Universidad de Alcalá

PhD. Program in Space Research and Astrobiology

Deep Neural Networks for Geomagnetic Indices Forecasting

PhD. Thesis Presented by

Armando Collado Villaverde

2025

PhD. Program in Space Research and Astrobiology

Deep Neural Networks for Geomagnetic Indices Forecasting

PhD. Thesis Presented by

Armando Collado Villaverde

Advisors

Dra. María Dolores Rodríguez Moreno
Dr. Pablo Muñoz Martínez

Alcalá de Henares, 2025

“Space weather forecasting: when you need a PhD to tell people it’s going to be "sunny".”

A mis padres.

Acknowledgements

The darker the night, the brighter the stars.

En primer lugar, quiero expresar mi más profundo agradecimiento a ti, mi codirector de tesis, Pablo Muñoz Martínez. Tu apoyo y tu confianza desde el principio han sido fundamentales para la realización de este trabajo y para mi desarrollo profesional. Tu conocimiento, tu dedicación y tu ánimo me han ayudado a superar cada uno de los retos que he encontrado en el camino, que no han sido pocos ni fáciles. Durante este tiempo, más que un director, has sido un amigo, y por ello siempre te estaré agradecido.

También quiero hacer una especial mención a Consuelo Cid, ya que sin sus enseñanzas la meteorología espacial habría sido como saber cuán soleado está el día. Muchas gracias por haberme guiado por este camino tan apasionante y sobre todo, muchas gracias por tu apoyo en los momentos difíciles.

Agradezco también a mis compañeros del laboratorio E-31, actuales (Hugo, Iván y Mario) y pasados (Fernando y Daniel), que han sido mi familia laboral. Las innumerables conversaciones y consejos han sido clave para completar esta tesis. Asimismo, quiero expresar mi agradecimiento a David Fernández Barrero, quien me introdujo en la investigación y la Inteligencia Artificial; sin ti, nada de esto hubiera sido posible.

Por supuesto, los verdaderos responsables de que se haya podido completar esta tesis y mi desarrollo profesional son mis padres, Pilar y Armando. Me han apoyado siempre incondicionalmente y son la razón por la que soy quien soy. Han sido mi apoyo en los buenos y malos momentos, siempre alentándome a seguir adelante. Esta tesis es tan vuestra como mía, y os la dedico con todo mi cariño.

No puedo olvidar a mis amigos, tanto los de siempre de Cifuentes como los de la Universidad, quienes han escuchado mis monólogos sobre la investigación y me han acompañado durante estos años. Vuestra compañía ha sido una fuente de alegría y motivación.

During this thesis, I spent some time in Darmstadt with the Space Weather team at ESOC. I want to thank all the members of the team, but especially Judit Palacios Hernández for all her help in making the stay a wonderful experience. Finally, I want to thank ESA's technical officer, Dr. Alexi Glover, for her continuous support and for making this PhD possible through the Open Space Innovation Platform programme.

Thank you all for being a part of this journey.

This thesis was supported by the European Space Agency under the Networking and Open Space Innovation Platform titled "Deep Neural Networks for Geomagnetic Forecasting", contract 4000137421/22/NL/GLC/my.

Resumen

La meteorología espacial se refiere a las condiciones ambientales en el espacio influenciadas por la actividad solar, que incluye las emisiones del Sol, el viento solar y la perturbación causada en la magnetosfera, ionosfera y termosfera de la Tierra. Uno de los fenómenos de clima espacial más impactantes es la tormenta geomagnética, que ocurre cuando el viento solar y el campo magnético interplanetario se intensifican y perturban el campo magnético terrestre. Las tormentas geomagnéticas son la fuente de las perturbaciones geomagnéticas más significativas y pueden causar una gran variedad de graves consecuencias, afectando sistemas tecnológicos críticos como redes eléctricas, comunicaciones por satélite, sistemas globales de navegación por satélite (GNSS) y aviación. A medida que nos acercamos a un máximo solar, se espera que aumenten la frecuencia e intensidad de estas tormentas, amplificando los riesgos asociados con la meteorología espacial. Dada la creciente dependencia de la tecnología y la mayor vulnerabilidad a la meteorología espacial, resulta esencial desarrollar sistemas de predicción precisos y confiables.

En este contexto, esta tesis se centra en la creación de un modelo de red neuronal profunda (DNN) para la predicción en tiempo real de índices geomagnéticos clave, en particular SYM-H y ASY-H, con un enfoque en el SYM-H. El modelo está diseñado para aprovechar los recientes avances en el aprendizaje automático (ML) y la disponibilidad de datos continuos sobre el viento solar y el campo magnético interplanetario (IMF) proporcionados por satélites como el Advanced Composition Explorer (ACE), ubicado en el punto de Lagrange 1 (L1). Esta combinación de técnicas avanzadas de ML y el extenso histórico disponible de datos proporciona una potente herramienta para predecir la intensidad y el momento de impacto de las tormentas geomagnéticas.

Una innovación clave de este trabajo es el establecimiento de un sistema basado en estadística para identificar, clasificar y delimitar tormentas geomagnéticas, aplicable tanto a los índices SYM-H y ASY-H, como a otros índices relevantes. Este sistema nos ha permitido expandir los conjuntos de datos existentes al incorporar tormentas con datos provisionales, lo que permite evaluar el rendimiento del modelo durante su operación en tiempo real, donde los datos no siempre están completos o no son de suficiente calidad. Además, hemos desarrollado una nueva métrica, llamada “Binned Forecasting Error (BFE)”, diseñada específicamente para evaluar el rendimiento del modelo con mayor precisión en este contexto, donde los valores extremos y poco frecuentes son más importantes

que los más comunes. Las métricas de regresión tradicionales, como el RMSE, no son del todo adecuadas para la evaluación de modelos de predicción de tormentas geomagnéticas debido a su incapacidad para capturar las características distintivas de los períodos de alta actividad, mientras que el BFE ofrece una evaluación más precisa.

El modelo DNN desarrollado supera los sistemas previos en la predicción de índices geomagnéticos, especialmente para eventos de tormentas extremas. Al integrar intervalos de confianza basados en cuantiles junto con las predicciones puntuales, el modelo ofrece una imagen más completa de la actividad geomagnética esperada, lo que es crucial para los responsables de la toma de decisiones en sectores que dependen de predicciones precisas de la meteorología espacial. Esta mejora aumenta la utilidad del sistema al proporcionar información sobre la incertidumbre asociada a cada predicción.

Más allá de los índices globales, este estudio ha adaptado con éxito el modelo DNN para predecir índices geomagnéticos locales. A través de un análisis de las perturbaciones locales, encontramos que la actividad geomagnética local puede diferir significativamente de la global, dependiendo de la posición geográfica de la estación y del Tiempo Local Magnético (MLT) en el que la perturbación impacta la Tierra. El modelo ha demostrado un sólido rendimiento en la predicción de estos índices locales y, en algunos casos, se ha aplicado incluso a estaciones fuera del conjunto de entrenamiento inicial con un éxito razonable, lo que destaca las capacidades de generalización del modelo.

En conclusión, esta tesis aporta varios avances clave en el campo de la predicción de tormentas geomagnéticas. Al ampliar el conjunto de datos e introducir la métrica BFE, el modelo DNN se ha refinado para proporcionar predicciones más precisas, especialmente durante eventos geomagnéticos extremos. La integración de intervalos de confianza y la exitosa aplicación tanto a índices globales como locales representan un paso significativo en el campo de la predicción de meteorología espacial con aplicación en entornos de tiempo real. Estos avances serán esenciales a medida que nos preparemos para un aumento en la actividad de tormentas geomagnéticas durante el próximo máximo solar, ofreciendo herramientas más confiables para mitigar los impactos de la meteorología espacial en sistemas dependientes de la tecnología.

Palabras clave: Índices geomagnéticos, aprendizaje automático, redes neuronales.

Abstract

Space weather refers to the environmental conditions in space influenced by solar activity, including the Sun's emissions, solar wind, and the responses in Earth's magnetosphere, ionosphere, and thermosphere. One of the most impactful space weather phenomena is the geomagnetic storm, which occurs when heightened solar wind and interplanetary magnetic field conditions disturb Earth's magnetic field. Geomagnetic storms are the source of the most significant geomagnetic disturbances and can cause a myriad of severe consequences, affecting critical technological systems such as power grids, satellite communications, Global Navigation Satellite System (GNSS), and aviation. As we approach a solar maximum, the frequency and intensity of these storms are expected to increase, amplifying the risks associated with space weather. Given the growing reliance on technology and the increased vulnerability to these disturbances, the development of accurate, real-time forecasting systems for geomagnetic activity has become essential.

In this context, this thesis focuses on creating a Deep Neural Network (DNN) model for real-time forecasting of key geomagnetic indices, particularly SYM-H and ASY-H, with an emphasis on the SYM-H. The model is designed to take advantage of recent advancements in Machine Learning (ML) and the availability of continuous solar wind and interplanetary magnetic field (IMF) data from spacecraft such as the Advanced Composition Explorer (ACE), located at the Lagrange 1 (L1) point. This combination of state-of-the-art ML techniques and real-time data forms a powerful tool for predicting the intensity and timing of geomagnetic storms

A key innovation of this work is the establishment of a statistically backed system for identifying, classifying, and bounding geomagnetic storms, applicable to both SYM-H and ASY-H indices, as well as other relevant indices. This system has enabled the expansion of existing datasets; moreover, we have incorporated storms with provisional data, improving the model's ability to function in real-time settings where complete data is not always available. Furthermore, we have developed a new metric, the Binned Forecasting Error (BFE), specifically designed to evaluate the model's performance more accurately in this context, where extreme, rare values are more important than common ones. Traditional regression metrics, such as RMSE, are not well-suited for geomagnetic storm forecasting, as they fail to capture the distinct characteristics of high-activity periods, while BFE offers a more targeted assessment.

The DNN model outperforms previous approaches in geomagnetic forecasting, particularly for extreme storm events. By integrating quantile-based confidence intervals alongside the point predictions, the model offers a more complete picture of the expected geomagnetic activity, which is crucial for decision-makers in sectors that depend on accurate space weather predictions. This addition improves the system's usability by providing insights into the uncertainty surrounding each forecast.

Beyond global indices, this study has successfully adapted the DNN model to forecast local geomagnetic indices. An analysis of local disturbances revealed that local geomagnetic activity can differ significantly from global averages, depending on the station's geographical position and the Magnetic Local Time (MLT) of the disturbance impact. The model has demonstrated strong performance in forecasting these local indices and, in some cases, has even been applied to stations outside of the initial training set with reasonable success, highlighting the model's generalization capabilities.

In conclusion, this thesis makes several key contributions to the field of geomagnetic storm forecasting. By expanding the dataset and introducing the BFE metric, the DNN model has been refined to provide more accurate forecasts, particularly during extreme geomagnetic events. The integration of confidence intervals and the successful application to both global and local indices represent a significant step forward in operational space weather forecasting. These advancements will be essential in preparing for the increase in geomagnetic storm activity during the upcoming solar maximum, offering more reliable tools for mitigating the impacts of space weather on technology-dependent systems.

Keywords: Geomagnetic indices, Machine Learning, forecasting.

Contents

List of Figures

List of Tables

List of Acronyms

ABG	Alibag.
ACE	Advanced Composition Explorer.
AD	Absolute Difference.
AI	Artificial Intelligence.
AIA	Atmospheric Imaging Assembly.
ANN	Artificial Neural Network.
ASC	ACE Science Center.
BFE	Binned Forecasting Error.
CDAWeb	Coordinated Data Analysis Web.
CDS	Coronal Diagnostic Spectrometer.
CELIAS	Charge, Element, Isotope Analysis.
CLF	Chambon la Forêt.
CME	Coronal Mass Ejection.
CNN	Convolutional Neural Network.
COSTEP	Suprathermal & Energetic Particle Analyser.
CRIS	Cosmic-Ray Isotope Spectrometer.
CUDA	Compute Unified Device Architecture.
CuDF	Cummulative Distribution Function.
DL	Deep Learning.
DNN	Deep Neural Network.
DSCOVR	Deep Space Climate Observatory.
EIS	Extreme-Ultraviolet Imaging Spectrometer.
EIT	Extreme UV Imaging Telescope.
EPACT	Energetic Particle Acceleration, Composition and Transport.
EPAM	Electron, Proton, and Alpha-particle Monitor.
EPIC	Earth Polychromatic Imaging Camera.

ERNE Energetic Particle Analyser.
ESA European Space Agency.
EVE Extreme Ultraviolet Variability Experiment.

FSS Forecast Skill Score.

GBM Gradient Boosting Machines.
GD Gradient Descent.
GIC Geomagnetic Induced Current.
GNSS Global Navigation Satellite System.
GOES Geostationary Operational Environmental Satellite.
GOLF Global Oscillations at Low Frecuence.
GPS Global Positioning System.
GPT-3 Generative Pre-trained Transformer 3.
GPU Graphical Processing Unit.
GSM Geocentric Solar Magnetospheric.

HDF Hierarchical Data Format.
HMI Helioseismic and Magnetic Imager.
HON Honolulu.
HPC High Performance Computing.
HRO High Resolution OMNI.
HSS High-Speed Stream.

IAGA International Association of Geomagnetism and Aeronomy.
IGRF International Geomagnetic Reference Field.
IMF Interplanetary Magnetic Field.
IMP-8 Interplanetary Monitoring Platform-8.
IRIS Interface Region Imaging Spectrograph.
ISEE-3 International Sun-Earth Explorer-3.

JAXA Japan Aerospace Exploration Agency.

LASCO Large Angle and Spectrometric Coronagraph.
LCi Local Current Index.
LDi Local Disturbance Index.
LRO Low Resolution OMNI.
LSTM Long Short-Term Memory.

