Forecasting geomagnetic storm disturbances and their uncertainties using deep learning

Research project PID2020-113135RB-C33 supported by MCIN/AEI/10.13039/501100011033

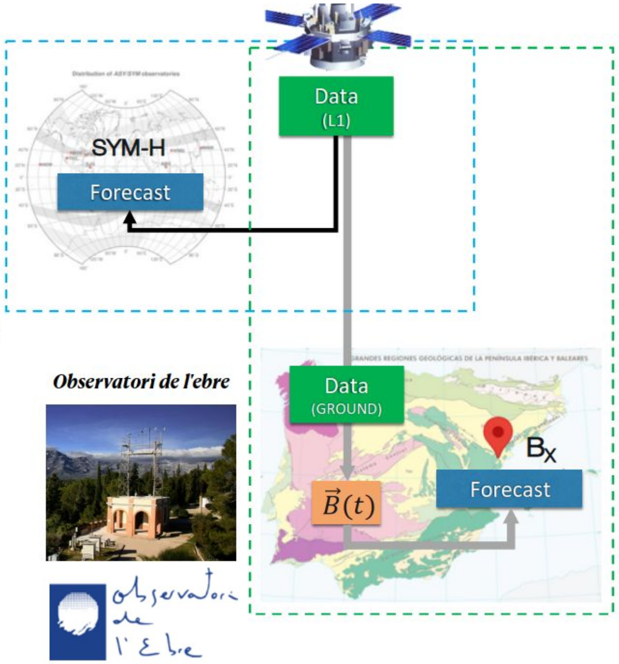
Florencia Castillo [2], Daniel Conde [1], Carlos Escobar [1], Carmen García [1], Jose Enrique García [1], Veronica Sanz [1], Bryan Zaldivar [1],

[1] Instituto de Física Corpuscular (IFIC), [2] Laboratory d'Annecy de Physics des Particules (LAPP)

@HEPHY 2024
Vienna, Austria
12/09/2024

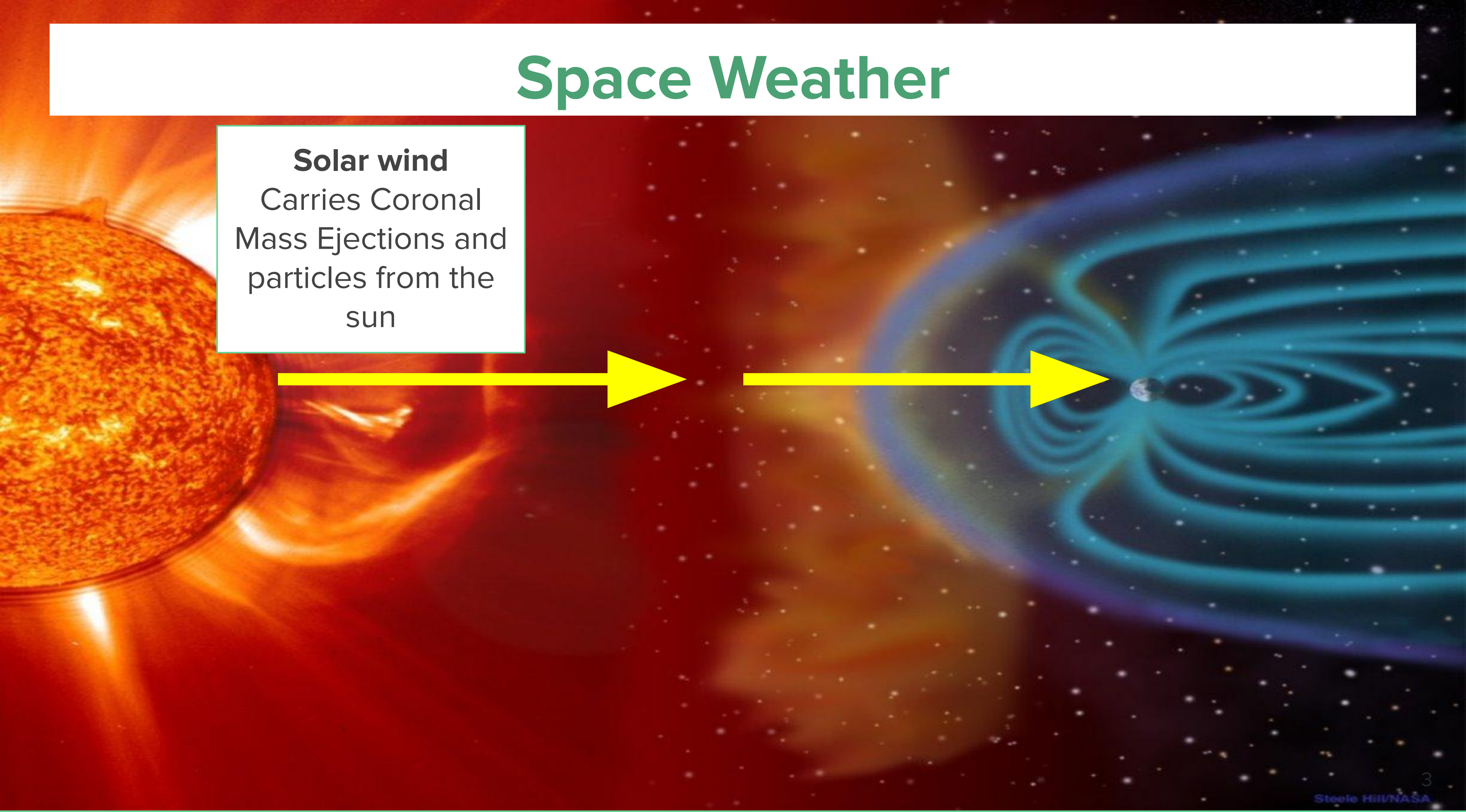
Overview

- **Introduction**
  - Space Weather
  - Objectives
  - Data used
- **Neural network architectures:**
  - LSTM as a recursive neural network
  - Hyperparameter optimization with Optuna
  - Feature importance
- **Forecasting of SYM-H**
  - Target variable predictions
  - Comparison with Siciliano et al using ACE data
  - Uncertainty analysis
- **Addition of solar wind feature variables**
  - Compare RMSE, PICP and feature importance
- **Forecasting of ground level magnetic field**
  - Data description
  - Target variable predictions and RMSE for test storms
- **Conclusions**



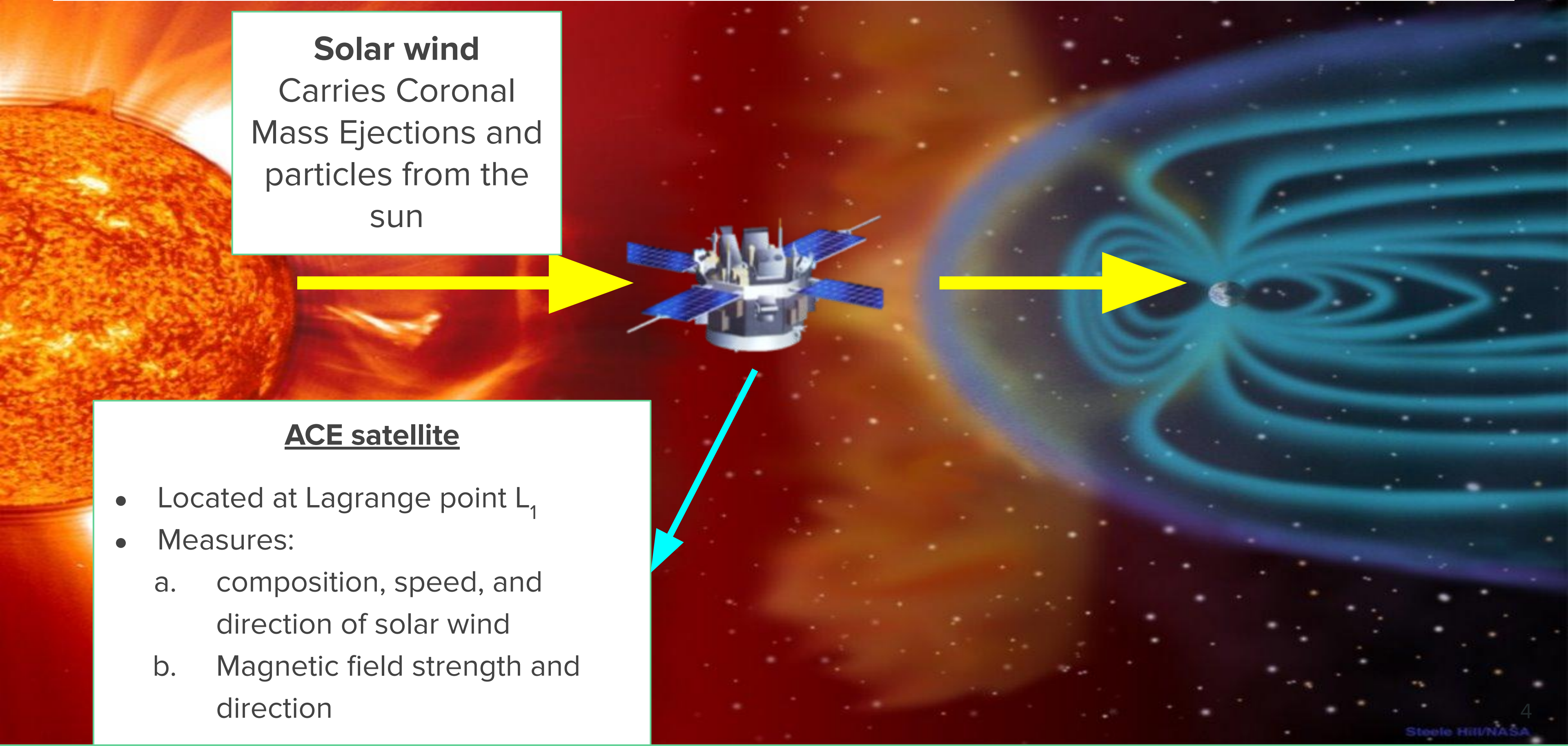
Space Weather

Solar wind
Carries Coronal Mass Ejections and particles from the sun



Space Weather

Solar wind
Carries Coronal Mass Ejections and particles from the sun

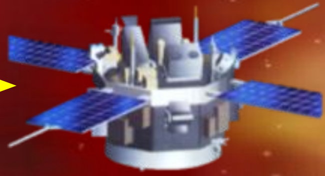


ACE satellite

- Located at Lagrange point L_1
- Measures:
  a. composition, speed, and direction of solar wind
  b. Magnetic field strength and direction

