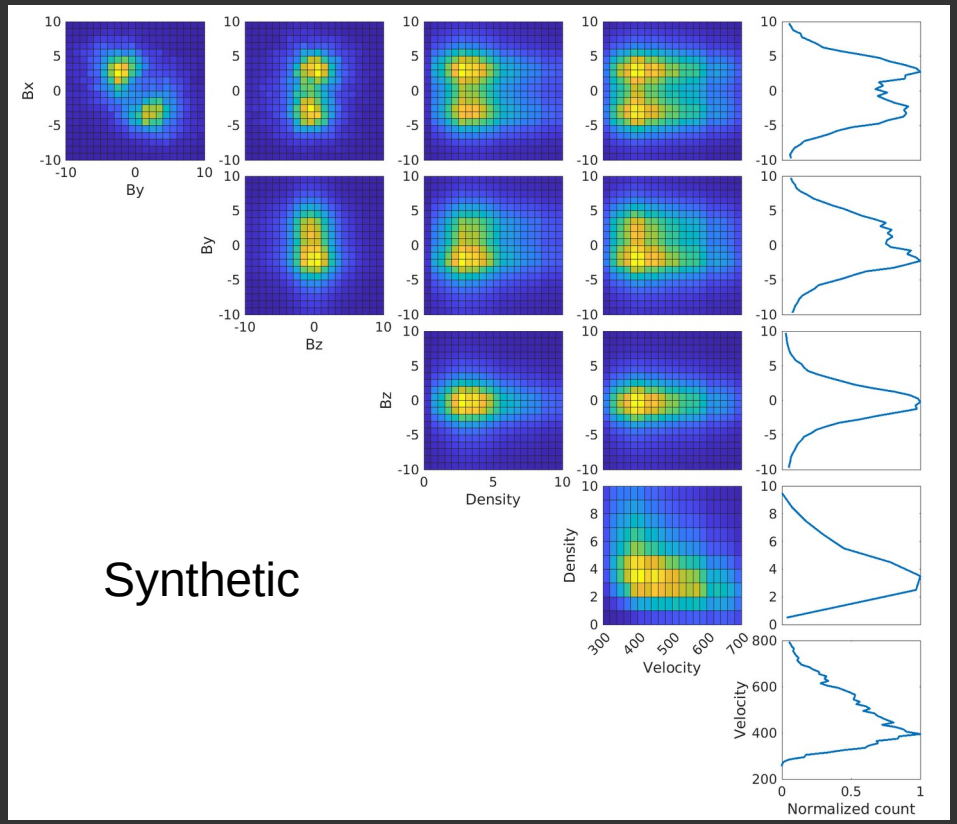
3) To run global **sensitivity** analysis of models (i.e., change one, two, …, five, parameters at a time)