MAD	Mean Absolute Difference.
MADev	Mean Absolute Deviation.
MAE	Mean Absolute Error.
MAG	Magnetometer.
MDI	Michelson Doppler Imager.
MFI	Magnetic Field Investigation.
ML	Machine Learning.
MLT	Magnetic Local Time.
MMB	Memambetsu.
MSE	Mean Square Error.
NASA	National Aeronautics and Space Administration.
NISTAR	National Institute of Standards and Technology Advanced Radiometer.
NLP	Natural Language Processing.
NN	Neural Network.
NOAA	National Oceanic and Atmospheric Administration.
OMNI	Operating Missions as a Node on the Internet.
PELT	Pruned Exact Linear Time.
PI	Prediction Interval.
PIAW	Prediction Interval Average Width.
PIBW	Prediction Interval Binned Width.
PICP	Prediction Interval Coverage Probability.
PINAW	Prediction Interval Normalized Average Width.
PlasMag	Plasma-Magnetometer.
RMSE	Root Mean Square Error.
RNN	Recurrent Neural Networks.
RTSW	Real-Time Solar Wind.
SDO	Solar Dynamics Observatory.
SGD	Stochastic Gradient Descent.
SHAP	Shapley additive explanation.
SIS	Solar Isotope Spectrometer.
SOHO	Solar and Heliospheric Observatory.

SOT	Solar Optical Telescope.
SPT	San Pablo-Toledo.
STICS	Suprathermal Ion Composition Spectrometer.
SW	Space Weather.
SWE	Solar Wind Experiment.
SWEPAM	Solar Wind Electron, Proton, and Alpha Monitor.
SWICS	Solar Wind Ion Composition Spectrometer.
SWIMS	Solar Wind Ion Mass Spectrometer.
SWPC	Space Weather Prediction Center.
TUC	Tucson.
ULEIS	Ultra-Low Energy Isotope Spectrometer.
USAF	United States Air Force.
UVCS	UltraViolet Coronagraph Spectrometer.
WAVES	Radio and Plasma Wave Experiment.
WDCG	World Data Center for Geomagnetism, Kyoto.
XRT	X-ray Telescope.

Chapter 1

Introduction

"When there's nothing left to burn you have to set yourself on fire." - Stars

In this chapter we present the foundations of this PhD thesis. First, we describe the motivation that conducts our research. Then, we define the objectives and structure of this dissertation. Finally, we enumerate the publications generated during this thesis.

1.1 Motivation

Space weather refers to the physical and environmental conditions in space primarily driven by solar activity, including variations in the solar wind, interplanetary magnetic field and their interactions with Earth's magnetosphere, ionosphere and thermosphere. It encompasses phenomena such as solar flares, geomagnetic storms and energetic particle events, all of which can significantly impact both space-based and ground-based technological systems. These phenomena can disrupt communication networks, navigation systems, satellite operations and power grids. As society becomes increasingly reliant on technology, understanding and predicting space weather events has become essential for mitigating their adverse effects. Among these phenomena, geomagnetic storms are particularly disruptive; they are large-scale disturbances of Earth's magnetosphere caused by enhanced solar wind-magnetosphere coupling, often following coronal mass ejections or high-speed solar wind streams. To monitor and quantify the resulting geomagnetic activity, scientists use indices such as Dst, Kp, or SYM-H, which provide standardized numerical measures of the disturbances in Earth's magnetic field. These indices are crucial for assessing and forecasting the intensity and potential impacts of geomagnetic storms on technological systems.

Among the available geomagnetic indices, this thesis focuses on SYM-H and ASY-H due to their higher temporal resolution and broader operational relevance. Compared to traditional indices such as Dst or Kp, which are calculated every 1 and 3 hours respectively, SYM-H and ASY-H offer data at a 1-minute cadence. This fine-grained resolution allows for the detection of rapid fluctuations in geomagnetic activity that lower-resolution indices may overlook [1], making them especially suitable for real-time forecasting applications. Furthermore, the SYM-H and ASY-H indices are derived from measurements taken

by magnetometers distributed across mid-latitudes, covering areas of high technological dependency, such as Europe or North America. This spatial configuration ensures that the indices provide a representative and reliable proxy for the disturbances experienced in the most densely populated and infrastructure-critical regions of the globe.

The motivation for this research arises from the need to enhance our predictive capabilities for space weather events. Historically, the scientific community has relied on physics-based models to forecast the impact of incoming geomagnetic storms. These models, grounded in heliophysics, have been invaluable in understanding the physical phenomena driving space weather. However, they require considerable amounts of memory and computational time, which can be critical in a real-time operational scenario.

With constant advancements in Machine Learning (ML), new opportunities have emerged to address these challenges. The ever-increasing amount of available data, combined with significant advances in both the ML field and computational capabilities, has made it feasible to use ML approaches to forecast geomagnetic indices. Unlike traditional physics-based models, ML models primarily require large resources during the training process, which is only needed once to build the model. After it has been trained, these models are considerably faster during operation, making them more efficient and scalable.

In recent years, ML-based models have demonstrated significant improvements in forecasting capabilities. These improvements have been driven by the development of advanced neural network architectures and the integration of diverse data sources. Surveys and guidelines have further aided the scientific community in properly applying ML techniques to space weather forecasting, covering applications such as forecasting geomagnetic indices, Coronal Mass Ejection (CME) propagation time, solar wind speed, and flare occurrences.

The evolution of ML has been marked by several key advancements that have made it a promising tool for space weather forecasting. Techniques such as Deep Learning (DL), particularly the use of Deep Neural Networks, have enabled the modeling of complex, nonlinear relationships between input variables and geomagnetic indices. These models can capture the intricate dynamics of the solar-terrestrial environment, leading to accurate and reliable forecasts, while also providing uncertainties about the forecast, which is critical for this scenario.

Furthermore, the computational capabilities required to train and deploy these models have also seen substantial improvements. High Performance Computing (HPC) resources, including Graphical Processing Units (GPUs) and cloud-based platforms, have drastically reduced the time and cost associated with training large-scale ML models. This has facilitated access to powerful computational tools, allowing researchers and organizations to develop sophisticated forecasting models without the need for extensive infrastructure.

The significance of accurately forecasting geomagnetic disturbances is of paramount importance. As society becomes increasingly dependent on technology, the ability to predict

and mitigate the effects of space weather events becomes ever more critical. Geomagnetic storms, in particular, can have severe consequences for a wide array of technological systems. These storms can induce strong Geomagnetic Induced Currents (GICs), which can disrupt power grids, leading to widespread blackouts and significant economic losses. For instance, a geomagnetic storm in 1989 caused a nine-hour province-wide blackout in Canada, affecting millions of people and resulting in substantial economic damage. In addition to power grids, geomagnetic storms can also impact communication systems, aviation and navigation. High-frequency radio communications, which are essential for aviation and maritime operations, can be severely disrupted by space weather events. Global Navigation Satellite System (GNSS) signals, used for navigation and timing, can also be perturbed, affecting everything from everyday Global Positioning System (GPS) services to critical infrastructure. Military systems are not immune to these effects either; historical incidents, such as the sudden detonation of sea mines in North Vietnam in 1972, highlight the potential risks posed by geomagnetic disturbances to national security and defense operations.

A recent example of geomagnetic storm consequences occurred in February 2022. Forty out of forty-nine Starlink satellites launched by SpaceX were impacted by a geomagnetic storm one day after their launch. The storm caused the atmospheric drag to increase up to 50% higher than during previous launches, causing the satellites to reenter the Earth's atmosphere, being destroyed in the process. In May 2024, a superstorm caused widespread auroras visible from lower latitudes, reaching as far south as Texas and Spain. This geomagnetic storm also caused significant disruptions in GPS functionality and other systems reliant on satellite technology, highlighting the broad and potentially severe impacts of space weather on modern infrastructure and navigation.

Given these widespread and potentially severe impacts, the development of reliable forecasting models is essential. By providing accurate and timely predictions of geomagnetic disturbances, we can implement measures to protect critical infrastructure, ensure the continuity of essential services, and mitigate the risks to human safety.

In light of the challenges and opportunities outlined above, using ML to address the problem of geomagnetic indices forecasting presents a promising line of work. Although early efforts in applying ML techniques to this domain began in the 1990s [2], [3], and even earlier attempts using statistical models [4], recent advancements in computational power, neural network architectures, and the availability of high-resolution, continuous space weather data have enabled a significant leap forward in performance. By continuously improving these models and integrating new data sources, we can enhance their accuracy and reliability, ultimately contributing to a more resilient society.

In conclusion, the increasing availability of data, coupled with advancements in ML and computational capabilities, has opened new possibilities for space weather forecasting. As dependence on technology continues to grow, the ability to predict and mitigate the

effects of geomagnetic disturbances becomes increasingly critical. Through this research, we aim to contribute to the development of more accurate and reliable forecasting models, ultimately enhancing our ability to protect and sustain technological systems in the face of space weather challenges.

1.2 Objectives

This PhD has a multi-faceted objective centered on the development, validation, and deployment of a neural network model capable of accurately forecast geomagnetic indices in real-time. Specifically, the focus is on improving the forecasting of the SYM-H and ASY-H geomagnetic indices. These indices are essential for assessing the intensity and impact of geomagnetic storms.

The primary objectives of this dissertation are outlined as follows:

1. Develop and validate a Neural Network-Based Model: Create a robust model for forecasting SYM-H and ASY-H indices using historical space weather data. The aim is to develop a model that can be directly compared to existing state-of-the-art forecasting systems, demonstrating superior accuracy and reliability.

2. Operational Deployment: Implement the developed model in a real-time operational environment by integrating live data from both satellites and ground-based magnetometers. This enables continuous updates and forecasting of geomagnetic indices, delivering timely and actionable insights.

3. Incorporate Prediction Uncertainty: Improve existing models by including the uncertainty associated with the predictions. In the context of space weather forecasting, understanding and quantifying the uncertainty of predictions is critical for decision-making and risk assessment. This involves enhancing the model's capabilities to provide reliable confidence intervals alongside the predictions.

4. Adaptation to Local Indices: Extend the applicability of the model to forecast local geomagnetic indices. This adaptation aims to enhance the model's versatility and provide more detailed predictions tailored to specific geographic locations.

In addition to these primary goals, the research also aims to contribute to the field through the following specific objectives:

5. Revise and Enhance the Training and Evaluation Framework: Conduct a thorough review of existing training and evaluation frameworks used in state-of-the-art forecasting models. Identify areas for improvement, such as incorporating a greater number of geomagnetic storms, including those with provisional data that more closely resemble real-time conditions.

6. Develop Specific Metrics for Geomagnetic Indices Forecasting: Create specialized metrics tailored to the unique challenges of geomagnetic indices forecasting. Traditional time series forecasting metrics may not fully capture the intricacies of this problem, necessitating the development of new evaluation criteria that better reflect the performance and reliability of forecasting models in this context.

By addressing these objectives, this dissertation aims to advance the field of space weather forecasting. The development of a high-performing neural network model, coupled with an enhanced evaluation framework and real-time operational deployment, will provide a robust foundation for future research and practical applications in mitigating the effects of geomagnetic storms on technological systems.

1.3 Structure

This section outlines the chapters of the dissertation:

- **Chapter 1:** Introduction. This chapter presents the motivation, objectives, and structure of the thesis.
- **Chapter 2:** State of the Art. This chapter provides a comprehensive review of the current literature on space weather and ML techniques used in forecasting geomagnetic indices. It covers missions monitoring the Sun, data sources, geomagnetic storms, and existing ML models in space weather.
- **Chapter 3:** Neural Network Architecture. This chapter details the design and development of the Deep Neural Network (DNN) model for forecasting the SYM-H and ASY-H indices. It covers the model architecture, training process, and feature selection.
- **Chapter 4:** Improving the Framework. This chapter focuses on improving the framework used to train and evaluate ML models for forecasting geomagnetic indices in the literature. It discusses the expansion of the dataset and the creation of specific metrics for evaluating the models in the context of geomagnetic indices forecasting.
- **Chapter 5:** Improving the network. This chapter focuses on further improving the network, measuring the impact of the expanded dataset on both traditional and new metrics, and providing a confidence interval alongside the forecasted value.
- **Chapter 6:** Local indices. This chapter extends the work done on global indices to local indices, particularly focusing on the Local Disturbance Index (LDi) of some of the observatories used to calculate the SYM-H.
- **Chapter 7:** Conclusions and Future lines of work. This chapter summarizes the research findings, discusses the implications of the results, and outlines potential directions for future work.

1.4 Publications

Along the development of this work some publications were generated. They are reported below.

Posters and Conferences:

- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Deep neural Networks With Convolutional and LSTM Layers for SYM-H and ASY-H Forecasting. *Poster at ML-Helio 2022*. March 2022.
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Neural Networks for operational SYM-H forecasting Using Attention and SWICS plasma features. *Poster at European Space Weather Week 2023*. November 2023.
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Geomagnetic storm resilience through local disturbances predictions with Machine Learning. *Poster at European Space Weather Week 2024*. November 2024.

Scientific Publications:

- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Neural Networks for operational SYM-H forecasting using attention and SWICS plasma features. *Space Weather*. Vol 21, no. 8, August 2023. doi:10.1029/2023SW003485
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Classifying and bounding geomagnetic storms based on the SYM-H and ASY-H indices. *Natural Hazards*. Vol 120, October 2023. doi:10.1007/s11069-023-06241-1
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. A Framework for Evaluating Geomagnetic Indices Forecasting Models. *Space Weather*. Vol 22, no. 3, March 2024. doi:10.1029/2024SW003868
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Comment on "Prediction of the SYM-H Index Using a Bayesian Deep Learning Method With Uncertainty Quantification" by Abduallah et al. (2024). *Space Weather*. Vol 22, no. 8, August 2024. doi:10.1029/2024SW003909
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Deep Neural Networks With Convolutional and LSTM Layers for SYM-H and ASY-H Forecasting. *Space Weather*. Vol 19, no. 6, June 2021. doi:10.1029/2021SW002748
- Armando Collado-Villaverde, Pablo Muñoz and Consuelo Cid. Operational SYM-H forecasting with confidence intervals using Deep Neural Networks. *Space Weather*. Vol 22, no. 10, October 2024. doi:10.1029/2024SW004039

Chapter 2

State of the art

I hate it when my mom says "get up it's already morning, sun is out". So what am I supposed to do? Photosynthesis?