Space Weather

Solar wind
Carries Coronal Mass Ejections and particles from the sun

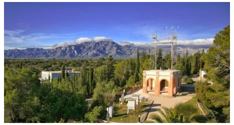


The electric field carried by the solar wind produces changes in magnetic field across Earth's surface, resulting in **Geomagnetically Induced Currents (GICs)**

ACE satellite

- Located at Lagrange point L_1
- Measures:
  a. composition, speed, and direction of solar wind
  b. Magnetic field strength and direction

Ground observatories:
Measure magnetic field strength and direction



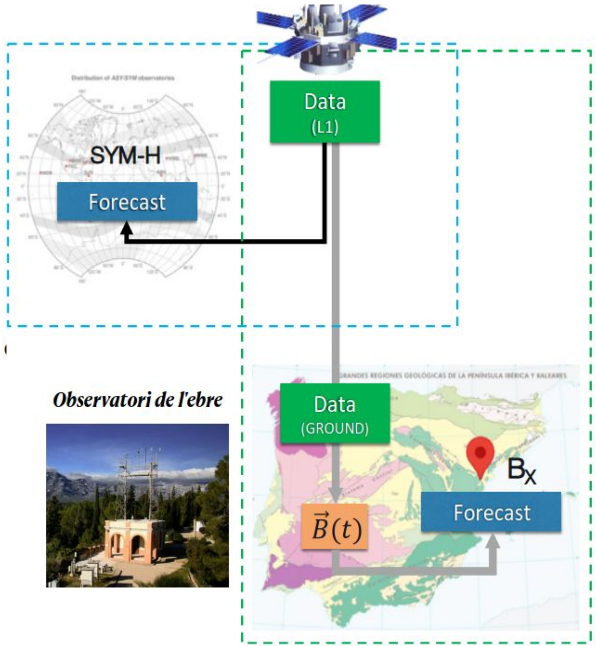
Steele Hill/NASA

Effects of GICs and goals

Impacts of GICs at ground level

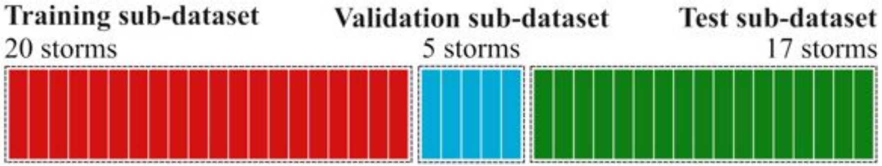
- Disrupt radio communication and navigation
- Impact on electric power grids

- Locations of high latitude are particularly at risk of the harmful effects of GICs.
    - Low latitude locations have had a history of GIC related events.
- The **objectives of this project** are to
    - **forecast SYM/H** (a magnetic field index) using data from ACE.
    - **forecast magnetic field at ground level** using data from ground observatories



Input variables

- We use the **same storms and time ranges** as Siciliano et al because we want to reproduce their results.
- These were chosen for their size and diversity in terms of quantity of peaks and shapes.



Data sample contains 42 of the most intense geomagnetic storms, distributed in two solar cycles (1998-2018)

Train

Label	Start Date	Days	SYM-H
T1	14/02/1998	8	-119*
T2	02/08/1998	6	-168*
T3	19/09/1998	10	-213
T4	16/02/1999	8	-127*
T5	15/10/1999	10	-218
T6	09/07/2000	10	-347
T7	06/08/2000	10	-235*
T8	15/09/2000	10	-196*
T9	01/11/2000	14	-174*
T10	14/03/2001	10	-165*
T11	06/04/2001	10	-275
T12	17/10/2001	10	-210
T13	31/10/2001	10	-320
T14	17/05/2002	10	-116*
T15	15/11/2003	10	-490
T16	20/07/2004	10	-208
T17	10/05/2005	10	-302*
T18	09/04/2006	10	-110*
T19	09/12/2006	10	-211*
T20	01/03/2012	10	-149

Test

Label	Start Date	Days	SYM-H
T1	22/06/1998	8	-120
T2	02/11/1998	10	-179*
T3	09/01/1999	9	-111
T4	13/04/1999	6	-122
T5	16/01/2000	10	-101*
T6	02/04/2000	10	-315
T7	19/05/2000	9	-159*
T8	26/03/2001	9	-437
T9	26/05/2003	11	-162*
T10	08/07/2003	10	-125*
T11	18/01/2004	9	-137*
T12	04/11/2004	10	-394*
T13	10/09/2012	25	-138
T14	28/05/2013	7	-134
T15	26/06/2013	8	-110
T16	11/03/2015	10	-234
T17	22/08/2018	12	-205

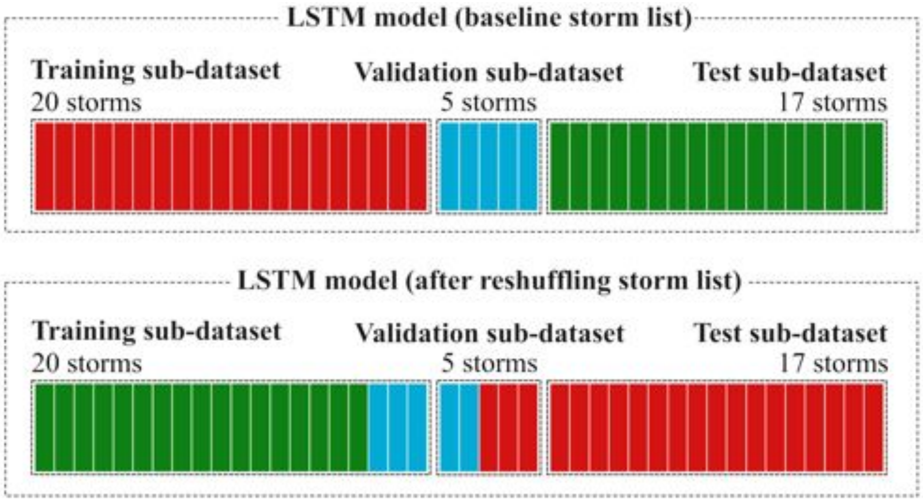
*storms with more than one peak

Validation

Storm N.	Start Date	Days	SYM-H
V1	28/04/1998	10	-268
V2	19/09/1999	7	-160
V3	25/10/2003	9	-432*
V4	18/06/2015	10	-207*
V5	01/09/2017	10	-146*

Input variables: alternative split

- A reordering of the storms within the splits was used to check the **robustness** of the algorithm with respect to the train-test-validation split.
    - Performance metrics in the second split were not different from the first
    - We may extend this idea and perform full k-fold cross validation



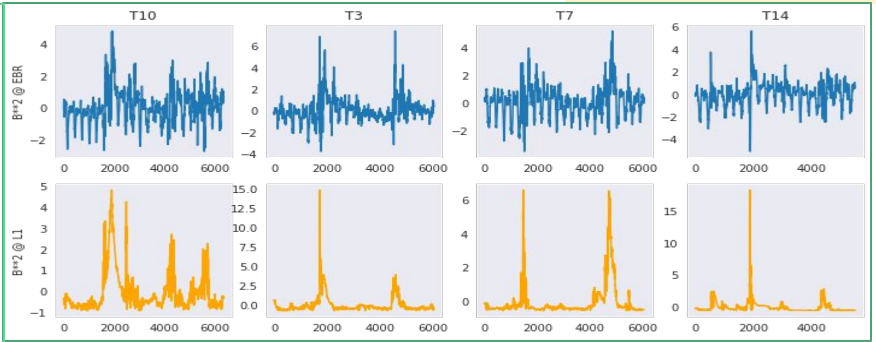
Training procedure

- For this research, **data** are taken from **ACE**, at the L1 Lagrange point, and from the **Ebre ground level observatory** in Spain

L1 variables only

- Feature variables:
  - SYM/H
  - By^2
  - B^2
  - B_z
- Target variable
  - SYM/H (future)