4) To run methods for **explaining** trained ML models

SynWind: synthetic solar wind

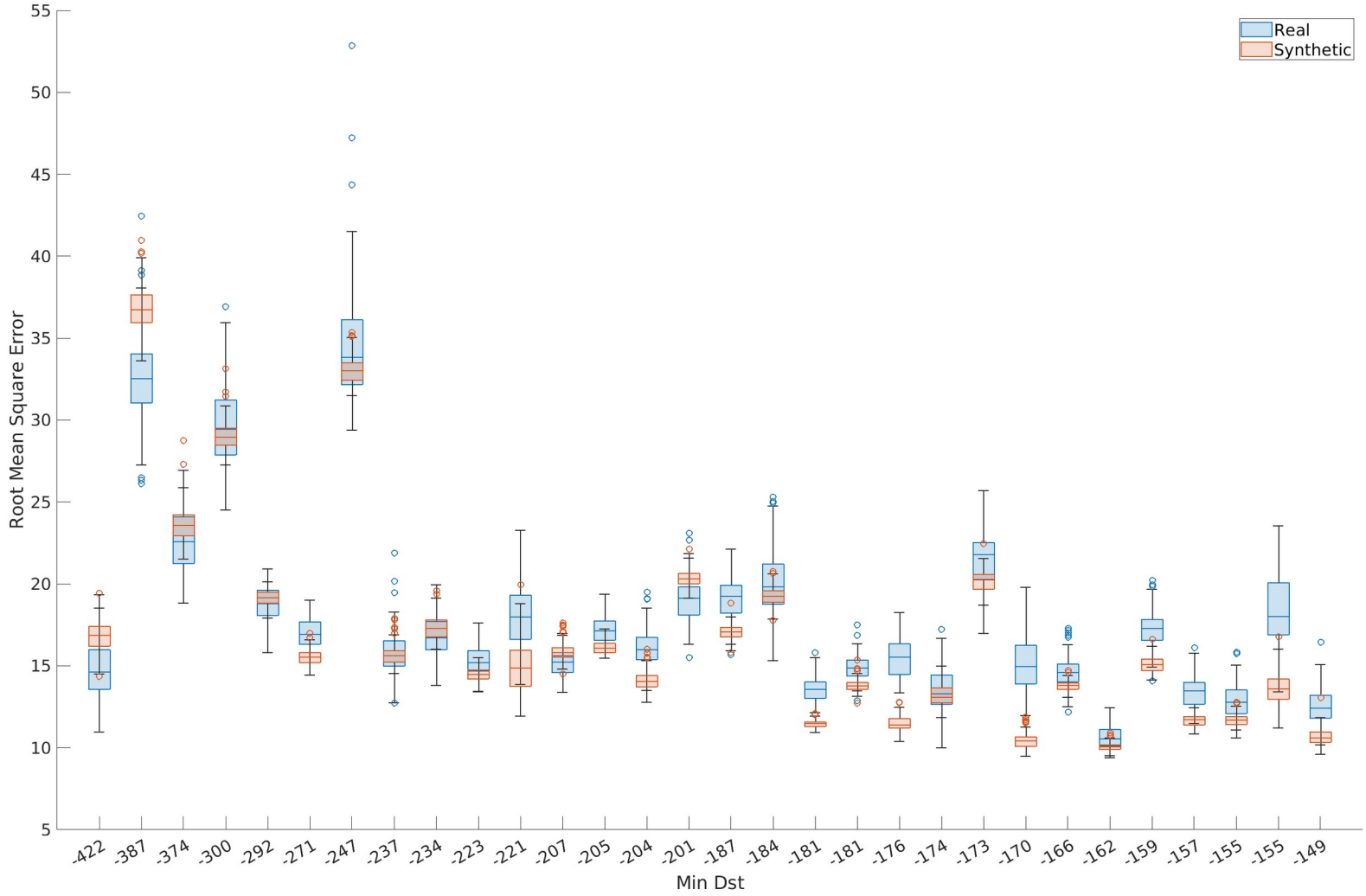
Trained on the most 50 intense storms, SynWind generates IMF Bx, By, Bz, Speed and Density