In this chapter, we present the state of the art and the theoretical knowledge required for the dissertation.

The main objective of this PhD is to develop a geomagnetic indices forecasting system based on ML techniques. The geomagnetic indices are a measure of geomagnetic activity; they represent the disturbance of the Earth's magnetosphere caused by charged particles emitted by the Sun, namely, the solar wind. The solar wind is measured by space probes, usually located between the Sun and Earth, which primarily measure the Interplanetary Magnetic Field (IMF) and plasma features (proton density, speed and temperature).

A reliable Space Weather (SW) forecasting system is becoming increasingly important due to the severe disruptions that intense geomagnetic storms could have on human activity, both in space and on Earth. It can severely impact technological and electrical systems, which are of paramount importance in today's society. For example, telecommunications systems, aviation, space exploration, and power grids are susceptible to solar events, requiring actions to prevent service disruptions or mitigate risks to human safety.

Initially, the scientific community developed physics-based models that relied on heliophysics to forecast the impact of incoming geomagnetic storms. These models, although conceptually robust and physically interpretable, posed considerable challenges due to the complexity and computational demands of simulating the coupled Sun-Earth system. In parallel, the development of machine learning approaches also began, but early models lacked the data availability and computational resources required to match the accuracy and reliability of physics-based methods. Only in recent years, thanks to significant advances in ML, particularly in Artificial Neural Networks (ANNs), and the availability of large, high-resolution datasets, have ML based models begun to demonstrate comparable or even superior forecasting capabilities. As a result, part of the scientific community has

increasingly turned to ML while continuing to refine traditional models, especially given ML's success in forecasting irregular magnetospheric and ionospheric processes that were difficult to model otherwise.

Nevertheless, ML-based models have some key drawbacks, the primary one being the *black box* nature of ANN-based models; there is no information regarding why these models produce their predictions, whereas physics-based models provide explanations of the physical phenomena behind their outputs. On the one hand, physics-based models require considerable amounts of memory and computational time, which can be quite high depending on the problem. On the other hand, ML-based models require significant resources primarily during the training process (which is only needed once to build and train the model) but are considerably faster during operation. Nevertheless, the scientific community has developed several models that have considerably improved the forecasting capabilities over the previous physics based models. The development of ML-based models has been aided by surveys and guidelines on properly applying ML techniques to the SW context, including forecasting geomagnetic indices, CME propagation time, solar wind speed, and flare occurrences.

This chapter provides an overview of SW, explaining what it is and why it is important. The missions that monitor the Sun are outlined and we introduce the data to be used in the dissertation. Section 2.3 briefly introduces ML, summarizing the most important points regarding ML and ANNs. Next, we review the state of the art for the geomagnetic indices forecasting. Finally, the chapter concludes with some key points.

2.1 Introduction to Space Weather

SW is a branch of space physics and heliophysics. It is defined as the physical conditions of the Sun, the interplanetary medium, the geomagnetic field, Earth's magnetosphere, and Earth's surface, all influenced by the Sun-Earth interaction. In particular, SW describes the phenomena that impact systems and technologies in orbit and on Earth.

The Sun constantly emits a stream of charged particles into space, known as the solar wind, which interacts with and disturbs Earth's magnetosphere. Most of the time, the Sun is in a "quiet" period, so the solar wind has a negligible impact on the magnetosphere. However, during "active" periods, the Sun often produces CMEs: explosive events where large amounts of ionized plasma are expelled at high speeds into space.

The impact of a CME on Earth's magnetosphere causes a geomagnetic storm: a compression of the magnetosphere among other phenomena followed by strong GICs, typically at high latitudes, producing beautiful effects such as the Aurora Borealis. While this rarely affects wildlife, it can significantly impact human activity, disrupting a wide variety of technological and electrical systems. In this regard, SW has gained significant relevance due to the great dependence of society on technology and electricity. Therefore, counter-

measures are required to minimize or even nullify such disturbances, as well as to avoid risks to human lives.

Geomagnetic storms can cause a myriad of different, severe, consequences. Especially in any technological system that relies on satellites, since they are less protected by the magnetosphere. One of the systems that can be heavily disrupted is GNSSs, whose signals can be perturbed by SW effects [5], causing serious problems for aviation and navigation systems, which rely on the geomagnetic field for orientation. Additionally, the consequences of geomagnetic storms can also be observed in ground-based technological systems, such as telecommunications and power grids [6], which can be damaged by GICs, leading to blackouts. The most notable GIC event caused a nine-hour province-wide blackout in Canada in 1989 due to a failure in the Hydro-Québec power system, with significant economic and social impact [7]. Later on, in October 2003, another GIC caused by a geomagnetic storm provoked a blackout in southern Sweden [8]. Moreover, GICs can also cause corrosion problems in pipelines as well as disturb any corrosion inspection [9], [10]. A solar event also caused several issues during the construction of an oil pipeline across Alaska in 1970 [11]. It has also been reported to pose a considerable risk to undersea cables, harboring the potential to damage the Internet infrastructure [12].

Geomagnetic storms also have a harsh impact on military systems, such as the sudden detonation of sea mines, as happened in North Vietnam in August 1972, when dozens of sea mines exploded due to the magnetic perturbations caused by a CME [13]. In addition, CMEs can disrupt radar activity, compromising any system that relies on radar for guidance [14].

Another recent example of the consequences of a geomagnetic storm happened in February 2022 [15]. 40 out of the 49 Starlink satellites launched by the north-American company SpaceX were impacted by a geomagnetic storm one day after they were launched. The storm caused the atmospheric drag to increase up to 50% higher than during previous launches, causing the satellites to reenter the Earth's atmosphere, being destroyed in the process.

Historically, the largest recorded geomagnetic storm is known as the Carrington Event [16]. It happened in September 1859 and its effects were seen all around the world in the form of auroras. It also caused failures in the telegraph system in Europe and North America; telegraph operators even received electric shocks. Additionally, multiple fires were caused by induced electrical currents at low altitudes. Nowadays, it has been estimated that the economic impact of another Carrington-like event could cause catastrophic damages that could take up to 10 years to recover from, estimating a cost of \$0.6 to \$2.6 trillion in the United States [17]. Geomagnetic storms have been classified as another natural hazard by some countries [18].

Despite the estimations mentioned, the intensity and effect of geomagnetic storms and SW processes are, in general, very complex to quantify. As such, the SW community has

developed different indices that are used to numerically describe the effect of the storms, each of them being focused on different parts of the magnetosphere.

In general, geomagnetic indices quantify the disturbance of the Earth's magnetosphere caused by the solar wind at different latitudes, condensing the status of the magnetosphere in a single number. They play a key role in describing how the magnetosphere has changed, and they have become the focus of SW studies since the 1950s. They can monitor the evolution of complex phenomena while at the same time offering reliable and concentrated information.

2.1.1 Missions monitoring the Sun

In addition to the information gathered by Earth-based magnetometers, solar wind data is also crucial for predicting geomagnetic activity. As such, there are several space missions that are observing the Sun and providing the data needed to develop any kind of SW system. Among them we can highlight:

1. Advanced Composition Explorer (ACE): A National Aeronautics and Space Administration (NASA) space probe launched in August 1997. It started its operations on January, 1998. It provides continuous SW reports and warnings of geomagnetic storms that can potentially reach Earth, disrupt communications, and harm astronauts in space. It orbits the L1 Lagrange Point and carries nine scientific instruments:

    - Magnetometer (MAG): It continuously measures the local magnetic field in the interplanetary medium, providing measurements with a one-second resolution.
    - Real-Time Solar Wind (RTSW): It continuously monitors the solar wind, producing warnings of incoming major geomagnetic activity, up to one hour in advance.
    - Electron, Proton, and Alpha-particle Monitor (EPAM): It measures a wide variety of energetic particles around it at a high time resolution. It measures from 50 keV to 5 MeV for ions, and 40 keV to about 350 keV for electrons, which are essential to understand the dynamics of solar flares.
    - Cosmic-Ray Isotope Spectrometer (CRIS): It studies and determines the isotopic composition of cosmic rays in an attempt to identify their origin.
    - Solar Wind Ion Mass Spectrometer (SWIMS) and Solar Wind Ion Composition Spectrometer (SWICS): These two instruments are mass spectrometers, each one intended for different measurements. They analyze the chemical and isotopic composition of solar wind and interstellar matter.
    - Solar Wind Electron, Proton, and Alpha Monitor (SWEPAM): It provides the bulk solar wind observations. It measures the solar wind plasma electron and ion fluxes. This information provides context for elemental and isotopic composition

measurements made by other instruments, as well as providing direct information of solar wind phenomena such as CMEs. SWEPAM also provides real-time solar wind observations, which are continuously delivered to the ground for SW studies and applications.

- Ultra-Low Energy Isotope Spectrometer (ULEIS): This instrument measures the flux of ions in the helium to nickel range. This is used to determine the characteristics of solar energetic particles and the mechanism by which they are charged by the Sun.
- Solar Isotope Spectrometer (SIS): It provides high-resolution measurements of the isotopic composition of energetic nuclei from helium to zinc over the energy range from ≈10 to ≈100 MeV/nucleon. During large solar events, SIS measures the isotopic abundances of solar energetic particles to directly determine the composition of the solar corona, allowing the study of particle acceleration processes.

2. Wind: Is a NASA space probe launched on November 1994, whose main objective is to observe the solar wind's plasma and radio waves. The original objective of the spacecraft was to orbit the Sun at the L1 Lagrange Point, similar to other spacecraft. However, the mission was delayed to study the magnetosphere near the lunar environment until it finally reached the L1 point in 2004. The scientific objective was to understand the behavior of the solar-terrestrial plasma environment to predict how Earth's atmosphere would respond to changes during intense solar wind conditions. The instruments on-board Wind are:

   - Magnetic Field Investigation (MFI): Consists of two magnetometers used to study large-scale structures and fluctuations in interplanetary magnetic fields. They have several measurement ranges with a high resolution.
   - Solar Wind Experiment (SWE): The instrument was designed to measure solar wind thermal protons and positive ions. It consists of a six-axis spectrometer that provides three-dimensional velocity distribution functions for ions and electrons with high temporal resolution.
   - Wind 3D Plasma Analyzer: Measures the three-dimensional distribution of plasma and energetic electrons and ions with high temporal, angular and energy resolution in the range from 10 eV to 5 MeV.
   - Radio and Plasma Wave Experiment (WAVES): Is targeted towards radio and plasma waves analysis. It measures the intensity and arrival direction of radio and plasma waves, which are originated in the near-Earth solar wind.
   - SWICS and Suprathermal Ion Composition Spectrometer (STICS): These instruments measure the ionic composition and electric charge of the solar wind, the velocity, density and temperature of He_4 ions, the average proton velocity of the solar wind and the energy distribution of some ionic species.

- Energetic Particle Acceleration, Composition and Transport (EPACT): Measures energetic particle acceleration and transport processes in solar flares, the interplanetary medium, the magnetosphere and cosmic rays.

3. Deep Space Climate Observatory (DSCOVR): Is a joint mission between the NASA, the National Oceanic and Atmospheric Administration (NOAA) and the United States Air Force (USAF). It is a SW station that monitors the solar wind, providing alerts and forecasts for geomagnetic storms that could be hazardous to the Earth's power grids, satellites, telecommunications, aviation and GPS. It orbits around the L1 Lagrange Point. It has three scientific instruments:

   - Plasma-Magnetometer (PlasMag): Consists of three instruments: A magnetometer, which measures the interplanetary magnetic field, a Faraday cup, which measures positively charged particles, and an electrostatic analyzer, to measure electrons. In general, it measures the solar wind for SW predictions.
   - Earth Polychromatic Imaging Camera (EPIC): Camera captures images of the sunlit side of Earth for various Earth science monitoring purposes in ten channels, ranging from ultraviolet to near-infrared. Ozone and aerosol levels are monitored along with cloud dynamics, properties of the land and vegetation.
   - National Institute of Standards and Technology Advanced Radiometer (NISTAR): Measures irradiance of the sunlit face of the Earth. That is, if the atmosphere of Earth is taking in more or less solar energy than it is radiating back towards the space.

4. Solar and Heliospheric Observatory (SOHO): It is a collaboration between the European Space Agency (ESA) and NASA to study the Sun, launched on December, 1995. It carries 12 scientific instruments and orbits at the L1 Lagrange Point. Its main scientific objectives are focused on researching the outer layer of the Sun, as well as studying the internal structure of the Sun. Additionally, it is also the main source of near-real-time solar data for SW predictions, along with the previously mentioned probes Wind, ACE, and DSCOVR. Its main instruments are:

   - Large Angle and Spectrometric Coronagraph (LASCO): Is one of the most important instruments of SOHOİt captures images of the corona and consists of three solar coronagraphs with nested fields of view. Two of them produce images of the corona over much of the visible spectrum, while the remaining one produces images of the corona in a number of very narrow visible wavelength bands.
   - Coronal Diagnostic Spectrometer (CDS): Measures density, temperature and plasma flows on the transition region and low corona.
   - Charge, Element, Isotope Analysis (CELIAS): Detects the solar wind as it passes SOHO, analyzing the density and nature of the charged particles. It provides a

brief warning of gusts in the solar wind, which arrive at SOHO 30–60 minutes before they reach the Earth.

- Suprathermal & Energetic Particle Analyser (COSTEP) and Energetic Particle Analyser (ERNE): Study the ion and electron composition of the solar wind.
- Extreme UV Imaging Telescope (EIT): Studies both the low coronal structure and its activity. It is sensitive to light of four different wavelengths, corresponding to light produced by highly ionized iron (XI)/(X), (XII), (XV) and helium (II).
- Global Oscillations at Low Frecuence (GOLF): Studies the internal structure of the Sun. It measures the spectrum of global oscillations in the frequency range 10^{-7} to 10^{-2} Hz.
- Michelson Doppler Imager (MDI): Is a helioseismology instrument. It measures the velocity and magnetic fields in the photosphere to learn about the convection zone. It also provides information about the magnetic fields which control the structure of the corona.
- UltraViolet Coronagraph Spectrometer (UVCS): Makes measurements of the solar corona between 2 and 10 solar radii from the Sun center with high spectral and spatial resolution. Its purpose is to provide a detailed description of the extended solar corona. That information can be used to address a broad range of scientific questions regarding the nature of the solar corona and the generation of the solar wind.