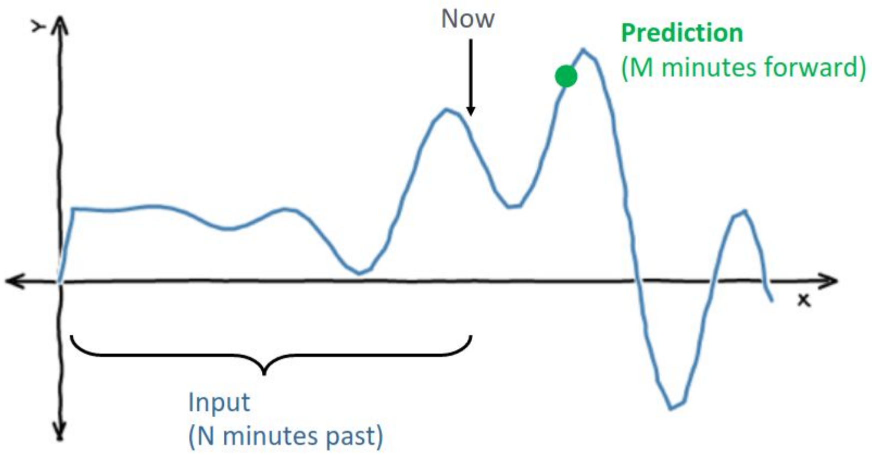
An example of how the magnetic field measurements at both locations compares for four storms:



*values of B**2 go below 0 because of standardization

Training procedure

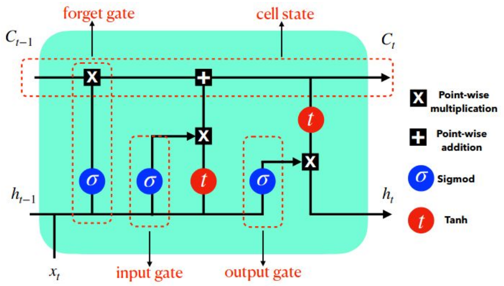
- The data has a sequence dependence because it is a time-series
- The benchmark we choose to use for forecasting is an hour
  - roughly what would be needed to respond to a GIC alert



Choice of machine learning algorithm

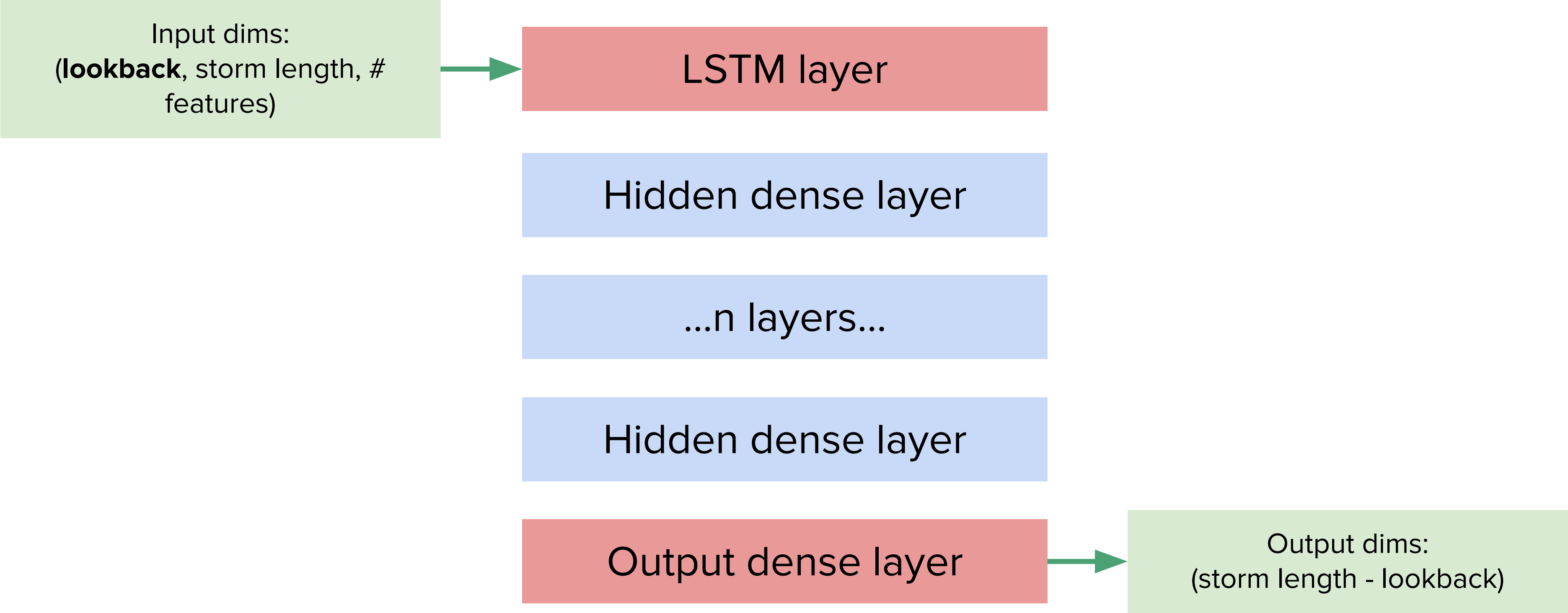
- **Long Short Term Memory neural network (LSTM)** are good for data with a sequence dependence, so they are very well suited to predict time-series like ours.
    - LSTM an improvement on the standard recurrent neural network because it solves the vanishing gradient problem by making its 'short term memory last a long time'
- **xLSTM and Transformers**: xLSTM is a rework of LSTM that includes new memory mixing and exponential gating, and it seems to perform as well as state-of-the-art Transformers. Any of these two are a natural next step after LSTM. We will be comparing results with the present work using these two architectures in the future.

For the rest of the research, **LSTM is ultimately used**



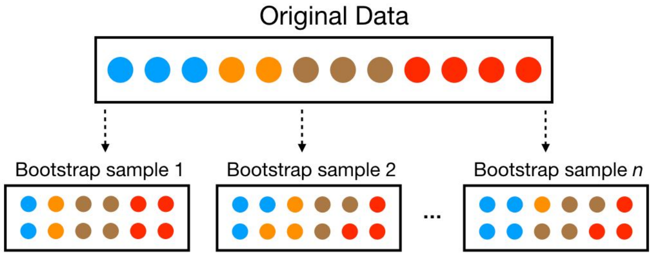
An LSTM cell

Model Architecture

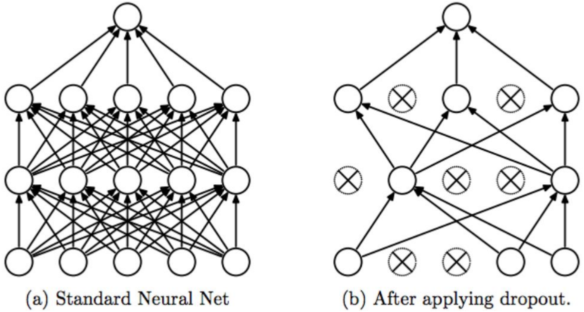


Uncertainty estimation: bootstrap and dropout

- Bayesian inference suggests that common **regularization techniques** in machine learning, like bootstrap, dropout and others, are already good at providing **uncertainty estimations** for final results and predictions.
- **Bootstrap vs dropout**
  - In the present study, **bootstrap uncertainty estimations** for predictions tended to include more of the test data around the peaks while giving larger mean square error (MSE) uncertainty.



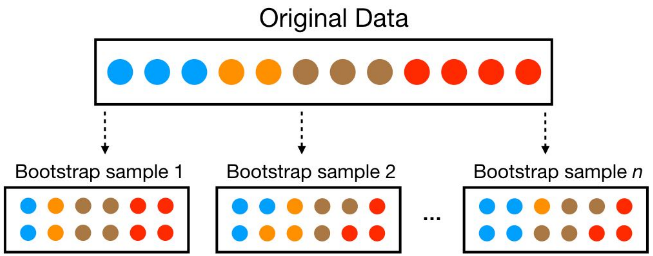
Bootstrap procedure



Dropout in different neural network layers

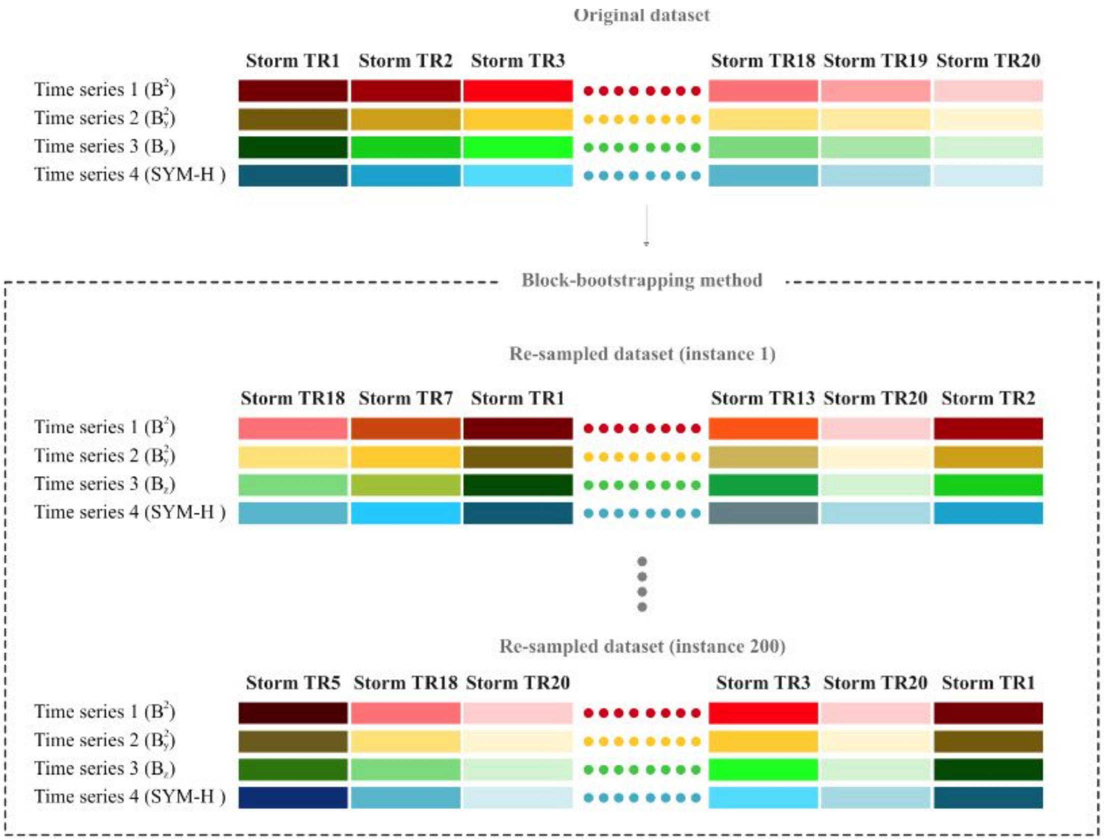
Bootstrap in this work

- Bootstrap,
  - Training is repeated on different samplings with replacements of the original dataset.