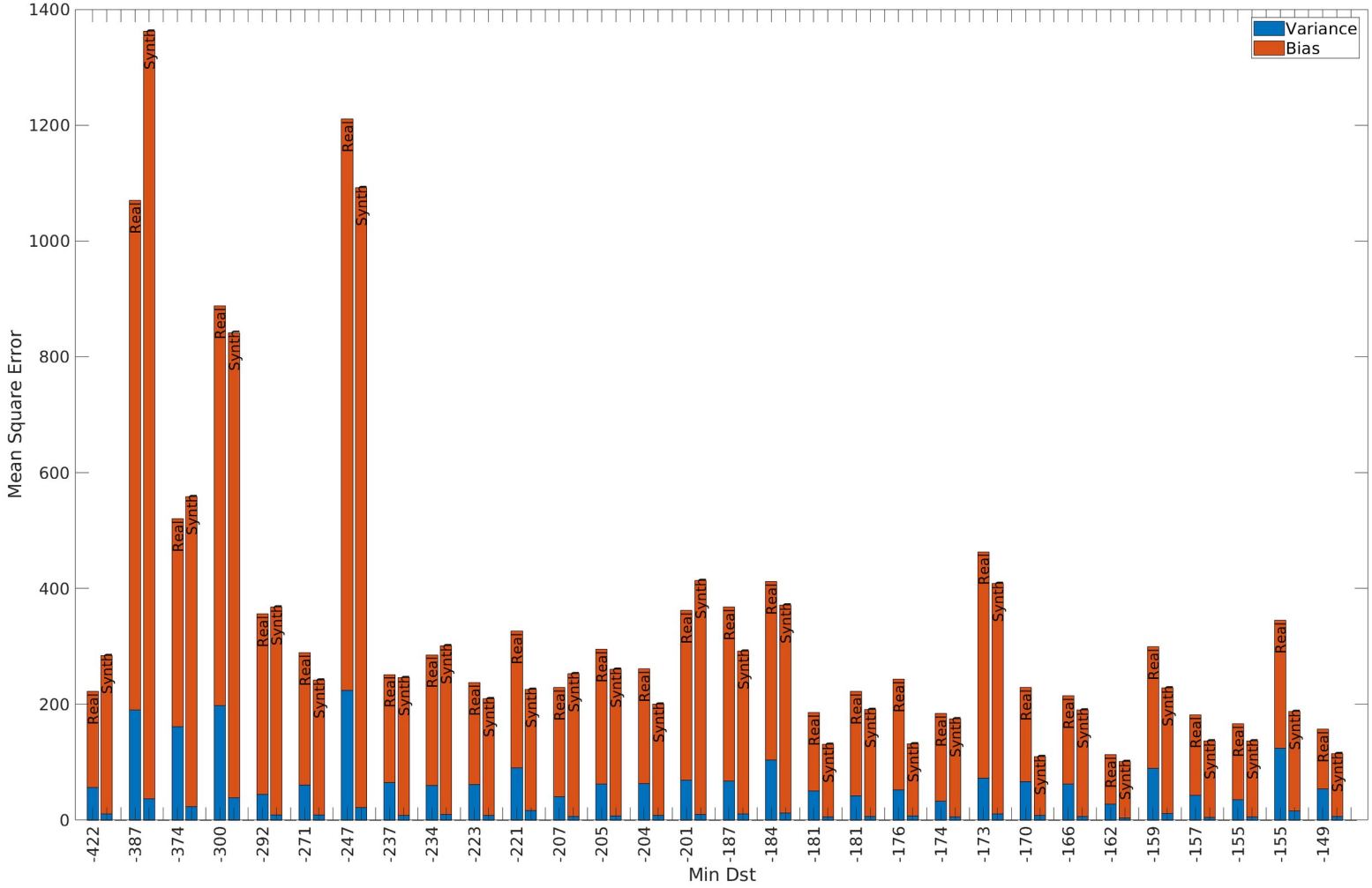




Train on synthetic, test on real!

- Benchmark: 3-hours ahead Dst prediction
- Generate 1M data points of SynWind (~114 years)
- Select 50 most intense storms (from real data)
- Leave one real storm out
- Train a nowcast model (no delay) for Dst and use that model to generate Dst associated to the 1M SynWind points
- Use that as a training set for 3-hrs ahead prediction
- Compare with a model trained on real data


Real
Synthetic
Root Mean Square Error
Min Dst
55
50
45
40
35
30
25
20
15
10
5
-422
-387
-374
-300
-292
-271
-247
-237
-234
-223
-221
-207
-205
-204
-201
-187
-184
-181
-181
-176
-174
-173
-170
-166
-162
-159
-157
-155
-155
-149


Variance
Bias
Mean Square Error
Min Dst
0
200
400
600
800
1000
1200
1400
-422
-387
-374
-300
-292
-271
-247
-237
-234
-223
-221
-207
-205
-204
-201
-187
-184
-181
-181
-176
-174
-173
-170
-166
-162
-159
-157
-155
-155
-149
Real
Synth

Conclusions

- Broader impact
  - ACCRUE
  - ProBoost

UQ and ML methods that are not confined to Space Weather (trans-disciplinary applications)

- Intellectual merit
  - State of the art forecasting models for geomagnetic indices, geoElectric field, radiation belts (with built-in UQ)
  - Pioneered ML-based equation discovery method
- Future work:
  - Improvement of ACCRUE and ProBoost
  - Solar wind prediction with physics-informed deep learning
  - Real-time prediction website to include Rad Belt, E-field predictions (investigating potential commercial applications...)
  - Leverage of Deep Learning Lab (CU SWx-TREC)