5. Solar Dynamics Observatory (SDO): Is a NASA mission that has been observing the Sun since 2010. Its main objective is to understand the influence of the Sun on the Earth. To do that, it studies the solar atmosphere on small scales of space and time and in many wavelengths simultaneously. SDO provides information about how the Sun's magnetic field is generated and structured. It also gives insight regarding how the magnetic energy is released into the space in the form of energetic particles, as well as monitoring variations in the solar irradiance. It has three main instruments:

   - Helioseismic and Magnetic Imager (HMI): Studies the solar variability and characterizes the Sun's interior and the various sources of magnetic activity. It takes high-resolution measurements of the longitudinal and vectorial magnetic field over the entire visible solar disk, as shown in Figure 2.1. It extends the capabilities of SOHO's helioseismology instrument, the MDI.
   - Extreme Ultraviolet Variability Experiment (EVE): Measures the Sun's extreme ultraviolet irradiance. It incorporates physics-based models in order to enable further scientific understanding of the relationship between solar extreme ultraviolet variations and magnetic variation changes in the Sun.

- Atmospheric Imaging Assembly (AIA): Provides an image of the solar disk in the various ultraviolet and extreme ultraviolet bands, with high temporal and spatial resolution. An example of such image can be seen in Figure 2.2.

6. Interface Region Imaging Spectrograph (IRIS): Is a NASA spacecraft whose primary goal is to understand how heat and energy move through the lower levels of the solar atmosphere. It is considered as a Small Explorer (less than $120 million). In general, IRIS is intended to advance Sun-Earth connection studies by tracing the flow of energy and plasma into the corona and heliosphere for which no suitable observations exist. It has an ultraviolet telescope combined with an imaging spectrograph:

   - IRIS's ultraviolet telescope's primary mirror can only see about one percent of the Sun at a time. It has enough resolution for the images to show features as small as 240 km on the Sun. The images record observations of materials at temperatures from 5,000 Kelvin to 65,000 Kelvin, in order to observe material traveling on the surface of the Sun. It captures a new image every five to ten seconds.
   - The spectrograph observes materials at temperatures from 5,000 Kelvin to 10 million Kelvin. This provides information on exactly how much light is visible from any specific wavelength, which represents how much material is present at specific velocities, temperatures and densities. It takes a new image about every one to two seconds.

7. Hinode: Is a collaboration between the Japan Aerospace Exploration Agency (JAXA), the United States and the United Kingdom. The main objective of the mission was to explore the magnetic fields of the Sun, studying the explosive solar activity that can interfere with satellite communications and electric power transmission grids. It has three scientific instruments on-board:

   - Solar Optical Telescope (SOT): Is focused on measuring small changes in the Sun's magnetic field, including how these changes match with dynamic events seen in the Sun's corona.
   - X-ray Telescope (XRT): Captures X-ray images of the Sun's corona, which is the spawning ground for the solar flares and CMEs that dominate the space between the Sun and Earth. The information obtained with this instrument allows the study of how changes in the Sun's magnetic field trigger these explosive solar events.
   - Extreme-Ultraviolet Imaging Spectrometer (EIS): The main objective of the instrument is to identify the processes responsible for the corona heating. It obtains spatially resolved spectra in two wavelength bands.

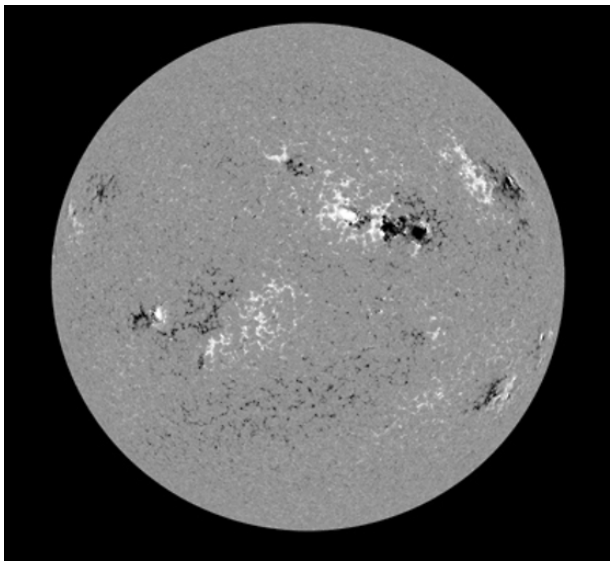


Figure 2.1: Magnetic field on the Sun's surface. Darker areas represent the magnetic field lines pointing away from the Earth, whereas the white areas show the magnetic field lines pointing towards the Earth. Extracted from NASA's website.

2.1.2 Data sources

All the measured data gathered from the previous missions are publicly accessible from several databases in different forms: raw measurements without processing (known as Level 1 data) and science-ready data (Level 2 and Level 3). In some cases, data-bases also include data measured by ground magnetometers and geomagnetic indices. The most relevant data-bases for SW studies are:

- Operating Missions as a Node on the Internet (OMNI)Web [19]: Is a database comprising the data measured by several space probes that can be used for solar wind studies and are relevant to any heliospheric task. An example of data hosted in the database includes L1 solar wind IMF and plasma data, energetic proton fluxes (ranging from 1 to >60 MeV), and the corresponding geomagnetic and solar activity indices.

  It offers data in multiple formats:

  1. The Low Resolution OMNI (LRO) dataset is primarily a 1963-to-current compilation of hourly-averaged, near-Earth solar wind magnetic field and plasma parameter data from several spacecraft in geocentric or L1 orbits. Mainly the International Sun-Earth Explorer-3 (ISEE-3), Wind and ACE spacecrafts. The data has been extensively cross-compared and, for some spacecraft and parameters, cross-normalized. It also contains several datasets related to heliophysics, such as sunspot numbers and geomagnetic activity indices: *Kp* (3 hours), *AE* (1 hour), *Dst* (1 hour) among others. When more than one spacecraft measures the solar wind, the sources are prioritized to maintain consistency and ensure

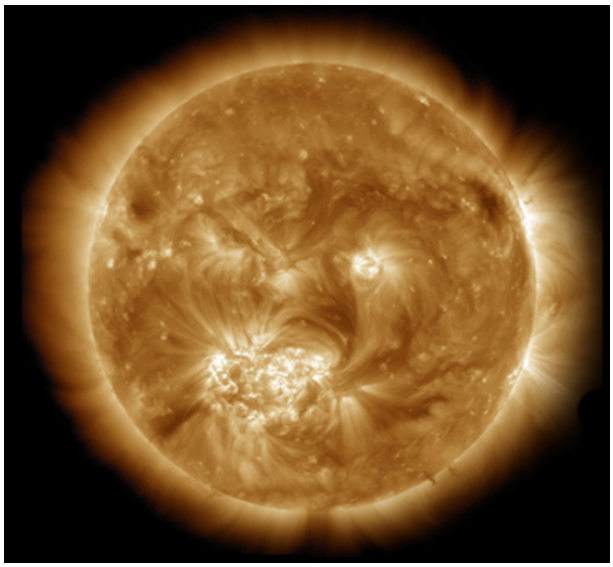


Figure 2.2: Image of the Sun captured by AIA at 193 Å, which is the iron (XII) at 1 million Kelvin and iron (XXIV) at 20 million Kelvin. Extracted from NASA's website.

maximum data quality. For example, Wind was prioritized over ACE through the end of 1999, then, ACE was prioritized and used as the primary source.

2. The High Resolution OMNI (HRO) dataset consists of 1-min and 5-min solar wind datasets at the Earth's bow shock nose. It contains IMF and plasma data from the ACE, Wind, Interplanetary Monitoring Platform-8 (IMP-8), Geotail, Geostationary Operational Environmental Satellite (GOES), and DSCOVR spacecrafts. Studies regarding the effects of solar wind variations on the magnetosphere and ionosphere are the main targets of this dataset. However, it is important to note that the solar wind measurements in the HRO dataset are post-processed. The data has undergone several correction processes, such as removing incorrect measurements and imputing data from other spacecraft if one fails.

3. Spacecraft specific datasets: There is also access to data measured by specific spacecraft, in their native resolutions. For example, the 16-second IMF data from ACE or 92-second plasma data from WIND.

- ACE Science Center (ASC). This database provides the data from the ACE spacecraft. It provides data at multiple levels: Level 1, Level 2, and Level 3. The first level consists of raw telemetry data sorted by instrument and formatted into Hierarchical Data Format (HDF) files. The second level is already preprocessed and corrected data, suitable for scientific studies without further processing. Finally, the level 3 offers post-processed data, plots, and lists. They are usually provided by members of the ACE team and others, in the hope that they will be useful to the community. The data can be downloaded on a per-instrument basis, with MAG and SWEPAM being the most relevant for this project.

- DSCOVR database. Similar to the ASC database, it provides the data from this specific spacecraft, being the faraday cup and the magnetometer the most relevant instruments for our work. However, since DSCOVR was launched in 2015, it has not yet recorded enough intense geomagnetic storms to be used to train ML systems.

Nevertheless, the available data is not complete. Due to instrument malfunctions or saturation during intense storms, some critical data is missing. Figure 2.3 shows the coverage of the solar wind parameters measured by ACE from its launch until the end of 2017. The IMF data has nearly 100% availability, and the small gaps can be filled using linear interpolation due to the high resolution of the magnetometer. Notwithstanding, the plasma variables present large data gaps starting from 2008, due to the aging of the instruments and saturation during intense geomagnetic storms. This was later on palliated thanks to an operational improvement, as informed by the SWEPAM instrument team [20]. Although the overall amount of missing data is not excessive, it is misleading because most data gaps occur during intense solar events, when forecasting geomagnetic conditions is critical.

As the data can contain missing or incorrect values, post-processing tasks are required. This can be done with a reprocessing toolchain on historical data, but it is not feasible in a real-time environment. Therefore, additional development will be needed to deal with such issue in this project.

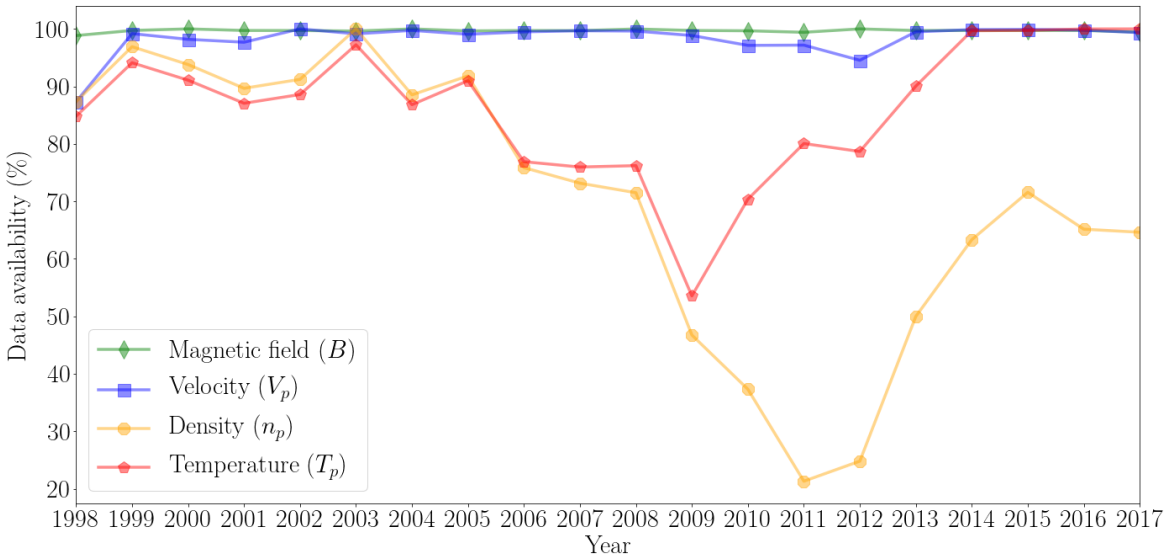


Figure 2.3: Data availability of the main solar wind parameters measured by the ACE spacecraft from the time it is operational until the end of 2017. Extracted from Larrodera and Cid [20].

2.2 Regional and global indices

When the solar wind, mainly the plasma and the IMF interact with the Earth's magnetosphere, a great amount of energy and particles are transferred inside the magnetosphere. The solar wind is highly variable, especially during geomagnetic storms, and it directly influences the magnetosphere, as the shape and size of the magnetosphere, along with the amount of energy transferred and dissipated, are closely related to the solar wind. All those processes are complex and difficult to quantify. Therefore, to provide a feasible measure of the magnetosphere disturbance, the geomagnetic indices are used.

In general, indices are composite statistics. They are widely used in different domains to measure and describe the evolution of a group of individual data points. As such, geomagnetic indices numerically describe the geomagnetic activity or some of its components, that is, the response of the Earth's magnetosphere and ionosphere to the solar wind. There are several geomagnetic indices that focus on different phenomena in the magnetosphere, ionosphere and deep in the Earth in their own way. The measured phenomena dictate the location of the measurement, its timing, resolution, and the method by which the index is calculated. They have become a key parameter in Solar-Terrestrial studies.