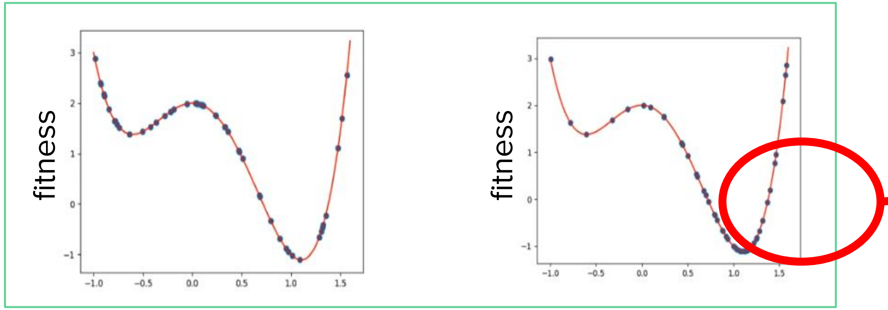
Bootstrap in this work

- Bootstrap,
  - Training is repeated on different samplings with replacements of the original dataset.
- Block bootstrap,
  - For time-series data, chunks of data need to be grouped in blocks to conserve time dependence



Hyperparameter optimization

- Hyperparameters are values that control the learning process. We identified four possibly important hyperparameters in our LSTM setup: **learning rate**, **look-back time**, **number of dense hidden layers** and **number of units in each hidden layer**.
- The choice of hyperparameter values is done by trial and error: one trains and tests data using different combinations of hyperparameters in their **multidimensional space** and optimises **fitness or objective function** result.
    - The possible combination of hyperparameter values is great or just infinite, so the choice of what combinations to try itself is done in different ways,

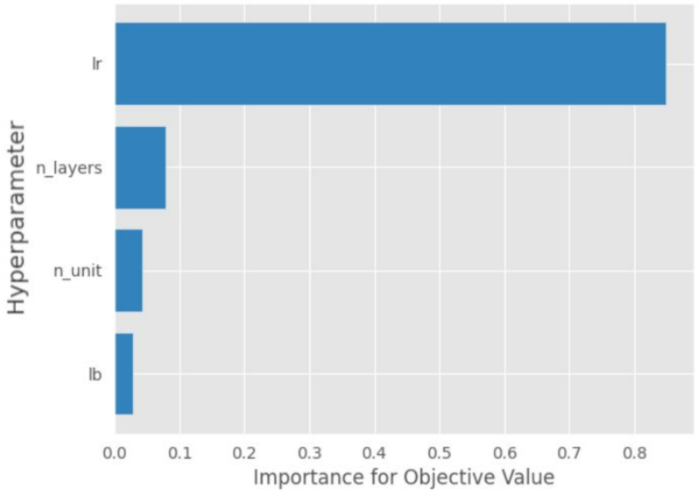


Random search   Bayesian optimization (Optuna)

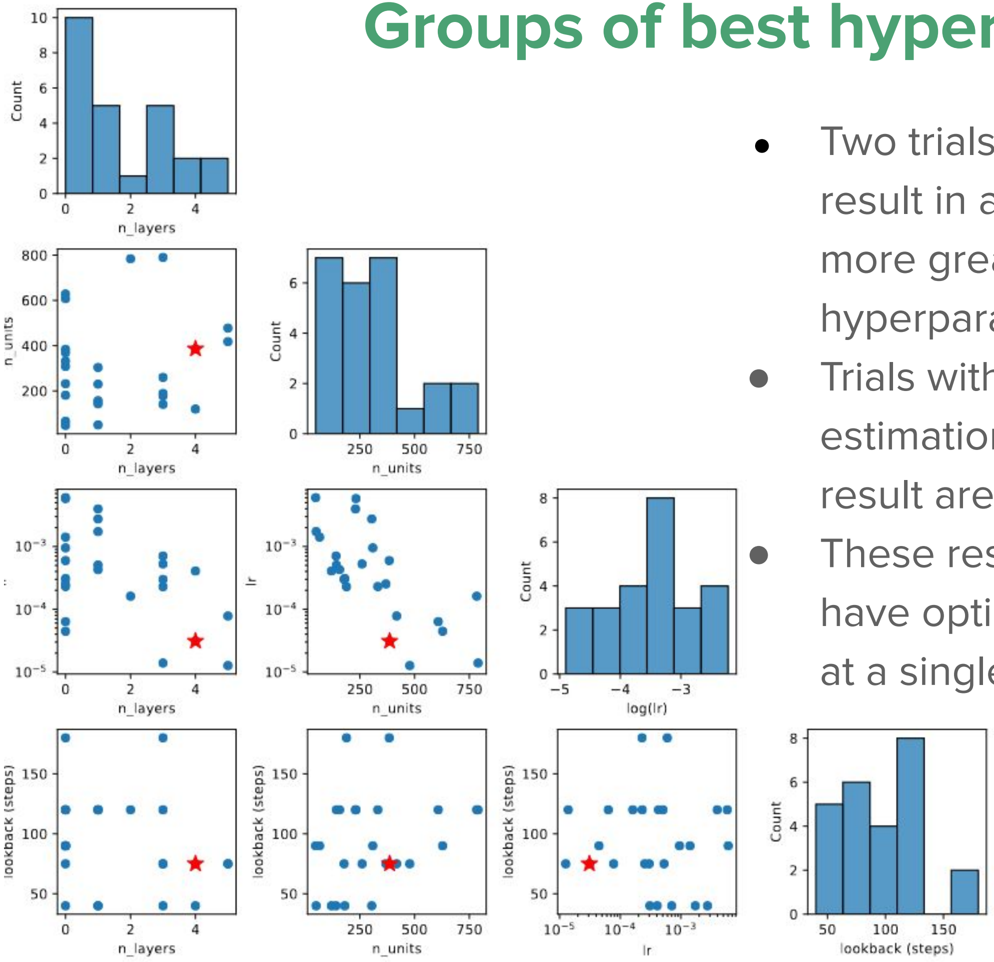
Bayesian optimization identifies potential wells

Hyperparameter optimization results

- **Optuna** offers a systematized way of searching the multidimensional hyperparameter space through **bayesian optimization**.
  - This is more efficient than grid search or random sampling
- The graph shows the relative importance of each hyperparameter with respect to the loss function
  - The **learning rate**, for all models, is always the **most dominant contributor**.



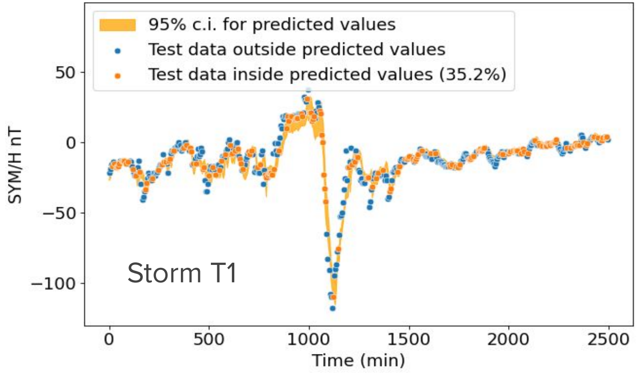
Groups of best hyperparameter sets

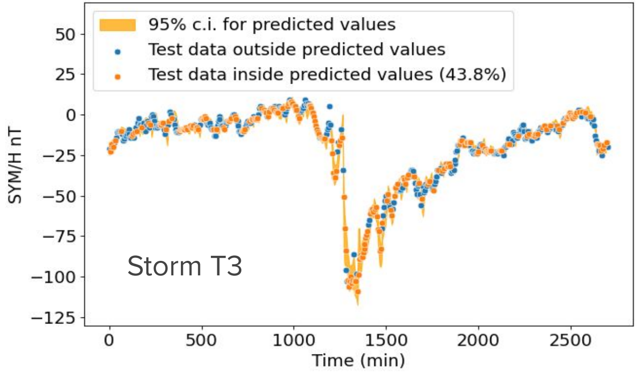


- Two trials with the same hyperparameter set can result in a value of the loss function that varies more greatly than trials with different hyperparameters.
- Trials with uncertainty intervals for the MSE estimation that overlapped with the best optuna result are labeled ‘best trials’
- These results show that some hyperparameters have optimal values at different regions, and not at a single optimal ‘well’

Predictions (SYM-H)

- Plots show the target variable **(SYM/H) prediction** with respect to time for two of the 17 test storms.
- The **orange bands** represent the **95% confidence interval** of the predicted value by our model.
- The dots represent the actual test values, orange if within the uncertainty band and blue if outside the band.
- **The percentage of predicted values** is the proportion of test values inside the uncertainty band. This was pivotal in choosing bootstrap over other ways to estimate uncertainty