The first attempt to outline geomagnetic activity started in 1885. This first approach was focused on estimating the geomagnetic disturbances that happened on a daily basis. Every day, the daily range, that is to say, the maximum difference between the recorded values on a given geomagnetic component, was calculated at the Greenwich observatory using the two horizontal magnetic components. H for the magnetic North and D for the magnetic East. This index was known as *"C"* and it has been calculated for more than 90 years, from 1884 to 1975. It is the first geomagnetic index defined at an international level; it judges the relative degree of disturbance of several magnetograms (quiet, moderately disturbed, disturbed), forming an international network. In order to calculate this index, each observer measures its disturbance, creating their own local C index (as in C0, C1, ...). Then the mean of the cooperating observatories is computed, resulting in the C_i index, as International C. However, due to the roughness and the potential to overlook irregularities by calculating the mean over a whole day, new indices that could allow more precise monitoring of the milder variations were introduced.

The next index, known as K, was introduced by Bartels et al. [21] in 1939, which later on, was combined to compute the Kp index, p as in planetary. It provides 3-hour quasi-logarithmic measures, calculated using the maximum variation of the H magnetic component, obtained from 13 observatories of sub-aural latitudes. In each observatory, a K index is calculated, then the Kp index is computed after standardizing each K index according to its latitude. This index is of great importance in the evolution of geomagnetism studies, but it is still not perfect, mainly due to the limited stations that were

available to provide the measurements at that time, the World War II, followed by the Cold War.

After the International Geophysical Year (1957), scientific exchange between the East and West resumed. Therefore, it became possible to develop new, more accurate indices thanks to the improvement of the existing observatory networks. The main objective for those indices was to describe the status of the magnetosphere in a more refined way, leading to a better understanding of the physics that governs the magnetosphere. The most relevant indices are summarized next:

- *Dst* index, introduced by Sugiura in 1964 [22]. This index reflects the intensity of the ring current, specifically, it describes the total amount of energy of the ring particles injected into the internal magnetosphere. That energy is proportional to the intensity of the storm. This hourly index is based on the H component measured from 4 low/medium latitude observatories distributed in longitude.

- *AE* index, introduced in 1966 by Davis and Sugiura [23]. This index is derived from 12 high latitude observatories. It provides a global, quantitative measure of the auroral zone magnetic activity. One of the main advantages of this index is that it can be derived on an instantaneous basis, or from averages of variations computed over any selected interval. It has been successfully employed both qualitatively and quantitatively as a correlative index in studies of substorms' morphology. In this context, a substorm refers to a time-dependent process in the Earth's magnetosphere that results in a sudden intensification of auroral activity, typically accompanied by rapid variations in the magnetic field observed at high latitudes.

- *am* index: Introduced by Mayaud in 1968 [24], the *am* index is a global geomagnetic index derived from the local K indices. It is computed using a worldwide network of observatories located near 50° geomagnetic latitude, within the sub-auroral zones. These stations are grouped into sectors based on longitude, five in the Northern Hemisphere and four in the Southern Hemisphere, to ensure longitudinal coverage. Within each group, the K indices are converted to amplitudes in nanotesla (nT), and weighted averages are calculated separately for the northern and southern hemispheres, producing the *an* and *as* indices, respectively.

- *aa* index: Introduced by Mayaud in 1972 [25], the *aa* index is another planetary geomagnetic index derived from the K index, designed to measure the amplitude of global geomagnetic activity in 3-hour intervals. It was specifically developed to provide the longest possible continuous record of geomagnetic activity, and it has been calculated from 1868 to the present day. In 1975, the International Association of Geomagnetism and Aeronomy (IAGA) officially recommended the *aa* index as a replacement for the earlier C_i index due to its improved consistency and reliability.

- *PC* index: The Polar Cap (PC) index was originally proposed by Troshichev et al. in 1979 [26] and further developed by Troshichev and Andresen in 1985 [27]. It monitors geomagnetic activity over the polar caps and is particularly sensitive to changes in the IMF, especially the B_z component, and the solar wind. The index provides valuable insight into polar cap convection and magnetic disturbances linked to solar wind-magnetosphere coupling.

The previous indices were designed when observatories could only operate using analogue magnetometers, greatly limiting their capabilities. However, as new technology was developed, digital magnetometers were installed in most observatories.

Thus, new indices that could make better use of the high-quality digital data were proposed, offering new insight into geomagnetic activity. From a variety of the recently developed indices, this work is focused on the *SYM* and *ASY* indices, which offer the longitudinally asymmetric and symmetric *H* component disturbances on a 1-minute interval, being available from 1981 onwards. These indices are calculated using the measurements from 6 observatories at mid-latitudes ($\pm 20 - 50^o$) evenly distributed in longitude. The main purpose of these indices is the description of the geomagnetic disturbances at mid-latitudes, in terms of longitudinally asymmetric (ASY) and symmetric (SYM) disturbances for both H and D components. This distinction is made because the disturbance in mid and low latitudes is not axially symmetric. In some cases, during the development phase of a geomagnetic storm, the asymmetric disturbance field can be even greater than the symmetric counterpart [28], [29].

The SYM-H has a similar elaboration procedure as the *Dst* index; thus, it is considered as a high-resolution *Dst* index, although there are minor differences in the coordinate system that each index uses. The main advantage of the SYM-H over the *Dst* index is the more trustworthy reflection of rapid variation of the solar wind parameters, such as the southward component of the IMF or the dynamic pressure. Additionally, the effect of substorms are better reflected on the SYM-H index thanks to its higher resolution.

However, the *Dst* index remains highly relevant in the field, used alongside SYM-H and ASY-H to characterize geomagnetic activity. Several works have established thresholds differentiating the geomagnetic activity, such as the ones in the work of Palacios et al. [30]. They classify the disturbance into four classes: Quiet-minor, Moderate, Intense and Superintense, according to the *Dst*, SYM-H and ASY-H index values, as shown in Table 2.1.

2.2.1 IAGA endorsed indices

The IAGA is a non-governmental, international scientific association that focuses on the study of terrestrial and planetary magnetism. They are responsible for developing and maintaining the International Geomagnetic Reference Field (IGRF), which is a standard

Table 2.1: Thresholds of geomagnetic activity for different indices. Dst thresholds are extracted from Palacios et al. [30], SYM-H and ASY-H thresholds from are extracted from Dremukhina et al. [31].

Index	Quiet-Minor	Moderate	Intense	Superintense
Dst [nT]	(∞, -50)	(-50, -100)	(-100, -250)	(-250, -∞)
SYM-H [nT]	(∞, -75)	(-75, -150)	(-150, -330)	(-330 , -∞)
ASY-H [nT]	(-∞, 20)	(20, 100)	(100, 400)	(400, ∞)

mathematical description of the large-scale structure of the Earth's main magnetic field and its secular variation. As such, the IAGA has endorsed several indices.

- ***Kp* index**: Introduced by Bartels on 1949 [32]. The *p* stands for planetary. This index is crucial for geomagnetic studies; it provides a geomagnetic activity measure on a planetary scale in 3-hour intervals using a quasi-logarithmic scale as a third of K units, and it has been computed since 1932. However, this index is not perfect, mainly because its values are heavily skewed toward Europe and Northern America due to the observatories used to calculate the index at that time (during the cold war).

- ***Dst* index**: This index is highly significant; it represents the axially symmetric disturbance magnetic field at the dipole equator on Earth's surface, caused by the ring current. It provides the disturbance in nT units with 1 hour intervals and it has been calculated from 1957 onwards. Major disturbances in the Dst index are measured as negative values, as the disturbance is southward. Given its importance, the index is provided in three classes by the World Data Center for Geomagnetism, Kyoto (WDCG): Quick Look Dst, Provisional Dst, and Final Dst, depending on the stage of data processing provided by the observatories.

- ***AE* index:** It monitors the magnetic signature of the eastward and westward auroral electrojets in the Northern Hemisphere at 1-minute intervals. Its value is derived from the geomagnetic variations in the horizontal component H using the measurements from 12 observatories in the Northern auroral zone. The AE index is calculated as the difference between the minimum (AU) and maximum (AL) H deviations.

- ***am* index:** This index is particularly useful for assessing planetary-scale geomagnetic conditions, including hemispheric differences and variations across all longitudes. The *am* index is computed as the average of *an* and *as*, providing a balanced measure of global geomagnetic activity. It is calculated at 3-hour intervals, like the original K index, but offers improved spatial representation and sensitivity by incorporating a more globally distributed set of stations.

- ***aa* index:** The index is based on geomagnetic measurements from two nearly antipodal observatories, located in England and Australia, normalized to a geomagnetic latitude of approximately $\pm 50^{o}$. For each 3-hour interval, K indices are recorded,

standardized, and converted into amplitude values using mid-class amplitudes. These are then averaged with specific weighting factors to account for hemispheric differences in geomagnetic disturbance levels, resulting in a global index suitable for long-term studies of geomagnetic activity and solar variability.

- ***PC* index:** The PC index is derived from geomagnetic field measurements at two polar cap stations—one in the Northern Hemisphere and one in the Southern Hemisphere. It is computed from deviations in the horizontal components of the magnetic field: H (northward) and D (eastward), relative to a geomagnetically quiet reference level. The index is available at a 15-minute resolution and was officially adopted in 2013, becoming a useful tool for monitoring high-latitude geomagnetic activity under varying solar wind conditions.

2.2.2 Non endorsed indices

Aside from those indices, there are other indices that are not officially endorsed by IAGA but are already widely used and hold substantial importance:

- $a\sigma$ **index:** this index is also a K derived index. It provides a characterization of local geomagnetic activity in 4 longitudes sectors and possible hemispheric discrepancies and it uses the same network of observatories as the am index as a base. This index has not been endorsed by IAGA, as it is primarily used for academic research.
- **Ap and ap indices:** Both indices are derived and closely related to the Kp index. The ap index is a planetary 3-hour range index that quantifies global geomagnetic activity on a linear scale, derived from the quasi-logarithmic Kp index. Because of the non-linear relationship of the K-scale to magnetometer fluctuations, it is not meaningful to take the average of a set of K-indices. Instead, every 3-hour the K-value will be converted back into a linear scale, being the ap index. The average from 8 daily ap is the Ap index of a certain day. It is expressed in units of nT and represents the average disturbance level across the same observatories as the Kp index ones. Although both indices are widely used and considered valuable for long-term and climatological studies, they are not officially endorsed by IAGA.
- ***SYM* / *ASY* indices:** These two indices were created to take advantage of the availability of high quality digital data, giving new insight into geomagnetic activity. They were introduced in 2010 [33] and are produced by the WDCG. Both indices compute their values in 1 minute resolution and are computed using 6 observatories evenly distributed in longitude. They describe the geomagnetic disturbances at mid-latitudes in terms of longitudinally asymmetric (ASY) and symmetric (SYM) disturbances for both H and D components, respectively parallel and perpendicular to the dipole axis. SYM-H is essentially the same as the Dst index with a different time resolution.

- **Hp30 and Hp60 indices:** Developed by the GFZ German Research Centre for Geosciences, these indices are advanced versions of the global Kp index, designed to measure geomagnetic disturbances on a planetary scale with higher temporal resolutions of 30 minutes (Hp30) and 60 minutes (Hp60). First created in 2022 [34], these indices utilize data from a network of ground-based magnetometers located at approximately 13 observatories worldwide, including sites in Europe, North America, and Australia. The Hp30 and Hp60 indices provide more detailed and frequent updates on geomagnetic activity, allowing for improved monitoring and forecasting of SW events compared to the information offered by the Kp index. They are particularly useful in real-time applications, enabling better preparedness and response strategies for mitigating the impacts of geomagnetic storms on technological systems. By offering a more granular view of geomagnetic disturbances, these indices enhance our ability to understand and predict SW phenomena, contributing to the overall resilience of critical infrastructure and services.

- **Local Disturbance Index (LDi):** A geomagnetic index specifically designed to quantify localized geomagnetic disturbances. Developed by Cid et al. [35], it captures geomagnetic variations at individual observatories by subtracting the solar quiet (Sq) variation and a baseline from the horizontal component (H) of the geomagnetic field. This method isolates disturbances locally generated at mid-latitude stations, providing a clearer representation of geomagnetic conditions. The LDi is particularly useful for understanding regional geomagnetic activity and offers more precise monitoring of local disturbances compared to global indices, making it valuable for both academic research and operational space weather forecasting.

Table 2.2 compares the key characteristics of the most relevant geomagnetic indices. From all of them, we have chosen to focus on the ASY and SYM indices for the global indices. They are both modern indices, with a time resolution of 1 minute, which makes them ideal for a ML forecasting system. Additionally, the measured geomagnetic area encompasses most of the developed countries (where the largest electrical infrastructures and technological factories are condensed) making it even more appealing as a forecasting target.

Table 2.2: Comparison of the different geomagnetic indices.

Section	Index	Observatories	Type	Resolution	Availability	Distributor	IAGA Endorsed
Global	aa	2 ±50º latitude antipodal magnetic observatories in UK and AUS	K-derived planetary	3 hour	1868 onwards	EOST, Strasbourg, France	IAGA Bulletin 37, 1975, p. 128, resolution 3
Global	am	5 observatories in +50º, 4 in -50º, in sub-aural zones	K-derived planetary	3 hour	1959 onwards	EOST, Strasbourg, France	IAGA Bulletin 27, 1969, p. 123, resolution 2
Global	Kp	11 northern and 2 southern stations between ±44–60º	K-derived planetary	3 hour	1932 onwards	GFZ Potsdam, Germany	IATME Bulletin 14, 1954, p. 368, resolution 6 and p.229
Equatorial	Dst	4 low latitude (SA, 2x USA, Japan)	Equatorial index horizontal component disturbances	1 hour	1957 onwards	WDC for Geomagnetism - Kyoto, Japan	IAGA Bulletin 27, 1969, p. 123, resolution 2
Polar	PC	2 Polar cap stations (Greenland, Antartica)	Polar Cap index horizontal component disturbances	1 minute	1975 onwards	*PCN*, Technical University of Denmark. *PCS*, Arctic and Antarctic Research Institute	IAGA 12th Scientific Assembly Resolution No. 3 (2013)
Auroral	AE, AU, AL, AO	12 observatories in the Northern auroral zone	Auroral index horizontal component disturbances	1 minute	1975 onwards	WDC for Geomagnetism - Kyoto, Japan	IAGA Bulletin 27, 1969, p.123, resolution 2
Regional	ASY, SYM	6 observatories distributes in longitude ±40º	Longitudinally asymmetric and symmetric horizontal component disturbances	1 minute	1981 onwards	WDC for Geomagnetism - Kyoto, Japan	No
Regional	$a\sigma$	based on the am network	K-derived in 4 Magnetic Local Time sectors	3 hour	1959 onwards	EOST, Strasbourg, France	No, mainly used for academic research

SYM-H and ASY-H indices are computed using the measurements from 6 observatories evenly distributed in longitude shown in Table 2.3. The stations are distributed into 6 groups:

- Honolulu
- Boulder and Tucson
- Fredericksburg and San Juan
- Chambom-la-Foret and Hermanus
- Urumqi, Alibag and Martin de Vivies
- Memambetsu

Table 2.3: Distribution of observatories used to compute the SYM/ASY indices.

Observatory name	**Code**	**Geographic Latitude**	**Geographic Longitude**	**Geomagnetic Latitude**	**Geomagnetic Longitude**	**Country**
San Juan	SJG	18.110	293.850	28.04	6.54	USA
Boulder	BOU	40.130	254.760	48.24	321.28	USA
Fredericksburg	FRD	38.200	282.630	48.14	353.93	USA
Tucson	TUC	32.170	249.270	39.73	316.74	USA
Honolulu	HON	21.320	202.000	27.71	270.27	USA
Memambetsu	MMB	43.910	144.189	35.63	211.74	Japan
Urumqi	WMQ	43.800	87.700	34.34	162.53	China
Alibag	ABG	18.638	72.872	10.37	146.55	India
Martin de Vivies	AMS	-37.796	77.574	-46.22	144.93	France
Hermanus	HER	-34.425	19.225	-34.08	84.63	South Africa
Chambom-la-Foret	CLF	48.025	2.261	49.75	85.80	France

In groups with more than one observatory, data can be exchanged between them; if one observatory has an issue, another can be used as backup. Figure 2.4 depicts each group and the location of each observatory.

The procedure for computing the index consists of four steps:

1. Subtraction of the geomagnetic main field and the solar quiet daily variation field to obtain the disturbance field component.
2. Coordinate transformation to a dipole coordinate system.
3. Calculation of the longitudinally symmetric components SYM-H and SYM-D. The disturbance component at each minute for the 6 stations is averaged. For the H component, a latitudinal correction is made on the obtained averages to obtain the SYM-H index, similar to the *Dst* index. For the D component there is no latitudinal correction.
4. Calculation of the asymmetric disturbance indices, ASY-H and ASY-D. The asymmetric component is obtained by subtracting the symmetric component from each disturbance field. They are defined as the range between the maximum and the minimum deviation at each moment for the H and the D component.

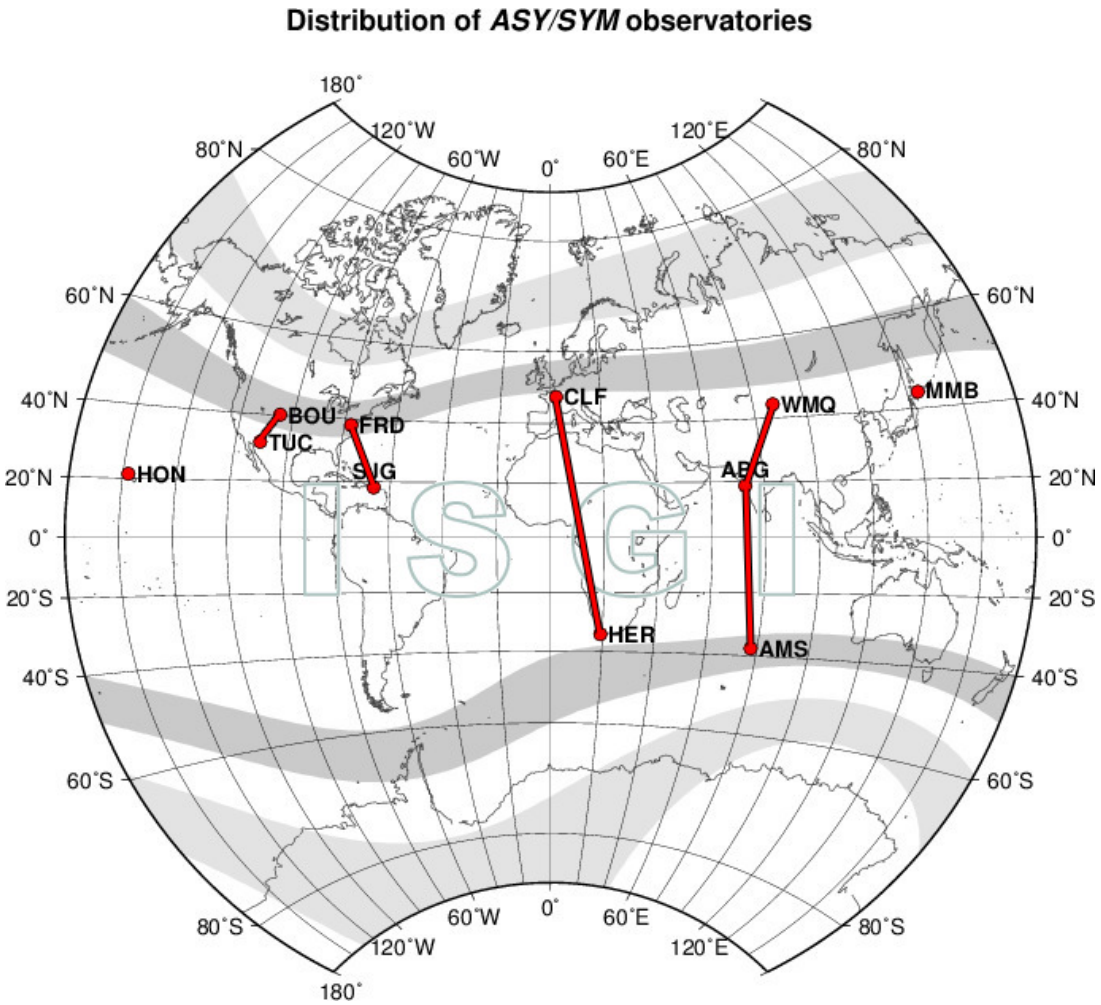


Figure 2.4: Observatories used to calculate the SYM and ASY indices. Data from stations connected by solid lines can be exchanged.

Conversely, the geomagnetic SYM-H and ASY-H indices that quantify the disturbance of the Earth's magnetosphere, jointly provide granular information on the state of the Earth's equatorial ring current. They represent its symmetric and asymmetric components, respectively. Notably, the ASY-H index provides information about the arrival and intensity of geomagnetic storms in standardized units. This aids interpretation of the potential impacts of one such event, especially as it is provided in 5 minutes long timesteps.

2.2.3 Geomagnetic storms

Despite the large amount of data available from both OMNIWeb and ASC, not all of it is equally useful for developing a geomagnetic indices forecasting system. This is due to the significant imbalance between quiet periods, when the solar wind has a negligible impact on the magnetosphere, and the number of moderate and intense geomagnetic storms that reach Earth, which are the main focus of the forecasting system.

Siciliano et al. [36] made a comparison of two popular ANNs architectures to forecast the SYM-H index. To train the networks, they made a selection of the 42 most intense geomagnetic storms that occurred between 1998 and 2018 (solar cycles 23 and 24). In all

42 storms, the SYM-H index reached values below -100 nT. Based on the intensity of the storms:

- 19 out of the 42 storms are considered intense, characterized by a minimum value of the SYM-H index less than -200 nT.
- The remaining 23 are considered moderate, with the SYM-H index reaching minimum values between -200 and -100 nT.

The date of each storm, along with the minimum SYM-H value, maximum ASY-H value, and the occurrence of multiple depressions in the SYM-H index, are shown in Table 2.4.

Considering the guidelines and best practices in any ML project, the available data has to be evenly split into three subsets:

- The Train subset, used to train the network.
- The Validation subset, used to stop the training.
- The Test subset, which will be used to evaluate the performance of the network on previously unseen events.

The three subsets must be evenly populated based on the minimum value of the indices and the occurrence of multiple depressions (MP in the Table 2.4). Sicliano [36] proposed splitting the 42 storms into the three mentioned subsets: the first 20 storms in Table 2.4 are used for training, the next 5 for validation, and the last 17 for testing. In order to ease the comparison of any developed model, it is advisable to use the same storms for each set. As a result, more recent studies have also used the same storms [37], [38]; these will be presented in Section 2.4.

As mentioned in the previous section, the plasma data presents a significant amount of missing values in the ACE database. Table 2.5 shows the percentage of missing plasma values for each storm. Although the percentage is not excessive in many cases, it is quite misleading. Most of the missing values occur during the most intense moments of the storms, when the measuring instruments are more likely to become saturated due to the high number of charged particles. This is of great concern because it is during those key moments when the forecast is most critical.

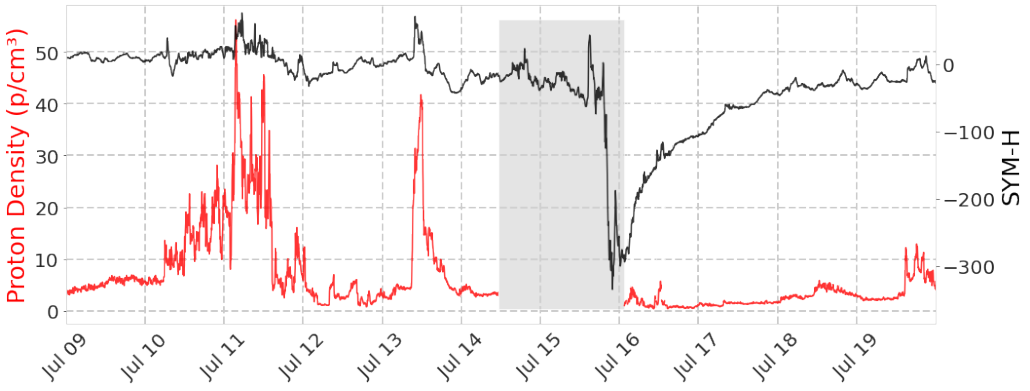
Figure 2.5 shows the measured plasma variables for storm number 6, which occurred in July 2000. This storm is an example of the measuring instruments becoming saturated and ceasing to provide valid measurements (shaded in grey) during the most intense period. In this case, the plasma variables were unavailable for more than 36 hours. Addressing such data gaps is crucial to achieve an accurate forecast of the index.

Table 2.4: Geomagnetic storms occurred between 1998 and 2018 used to train, validate and test DNN models, as proposed by Siciliano et al. [36]. SYM-H and ASY-H values extracted from the `OMNI_HRO_5MIN` dataset.

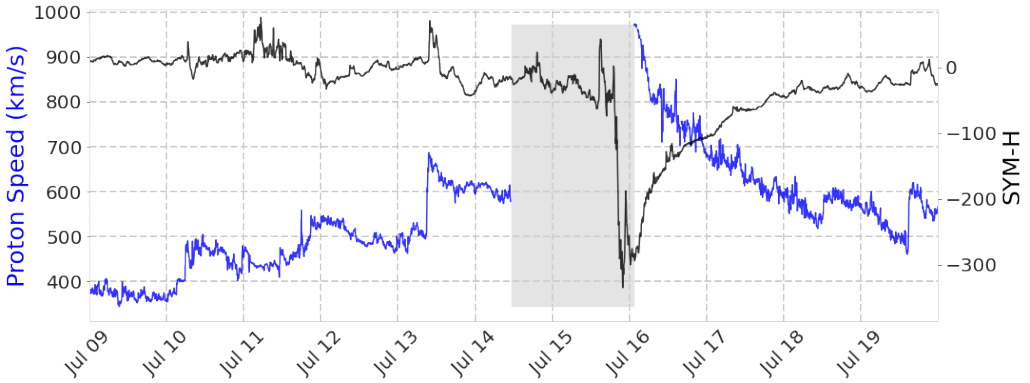
Set	Storm	Start Date	End Date	MP	Min SYM-H (nT)	Max ASY-H (nT)
Train	1	14/02/1998	22/02/1998	Y	-119	188
	2	02/08/1998	08/08/1998	Y	-168	136
	3	19/09/1998	29/09/1998	N	-213	395
	4	16/02/1999	24/02/1999	Y	-127	246
	5	15/10/1999	25/10/1999	N	-218	214
	6	09/07/2000	19/07/2000	N	-335	380
	7	06/08/2000	16/08/2000	Y	-235	206
	8	15/09/2000	25/09/2000	Y	-196	219
	9	01/11/2000	15/11/2000	Y	-174	130
	10	14/03/2001	24/03/2001	Y	-165	225
	11	06/04/2001	16/04/2001	N	-275	422
	12	17/10/2001	22/10/2001	N	-210	154
	13	31/10/2001	10/11/2001	N	-313	329
	14	17/05/2002	27/05/2002	Y	-113	204
	15	15/11/2003	25/11/2003	N	-488	374
	16	20/07/2004	30/07/2004	Y	-208	294
	17	10/05/2005	20/05/2005	N	-302	250
	18	09/04/2006	19/04/2006	N	-110	162
	19	09/10/2006	19/12/2006	N	-206	267
	20	01/03/2012	11/03/2012	Y	-149	229
Validation	21	28/04/1998	08/05/1998	N	-268	415
	22	19/09/1999	26/09/1999	N	-160	157
	23	25/10/2003	03/11/2003	Y	-427	828
	24	18/06/2015	28/06/2015	Y	-207	348
	25	01/09/2017	11/09/2017	Y	-144	230
Test	26	22/06/1998	30/06/1998	N	-120	127
	27	02/11/1998	12/11/1998	Y	-179	196
	28	09/01/1999	18/01/1999	N	-111	147
	29	13/04/1999	19/04/1999	N	-122	138
	30	16/01/2000	26/01/2000	Y	-101	99
	31	02/04/2000	12/04/2000	N	-315	612
	32	19/05/2000	28/05/2000	Y	-159	225
	33	26/03/2001	04/04/2001	N	-434	352
	34	26/05/2003	06/06/2003	Y	-162	377
	35	08/07/2003	18/07/2003	Y	-125	172
	36	18/01/2004	27/01/2004	Y	-137	150
	37	04/11/2004	14/11/2004	Y	-393	339
	38	10/09/2012	05/10/2012	N	-138	130
	39	28/05/2013	04/06/2013	N	-134	154
	40	26/06/2013	04/07/2013	N	-110	165
	41	11/03/2015	21/03/2015	N	-233	250
	42	22/08/2018	03/09/2018	N	-205	197

Table 2.5: Percentage of missing values for the plasma variables of the selected geomagnetic storms (see Table 2.4).