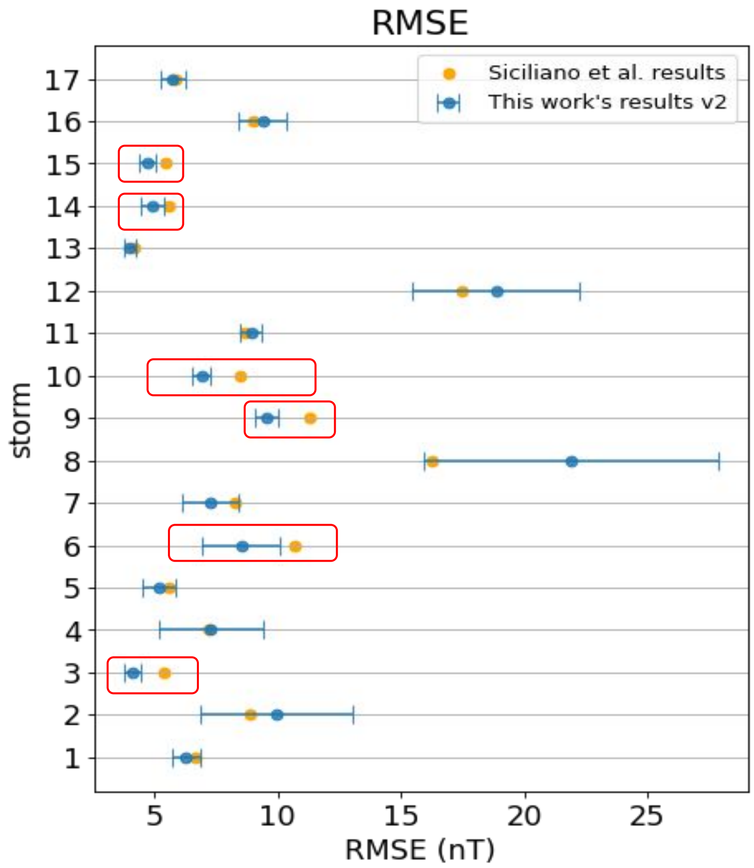


Uncertainty bands: Bootstrap with 200 runs

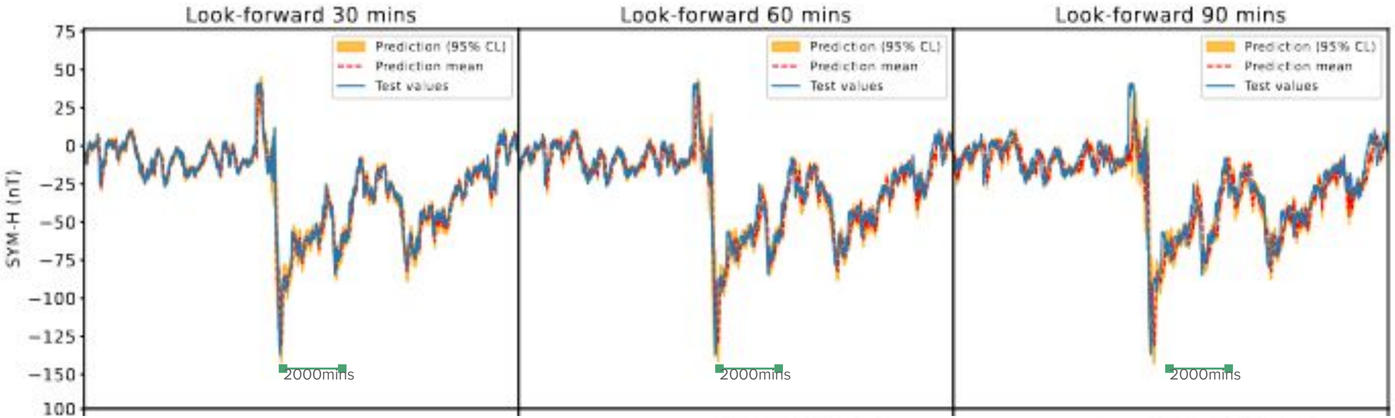
look-forward : 60 mins

Predictions (SYM-H)

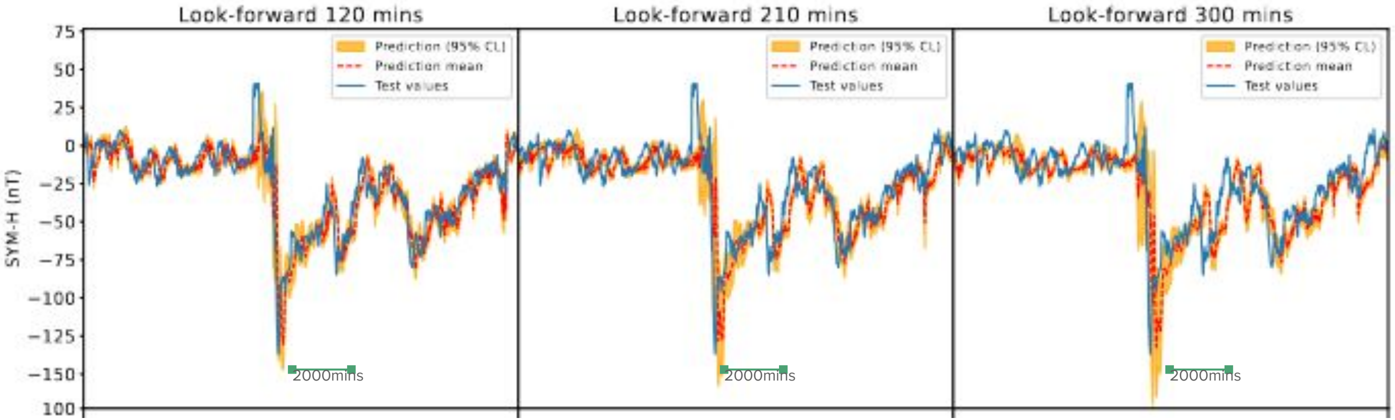
- **Comparison between our prediction and the model from Siciliano et al.**
- For RMSE of the target variable, all but six of the referenced values are within the 95% confidence intervals for the RMSE obtained with this work's model.
- In all of those six cases, **our present model gives RMSE values that are lower with 95% confidence**.



Predictions (SYM-H)

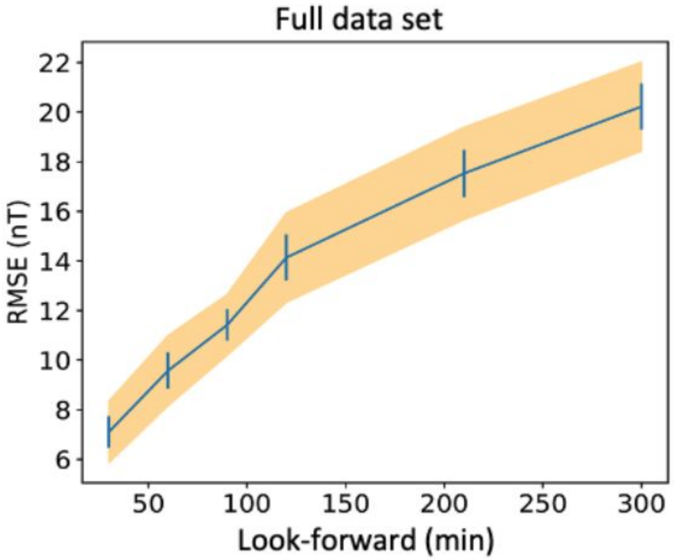


- Multiple-hour predictions for storm 11.



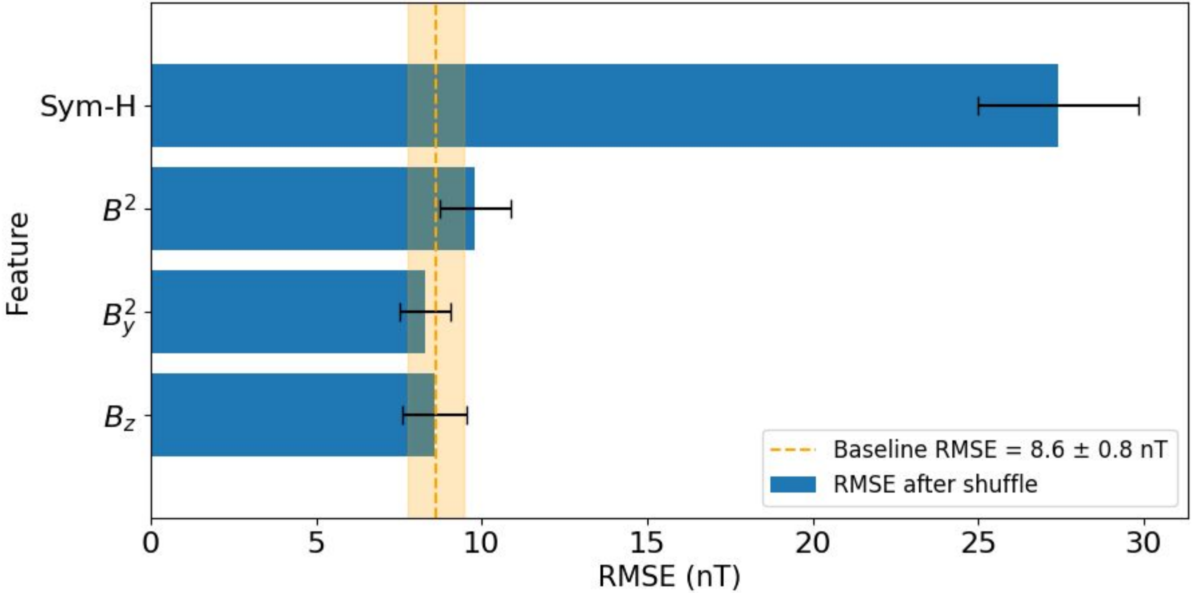
Predictions (SYM-H)

- Multiple hour predictions for storm 11, storm 12 and all storms:
  - RMSE increases with respect to look-forward
  - RMSE uncertainty increases with respect to look-forward



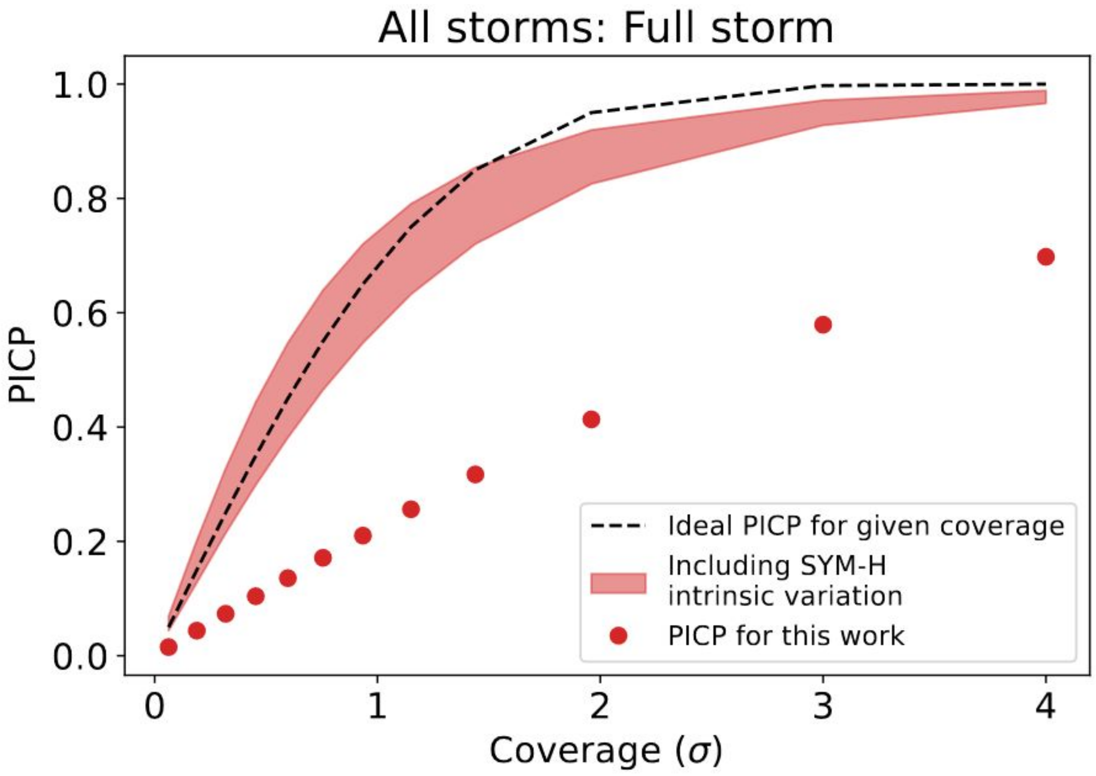
Feature Importance

- Values or a particular variable are shuffled and the rMSE is calculated.
- A bigger error bar signifies that the shuffled variable is more important, as its spoiling results in large error.



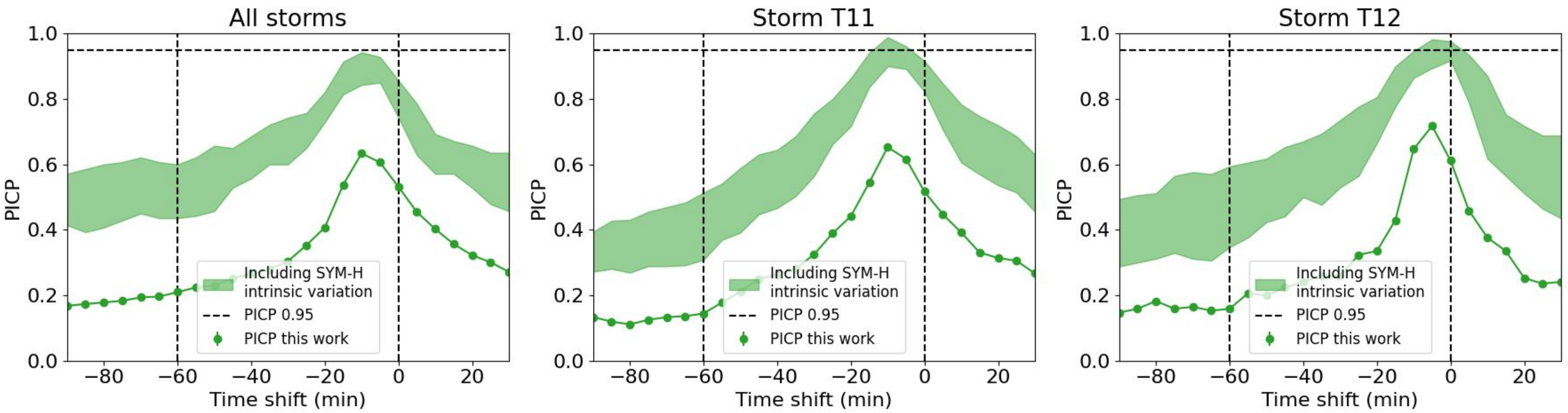
Uncertainty analysis: intrinsic variation

- **PICP**: the prediction interval coverage probability gives us an evaluation of how much of the data is within uncertainty bounds.
- By itself, the estimated uncertainties seems to be **underestimated**.
- By including SYM-H **intrinsic variation**, PICP approaches ideal value



Uncertainty analysis: time mismatch

- By shifting the predictions with respect to the true values in time, we observe that the PICP improves when the predictions are shifted back.
  - This is a inherent feature of RNN architectures forecast models, and so **time mismatch is a source of error**.
- This, together with the inclusion of the intrinsic variation of SYM-H might completely correct the missing uncertainty estimation.



Adding more variables

- Other works forecasting SYM-H include more variables from the L1 ACE satellite. With their inclusion, we may compare
  - Feature importance plot
  - Uncertainty estimation (PICP)
  - Time mismatch

Original variables

- Feature variables:
  - SYM/H
  - B_x
  - B_y
  - B_z

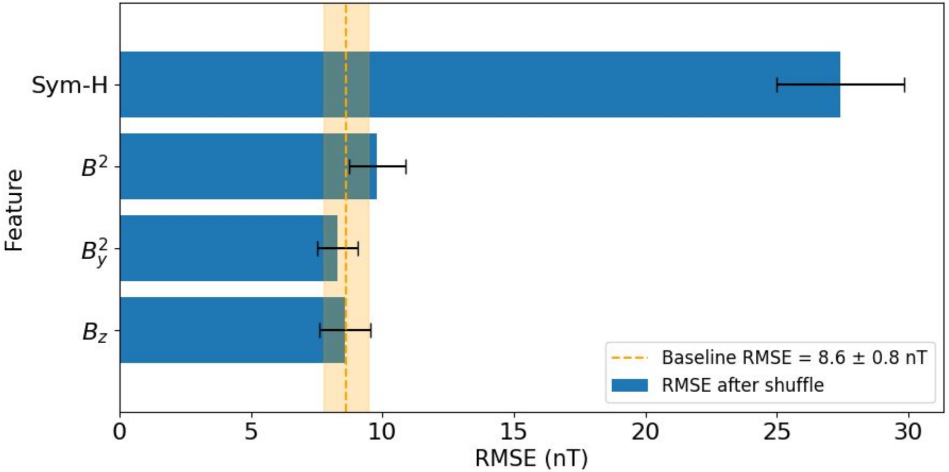
Solar wind additional variables

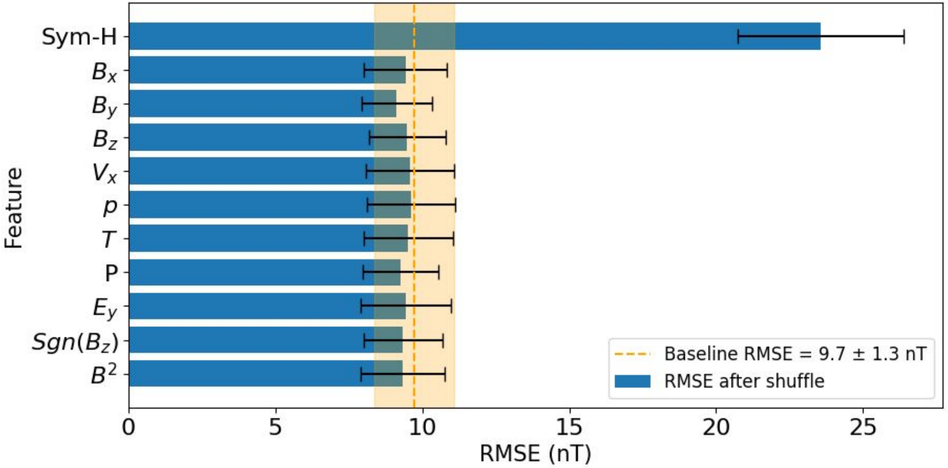
- Feature variables:
  - V_x: velocity
  - ρ: density
  - T: temperature
  - P: dynamic pressure
  - E_s: electric field