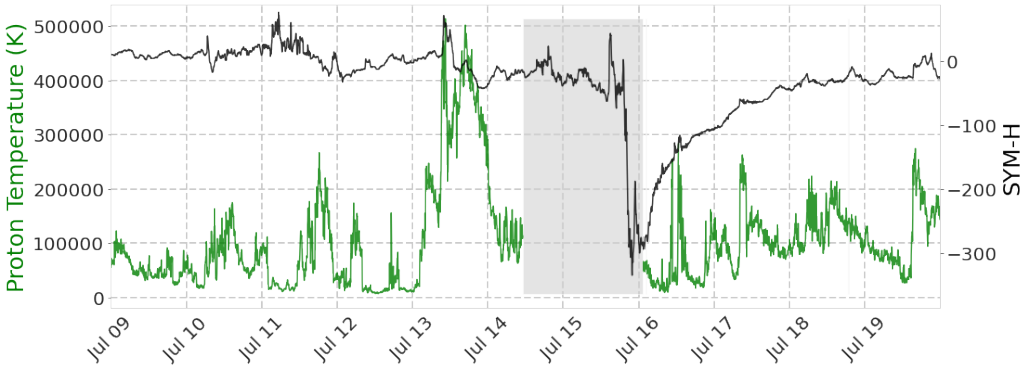
Set	Storm	% of missing plasma values		
		Density	Temperature	Speed
Train	1	24.73	24.73	24.73
	2	0.15	0.15	0.15
	3	0.32	0.32	0.03
	4	0.35	0.39	0.35
	5	8.11	8.11	0.09
	6	14.80	14.80	14.80
	7	2.65	2.65	2.65
	8	1.70	1.74	0.07
	9	11.60	11.60	9.42
	10	44.88	44.88	0.29
	11	0.35	0.35	0.32
	12	43.63	43.63	0.35
	13	35.39	35.39	10.35
	14	2.24	2.24	0.22
	15	1.01	1.01	1.01
	16	3.69	0.16	0.13
	17	0.76	0.82	0.76
	18	2.27	1.80	0.32
	19	7.32	7.32	6.98
	20	83.43	26.01	20.90
Validation	21	12.09	12.09	0.60
	22	0.00	0.00	0.00
	23	32.01	32.01	32.01
	24	36.99	6.60	2.78
	25	25.92	15.47	12.82
Test	26	0.54	0.54	0.54
	27	0.57	0.57	0.57
	28	1.49	1.49	1.49
	29	0.25	0.25	0.25
	30	12.66	12.66	0.13
	31	0.57	0.57	0.28
	32	0.62	0.62	0.62
	33	4.17	4.17	0.22
	34	0.09	0.09	0.09
	35	1.48	1.48	0.09
	36	0.03	0.03	0.03
	37	4.70	4.70	0.03
	38	92.57	39.32	7.36
	39	4.30	0.48	0.48
	40	13.31	0.50	0.50
	41	59.00	7.26	0.35
	42	29.43	2.43	0.13



(a) Proton density



(b) Proton Speed



(c) Proton temperature

Figure 2.5: ACE Measured plasma for the storm of July 2000 (Storm 6). Data gaps are shaded in gray.

2.3 Fundamentals of Machine Learning

ML is an application of Artificial Intelligence (AI) where a computer learns to calculate an output given an input (such as text, audio, images, numbers, etc.) without being explicitly programmed to do so. ML models learn by being exposed to enough representative examples (known as training data) and are tested by evaluating their performance on previously unseen examples [39]. There are several branches of ML, depending on the nature of the data and the techniques used. This work focuses on ANNs.

ANNs are computing systems inspired by the biological neural networks that constitute brains [40]. Nowadays, several problems that were hard to solve have been approached from the ANNs perspective with great success. For example, Generative Pre-trained Transformer 3 (GPT-3) [41] is a high-quality Natural Language Processing (NLP) model, trained using Internet data to generate various types of text. Just looking at the output of this model can be difficult to determine whether or not it was written by a human. RankBrain is a ML based solution used by Google to help processing search queries and to provide more relevant results for the user [42]. Other popular applications are object recognition in images [43].

ANNs are, in essence, a collection of connected units, called artificial neurons, that are organized in layers. The whole collection of layers is known as a model. These neurons attempt to imitate the behavior of real neurons in a biological brain. The human brain has billions of interconnections; as we learn, those connections can be formed, removed or changed. The greater the number of learning instances for a particular task, the more refined the connection between the neurons will be, becoming better at whichever task the brain is learning. The neurons process the signal they receive and then pass it along to a subset of network neurons they are connected to, depending on the stimulus; that connection is learned. In the ANN implementations, the aforementioned stimulus that is passed along is a real number, this number can either be the original input fed to the network or a computation made by another neuron.

This computation is regulated by several parameters (real numbers) that govern the behavior of the network. These parameters control which neurons are activated and their degree of activation. Regular artificial neurons have two parameters, known as weight and bias, that can be adjusted until the desired output is achieved. However, other types of neurons may have additional parameters depending on their complexity. The process of adjusting the trainable parameters to values that produce the desired network output is known as network *training*. The most common approach is to progressively find which neurons caused the output to be wrong, and then, try to change their parameters so the final output is closer to the desired value.

At a fundamental level, each artificial neuron applies a simple mathematical operation to its inputs. Consider a basic fully connected neural network with two hidden layers. Let the input vector be denoted as $\mathbf{x} \in \mathbb{R}^n$. The first layer computes an output $\mathbf{h}_1$ by applying a linear transformation followed by a non-linear activation function: $\mathbf{h}_1 = f(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$, where $\mathbf{W}_1$ is a matrix of learnable weights, $\mathbf{b}_1$ is a bias vector, and f is the activation function (such as ReLU or sigmoid). This output is then passed to the second layer: $\mathbf{h}_2 = f(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2)$, resulting in the final layer output. In a regression task, this might be directly used as the prediction, whereas in classification, further processing (like a softmax layer) may be applied.

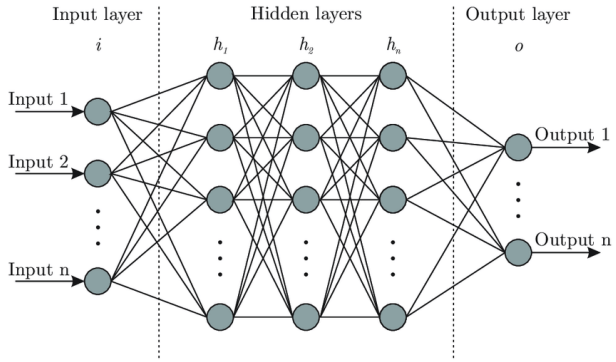


Figure 2.6: Artificial Neural Network architecture.

These calculations illustrate the compositional nature of neural networks: each layer applies a transformation to the output of the previous one, progressively refining the information until a final prediction is made. The set of all weight matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots$ and biases $\mathbf{b}_1, \mathbf{b}_2, \ldots$ constitutes the network's parameters. During training, these are adjusted to minimize the loss function, effectively learning the best way to map inputs to the desired outputs based on the examples provided in the training dataset.

The training process of a network follows an iterative approach, in which each iteration is known as *epoch*. In each epoch, the network processes the entire training set. Then, the parameters are updated in search of better accuracy. Sometimes, when we update the parameters, we experience a decrease in the network's accuracy. Thus, it is common to save a copy of the parameters that have performed best until we stop iterating through the dataset.

Figure 2.6 depicts the base architecture of an ANN. We can see that the network is split into three sections:

1. *Input Layer:* It is the original input data that will be fed to the network, it can be an image, text, time-series data, etc.
2. *Hidden layers:* The input then will be processed across several intermediate layers, performing the required operations to calculate the desired output based on the input data.
3. *Output layer:* Gathers the processed information from the previous layers and provides the solution that the network has learned for the given input data.

The performance of the network is evaluated using *loss* functions. They evaluate how well the network predicts the expected values, comparing the values obtained by the output layer to the expected values. If the discrepancy between the output values and the target ones is large, the loss functions will produce high values. Broadly speaking, loss functions can be classified into two major categories depending on the data and the learning task: Regression losses and Classification losses.

- In classification tasks, the objective is to predict which class the input data belongs to. For example, given a large amount of handwritten digits, categorize them into one of the 0-9 digits. In general, the score of the correct category should be greater than the sum of scores of all incorrect categories by some safety margin. For this

type of classification tasks, the Cross-Entropy Loss is the most common one. An example of a classification task related to SW would be classifying the category of a CME.

- In regression tasks, the objective is to predict a continuous value. One of the most usual tasks is to estimate the price of a house or predicting stock prices. The most common loss functions for such tasks are the Mean Absolute Error (MAE) and Mean Square Error (MSE). As their names suggest, the MAE takes the average sum of the absolute differences between the real and predicted values, whereas the MSE takes the squared differences. In practical terms, the difference between them is that the MSE makes emphasis on larger errors, while giving less importance to smaller ones; whereas the MAE gives an overall similar importance to all of them.

This project falls under the regression category. We will forecast the values of a geomagnetic index, and as such, we will use regression losses to evaluate the performance of our networks.

The term *Deep*, commonly used along with ANN as DNN, usually refers to the number of different layers stacked in the hidden layers section. The amount of computations needed to calculate the final output are directly related to the amount of intermediate hidden layers. On very large or deep networks, the original input will undergo a lot of computations until the final output values are achieved [44]. The simplest way to estimate a network's complexity and computational cost is by examining the number of trainable parameters, as it is related to the number of layers in the network. A simple network could work with just 20 trainable parameters, but very complex and deep networks can have up to millions of trainable parameters. However, this metric is not always accurate, as networks can be highly complex by using special layers or loop connections without an excessive number of parameters.

As mentioned earlier, these networks learn to solve the problem by being exposed to sufficient representative examples of the problem that they will learn to solve, as in input and expected output pairs. The deeper the network is, the more computations will be needed to make the predictions. However, recent advances in technology have made possible to train very deep networks, with a massive amount of trainable parameters, by exposing them to a huge amount of examples in an affordable amount of time. This is possible mainly thanks to the parallelism of the training process; the networks are exposed to several inputs at the same time, known as *batch*. Then, the parameters are updated using the loss function of the whole batch of data. This parallel computation is often performed in the GPUs, due to their high amount of matrix computation cores. To do so, the most common framework is Compute Unified Device Architecture (CUDA), used by several libraries such as TensorFlow [45] or PyTorch [46].

Nevertheless, ANNs are not perfect, as the main drawback of this approach is that ANNs work as a *black box*, i.e. we can't know what exactly happens inside the network for it to reach its conclusion, unlike the *white box* solutions. White box models give information about the steps taken to reach the conclusion, such as decision trees [47]. In this regard, researchers are doing considerable efforts regarding the interpretability of ANNs [48].

In general, ANNs are developed using one or more hidden layers for mapping the function that can produce the desired output based on given inputs. There are a wide

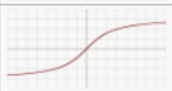
Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU)		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU)		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1+e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$

Figure 2.7: List of the most common activation functions.

variety of layers that perform different operations tailored to specific needs: the problem that they aim to solve or how the input data is organized (scalar values, time-series, images, video, words, etc). The most common layers are:

- Dense layers: They are the most used layers in ANNs and they serve as a general purpose layer. It is a fully connected layer that connects every input, which can be either a previous layer or the raw input, to every neuron on its layer. Then, the following formula is applied to each neuron: $y = f(w * x + b)$, where x is the input to the neuron, w and b are the parameters that we will need to train, w is known as *weight* while b is known as *bias*. Finally f is known as the *Activation function*. There are several functions, being the most common ones the ones shown in Figure 2.7. Nevertheless, the most used ones are the *ReLU* and the *Linear* functions given their overall good performance and their low computational cost.

- Convolutional layers: Are most commonly used when working with images. In images, the data is organized in two-dimensions. To exploit this, Convolutional Neural Networks (CNNs) are able to capture spatial dependencies [49]. Although the main purpose is related to image processing, they have also been widely applied to time-series related works. In these cases, temporal dependencies are identified instead of spatial ones, and they are used to make the prediction.

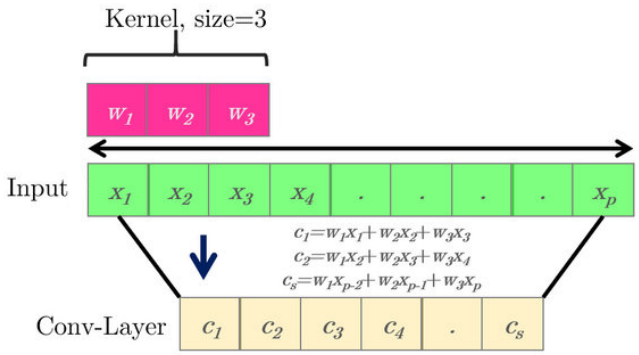


Figure 2.8: 1 dimension convolutional layer operation diagram.

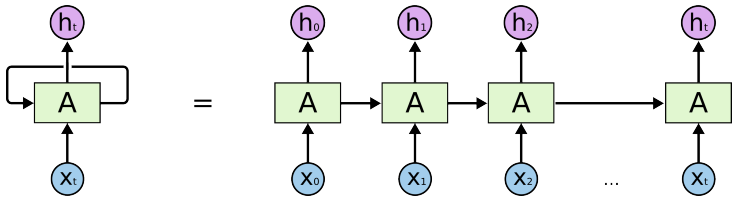


Figure 2.9: Unrolling of a recurrent neural network.