Target variable
SYM/H (future)

Adding more variables: Feature importance

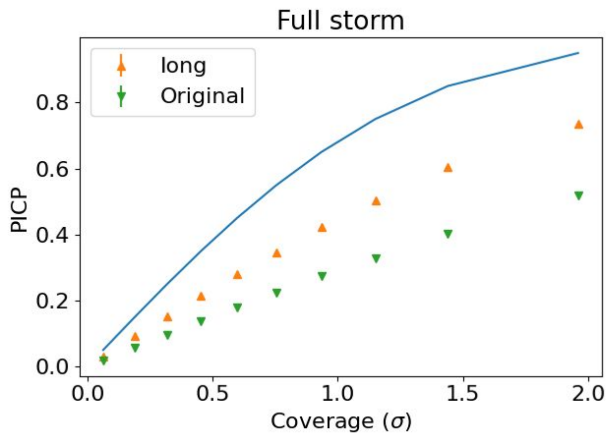
- In terms of RMSE, new variables don’t seem to pose an improvement.
- The baseline values are compatible.
- The uncertainty estimation is bigger with the additional variables





Adding more variables: PICP

- With the addition of new variables, the PICP values are closer to the ideal curve. Uncertainty estimation is bigger.
  - This could mean that uncertainty estimation is more ‘truthful’, or simply that the forecasting is ‘messier’



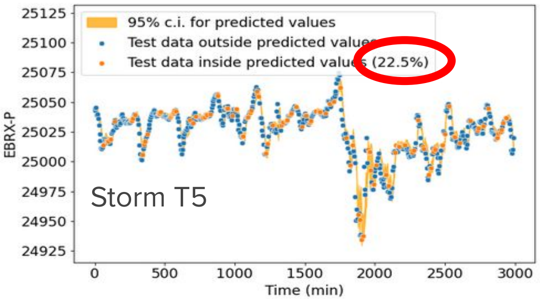
Future work: ground level forecasting

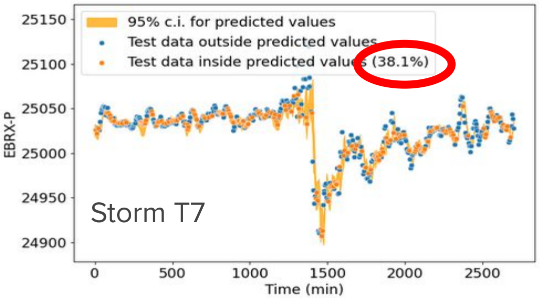
- For the **second analysis** in this work, we aim to predict **magnetic field at ground level** using data taken only from the **Ebre observatory** at ground level.
- We are interested in forecasting the horizontal geomagnetic field. $\mathbf{B_x}$ is chosen as the target variable. An advantage of this choice over B_y is that B_x has the larger influence on the appearance of GICs.

Preliminary results: Ebre predictions

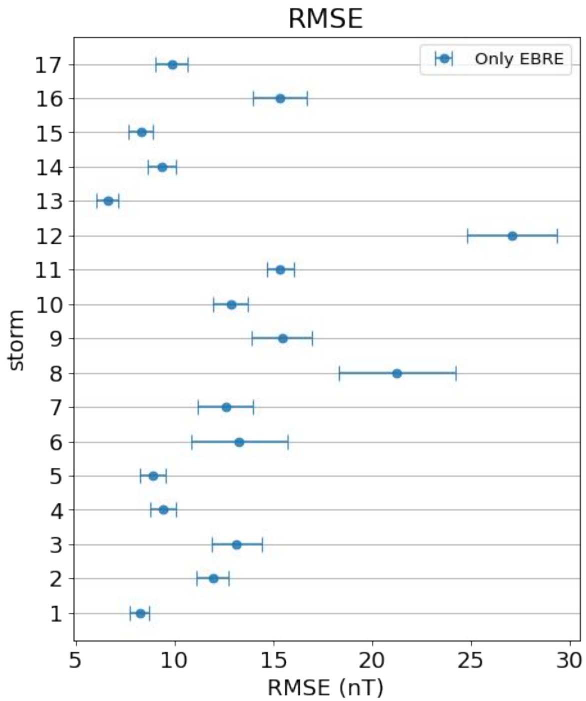
- We obtain an analogous **model for ground level prediction**, which gave a forecasting that contained less of the data and bigger RMSE in nT than the SYM/H model in comparison

Uncertainty bands: Bootstrap with 200 runs





look-forward : 60 mins

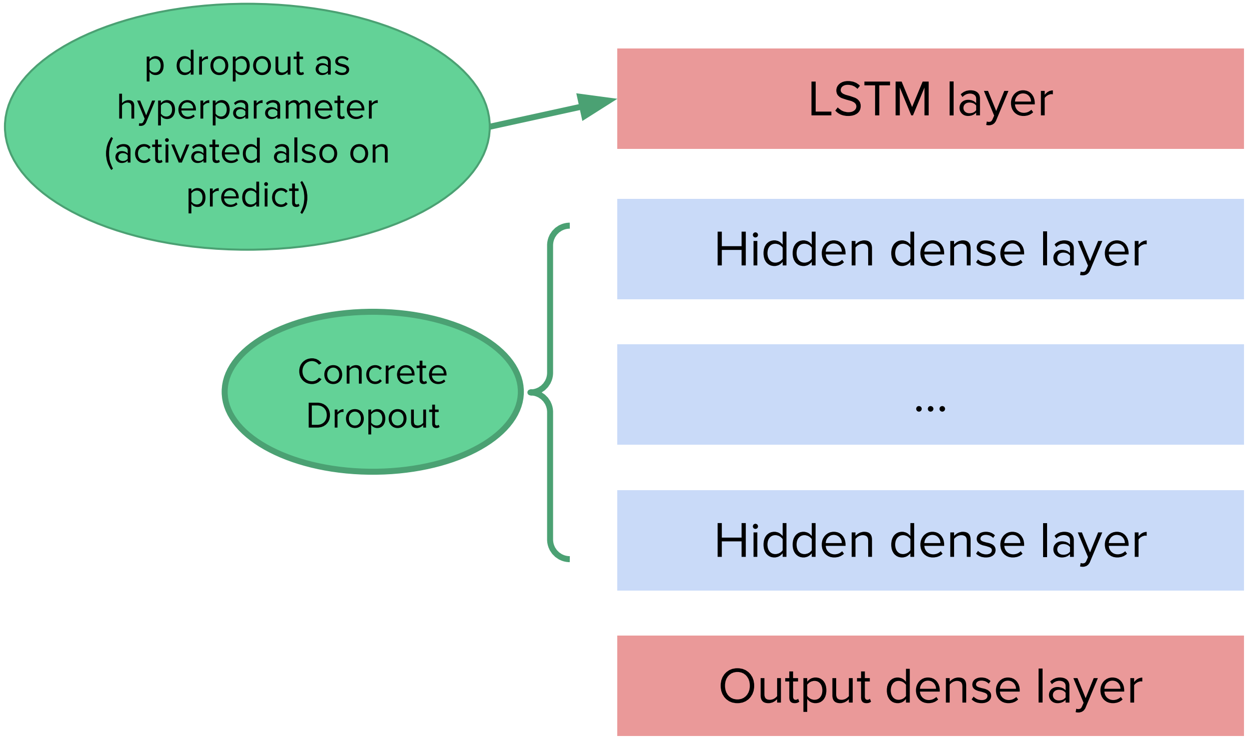


Conclusions

- We obtained **a forecast model** for **SYM/H** which features **uncertainty** measures via **bootstrap** and dropout.
- **RMSE** results for L1 are either **compatible** with Siciliano et al in most cases or **better** in the ones that are statistically different. The improvement can be mainly explained by the **hyperparameter optimization** via Optuna and the betterment of the initial dataset.
- We observe that both **RMSE values and their uncertainties grow with higher look-forward values**, making forecasting increasingly more unreliable.
- Considering an ideal PICP **uncertainty values appear to be underestimated**. The intrinsic variation of SYM-H and the inherent systematic time mismatch of RNN architectures may explain this underestimation
- **The addition of other solar wind variables** does not improve either RMSE or time mismatch. Uncertainty estimation is larger, which impacts positively on PICP.
- We obtain an analogous **model for ground level prediction**, which gave a forecasting that contained less of the data (smaller PICP) and bigger RMSE than the SYM/H model.

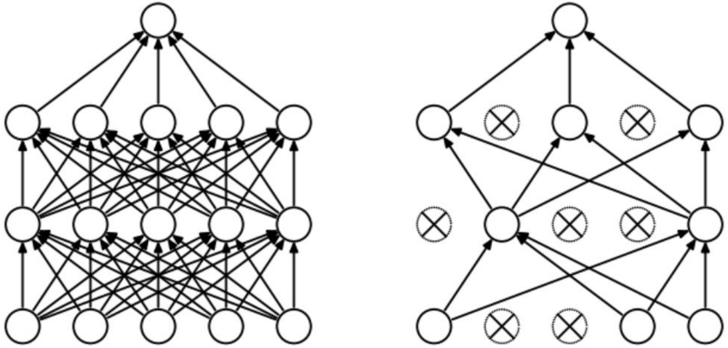
Thank you

Model Architecture for Dropout



Bootstrap and Dropout in this work

- Dropout,
  - A set proportion of random units in the neural network are turned off every time data is predicted with the model. This proportion is represented by p.
- Concrete dropout
  - Continuous approximation of the effect of dropout on the loss function can be automatically optimized for the dropout p.



(a) Standard Neural Net (b) After applying dropout.

Image source: Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting", JMLR 2014

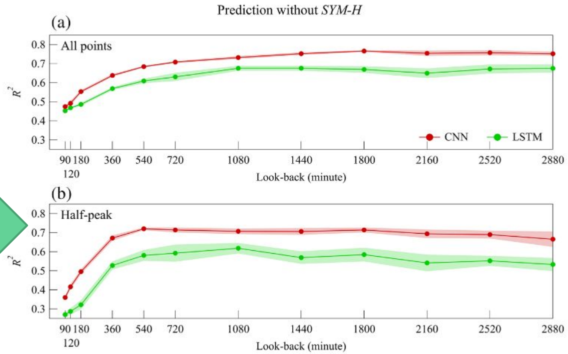
Identified hyperparameter

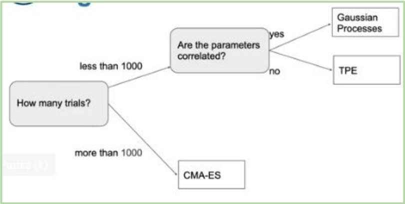
- Hyperparameters are values that control the learning process. We identified four hyperparameters in our LSTM setup:
    - **Learning rate**. In this work, this value changes in a triangular cycle throughout epochs, and the value to be optimized is the central value of the cycle, with its width given by the standard deviation of the chosen trial.
    - **Look-back**, relevant during the preparation of data for LSTM algorithm.
    - **Number of dense neural layers** after the LSTM layer and before the output dense layer.
    - **Number of units** in the inner dense layers. We simplified this to mean the same number for all dense layers.

Hyperparameter search

- The choice of what combinations to try itself is done in different ways:
  - Grid search
  - Random Sampling
  - Optuna: Bayesian optimization flavour called ``Tree-structured Parzen Estimator'' (TPE)
  - Genetic algorithm
  - Etc...
- Optuna offers different search algorithms. The one chosen was the default one, TPE, as we didn't want to do more than 1000 trials and the hyperparameters are [almost] uncorrelated.

An example of a 1 hyperparameter grid search in Siciliano et al

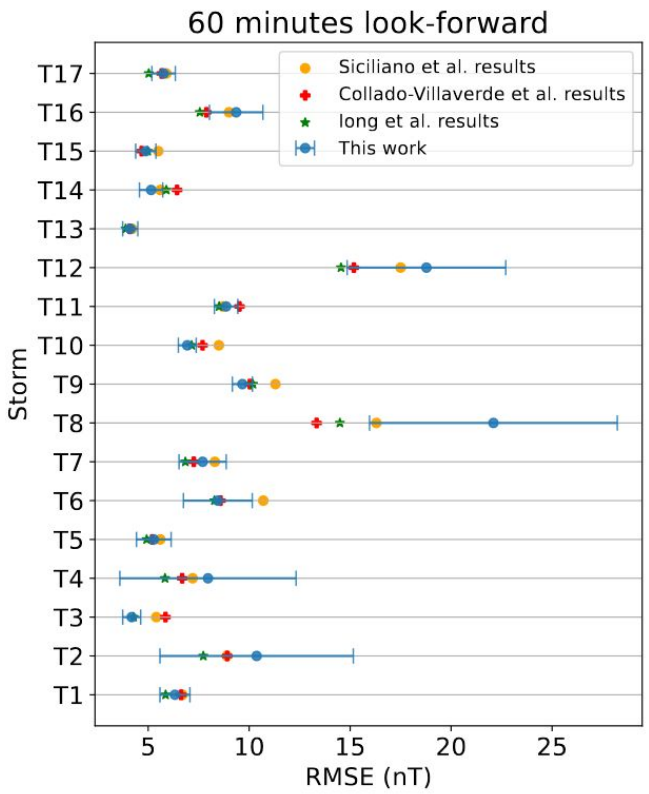




Optuna cheat sheet taken from tutorial video

Comparison of RMSE with other references

- Newer works have also reproduced Siciliano et al results, with considerable improvements
- An important observation is their inclusion of other ACE satelite variables that we chose to omit
- Our data is compatible up to 2 sigma with their results, except in storms T17, T16, T12 and T8 where they outperform us.

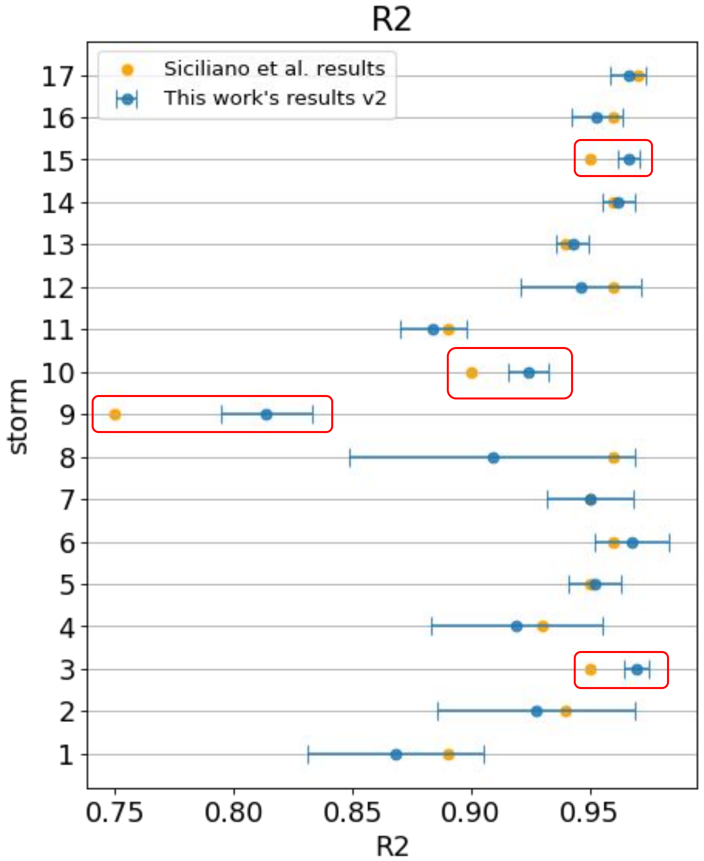


Ground level forecasting

- Data from the Ebro in the 90's and early 2000s had many missing observations due to a nearby railway. The team at Ebro complemented this data by referring to the San Pablo de los Montes-Toledo observatory to compare and interpolate in case of gaps in the Ebro observations.

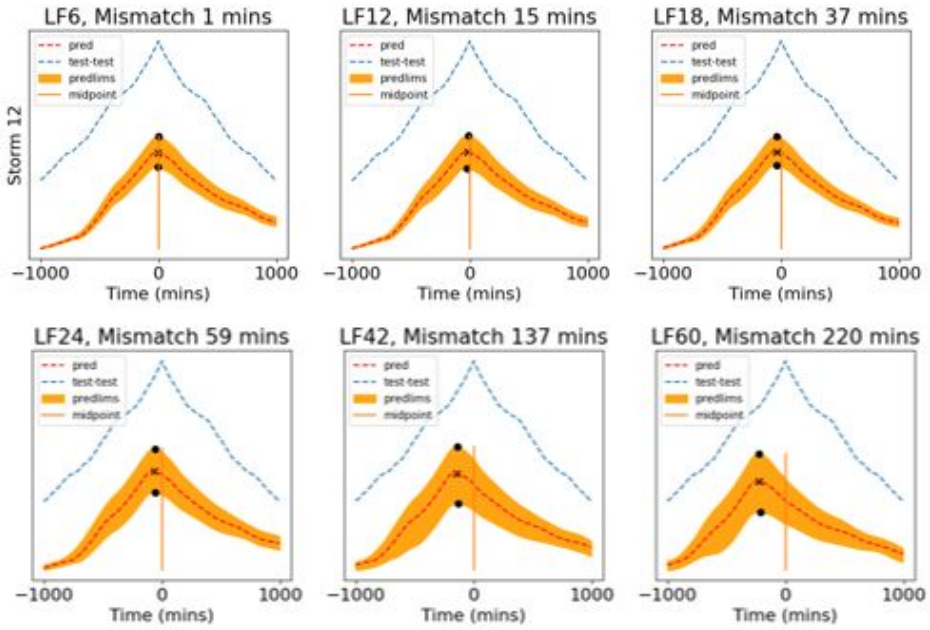
This work's R^2 results vs Siciliano

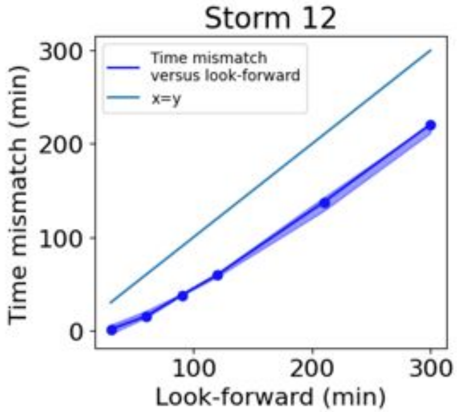
- For the coefficient of determination, R^2:
  - 13 of the referenced values are within our 95% confidence intervals
  - 4 of the referenced values are below our 95% confidence intervals



L1 predictions (SYM-H)

- **Cross-correlation** between prediction and true values shows a consistent time mismatch throughout all look-forward predictions
  - This time mismatch is lower than the look-forward, which means some predictive power is gained





Next steps

Prediction of derivative of SYM/H.

- Different time derivative calculation methods
- More noise and tighter look-forward window