Regarding how CNNs work, they involve a convolution function, as shown in Figure 2.8, where the *Kernel*, also known as *Filter*, is represented in pink. This kernel has a fixed, customizable, length. It goes through the input vector, represented in green, analyzing it in *"mini-packs"*. This process allows the layer to capture information than could not be obtained if the time series was analyzed individually. This approach relies on the fact that most of the data is organized in a hierarchical pattern. Finally, the result of the convolution function is represented in the yellow color in the image.

- Recurrent layers: Are specifically designed to solve the problem of the temporal dependencies in the input data. Although the previous convolutional layer could capture temporal dependencies, it is not capable of retaining such information. In this regard, the recurrent layers are specifically designed to tackle that problem.

  If we make an analogy of how a human thinks and how a neural network operates, we will see the fault in the previous layers. Both the dense and convolutional layers start from scratch every iteration, giving always the same output to the same singular input. However, humans do not restart their thinking every second, while you are reading something, your understanding of each word is based on the previous word understanding. The thoughts have persistence, and they are not started from scratch every time.

  Recurrent Neural Networkss (RNNs) were developed to face this issue. These networks have loops inside, allowing the persistence of important information. The main idea is shown in Figure 2.9. In the diagram, we can see that on the subsequent executions of the network, some kind of information is passed from the previous iteration (h in the image), one step at a time. This flow allows the persistence of important information and its usage in the future.

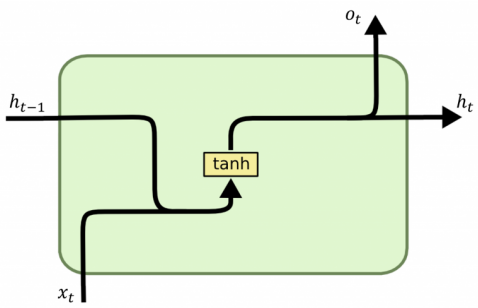


Figure 2.10: Core of a recurrent neural network.

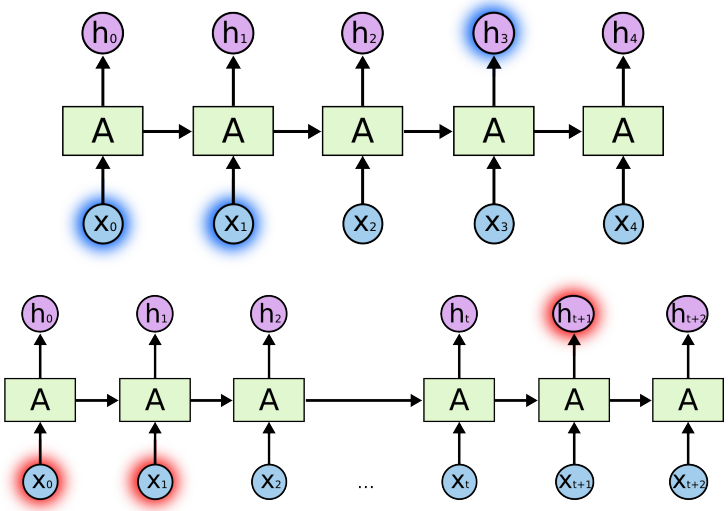


Figure 2.11: Only dependencies not too far apart are properly used.

The main appeal of the RNNs is their ability to associate previously processed information to the current state. In general, this works generally well using simple recurrent layers such as the one in Figure 2.10. However, this memory only lasts for a finite amount of steps. Therefore, it only works if the place where we need to use the saved memory is relatively close to when it was processed. As shown in Figure 2.11-top, when the critical point where we need to use the previous information (X_0, X_1) is close to when we processed it, the expected output (h_3) is produced. But when they are too far apart (Figure 2.11-bottom), the information (X_0, X_1) has already been forgotten, thus the expected output is not achieved (h_{t+1}).

Due to this problem, long term dependencies can not be solved using the previously shown layer in Figure 2.10. In order to be able to solve this kind of long term dependencies, a new, more complex layer needed to be developed. This new layer is called Long Short-Term Memory (LSTM). This layer was initially introduced by Hochreiter & Schmidhuberare [50]. It is a special kind of RNN capable of coming out on top when fighting this long term dependency challenge. Instead of having only one operation at its core like in Figure 2.10, and passing its information to the following layer, LSTMs have four layers, forming three groups, known as *gates*, as shown in Figure 2.12. These gates are organized in a very particular, yet successful way, and can preserve the critical information for a longer time.

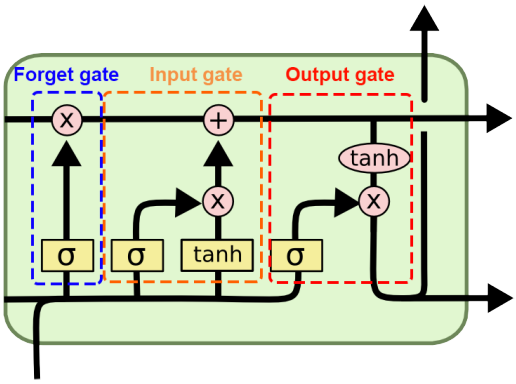


Figure 2.12: Core of an LSTM layer.

The novel concept behind LSTMs is the *cell state*, represented by the horizontal line on top of the Figure 2.12. This line represents the long-term information that is passed along to the next time-step, whereas the bottom line, the *hidden state* represents the short-term information. The behavior of each state is regulated by the LSTM layers, that is, which information is added or removed to each state, and how the new inputs are processed.

If we walk through all the layers, represented as yellow boxes in Figure 2.12, we can understand how the LSTMs work. Yellow rectangles represent Dense layers with an activation function (σ and *tanh* represent the Logistic and Hyperbolic tangent activation functions, third and fourth in Figure 2.7), whereas the red circles represent operations. The first step, known as the *forget gate layer*, decides what information will be thrown away from the previous cell state (represented by the first σ yellow square on the diagram). The next step, the *input gate*, comprises two layers (second yellow σ square, and the tanh square). They decide which new information will be added to the cell state (the + red circle in the top center of the diagram). Finally, the *output gate* computes the output of the current time-step, using a filtered version of the already processed cell state (tanh red circle), combined with the output of the last σ layer. This regulates how much information from the long-term section will be part of the output.

Those are the most common layers, and are the ones that have been used in the recent geomagnetic indices forecasting works that will be presented in Section 2.4. However, the ML field is rapidly evolving; new layers, activation functions and procedures that enhance the capabilities of the ANNs are continuously appearing. Problems that previous architectures could not solve, are being solved and the models that were considered the state of the art are constantly being outperformed.

One of the most important layers recently developed are the ones based on *Attention*. This technique mimics human cognitive attention. The layer enhances some parts of the input data while diminishing other parts, following the hypothesis that the network should focus on the small but important part of the data. This new approach has been used with great success in the NLP field [51], in computer vision [52] and time-series forecasting [53].

However, this relation is highly dependent on the context and may not work for every situation.

Another key factor in the development of ANNs is the optimizers. We have previously mentioned that the parameters of the networks are progressively updated so the output values of the networks are closer to the desired target values. As such, optimizers are algorithms and strategies that change those parameters in order to reduce the loss function of the network, until it converges, that is, ceases to improve.

In general, the optimizers rely on the backpropagation algorithm. This algorithm is based on computing the gradient of the loss function with respect to the weights of the network, by calculating the derivative of the loss function.

The most basic optimizer is the Gradient Descent (GD) algorithm, which depends on the first-order derivative of the loss function, calculating how the weights should be modified to minimize the function. Through backpropagation, the gradient is calculated for the previous layers and updated accordingly.

The main drawback of the GD is that it requires a lot of memory to load the entire dataset and compute the derivatives, and requires a lot of iterations to converge. In this regard, the Stochastic Gradient Descent (SGD) algorithm is an extension of the GD that can update the weights of the network using an arbitrary number of training samples, instead of requiring the whole dataset.

Another key factor is the *learning rate*, which refers to how much the weights are updated in each epoch. Meanwhile, the learning rate in the GD algorithm remains constant, other optimizers adapt it during the training process. This parameter is key, since a very low learning rate will lead to a very slow convergence. The convergence is defined as when the network achieves a state during training in which loss settles and additional training will not improve the model.

All other optimizers are based on the SGD algorithm, incorporating additional strategies and constraints to accelerate the training process, avoid local minima, and facilitate convergence. One of the most important additions is the usage of the *momentum*. The momentum's objective is to decrease the noise and the impact that a single sample could have. That is achieved by aggregating the previous gradients to update the weight. Another common modification is to have an adaptive learning rate, progressively decreasing it with the number of iterations.

Nowadays there are several optimizers that are tailored to specific problems, layers, or data, such as Adadelta [54], AdaMax, RMSProp [55] or Adam [56], among others.

2.4 State of the art of Machine Learning in Space Weather

Geomagnetic indices quantify disturbances in the Earth's magnetosphere caused by solar wind at different latitudes, providing a single-number summary of the magnetosphere's status. They can monitor the evolution of complex phenomena while at the same time offering reliable and concentrated information. Accurately forecasting geomagnetic indices, which encompass significant information, can enhance our understanding of the magnetosphere's dynamics and help mitigate the consequences of geomagnetic storms.

Accurately forecasting the geomagnetic indices is an ongoing problem. This challenge has been approached from both physics-based models and, more recently, using ML techniques. On the one hand, the ML techniques rely on exploiting Sun data gathered from different space probes during the last decades. On the other hand, the physics approach enables a comprehensive understanding of the different elements involved in a geomagnetic storm. However, the problem is extremely complex, involving magnetic entanglements, plasma physics, orbital dynamics, and so on. Thus, developing reliable forecasting models based on a pure physics approach is almost unfeasible nowadays. Since there are large datasets of Sun observations, it is feasible to use ML techniques. Although these models operate as black boxes and it remains uncertain whether they truly capture the underlying physical mechanisms, they are capable of providing useful and often accurate predictions. This is a current trend due to the promising results that can be found in the literature. Camporeale [57] provides a comprehensive survey on the current state of ML applications in SW nowcasting and forecasting, which will be referenced throughout this section. In that work, he describes and reviews the state of the art for the activities in which ML has been applied with success in the SW context: the forecasting of geomagnetic indices, the estimation of relativistic electrons at geosynchronous orbits, solar flares occurrence classification, forecasting the CMEs propagation time, forecasting of solar wind speed, estimation of the plasmaspheric electron density, etc. There is also an overview of the most used databases and a review of the appropriate performance metrics for each task.

From a ML time-series perspective, the geomagnetic indices forecasting problem can be summarized as a multivariate time series forecasting task, in which the target variable—the geomagnetic index—is closely related to the solar wind parameters measured by space probes. The most important features gathered by these probes are the IMF and plasma properties. The IMF includes the magnitude of the field and its components B_x, B_y, and B_z, while the plasma features comprise proton density, speed, and temperature. These variables are not only strongly correlated with the geomagnetic indices, especially during storm periods [58], but also allow for the derivation of other physically meaningful parameters, such as the solar wind dynamic pressure and the interplanetary electric field, which are themselves significant for space weather forecasting. The time resolution of the measured variables depends on the instruments onboard the spacecraft (see Section 2.1.1). For instance, the ACE spacecraft measures the IMF variables with a 1-second resolution and the plasma variables with a 64-second one. The IMF features present an almost complete availability, and the few data gaps are resolved when the data is resampled to match the resolution of the geomagnetic indices. However, this is not the case for the plasma features, which contain a considerable number of data gaps (as shown in Table 2.5). How to deal with missing data in the input features will be addressed in Section 3.2.

Liemohn et al. [59] also present a series of guidelines specifically tailored towards geomagnetic index forecasting, assessing which metrics are the most appropriate for each index. Their work recommends metrics to measure both the fit of the predictions, that is to say, how close is our prediction to the original values; and the event detection performance, which indicates how many events would have been detected by the usage of our prediction. In addition, a showcase of the different state of the art forecasting models for each index is shown, as well as their performance using the recommended metrics for each index. In general, that work gives an accurate overview of the current level of accuracy for the prediction of each index.

To properly evaluate the performance of the forecasting models and compare them, there are several metrics that can measure the quality of the prediction. Which metrics are the most appropriate to evaluate the models' performance is a topic that has already been addressed in the literature, through the different surveys and works [36], [37], [57]. They are pretty much set in stone: the RMSE and the coefficient of determination R^2 calculated on the predictions. In most ANN-related problems, it is standard practice to scale both input and output data to values close to zero due to the behavior of most activation functions. However, to compare models accurately, the predictions must be unscaled before computing metrics. Using other metrics while not computing the standard ones would difficult the comparison of different models, and, therefore, is a practice to avoid.

RMSE, defined in Equation 2.1, provides information regarding how much the predictions made by the model deviate from the real observations of the target variable. It penalizes larger differences between the observed values (y_i) and the predictions ($\hat{y}_i$). In our SW context, this error presents the highest values during the peaks of the storms.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.1}$$

R^2, defined in Equation 2.2, assesses the amount of variance of the observed data explained by the predicted value, that is, the strength of the relationship between the predicted value and the real one. This metric usually ranges between 0 and 1, being 1 a perfect prediction, whereas 0 indicates that the model is equivalent to predicting the mean of the target variable ($\bar{y}$).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.2}$$

Regarding the evaluation of the models, they are usually compared to a persistence model. In the time series forecasting context, a persistence model predicts that the next value will be equal to the current one, it gives an idea of how well the models being evaluated actually performs on the problem, and it is a clear metric that should be overcome by the developed models.

The official values of several indices, including *Dst* and SYM-H, are obtained from ground magnetometers and provided through the WDC Kyoto [1]. Additionally, provisional *Dst* is available with one-hour latency, while SYM-H is not available in real-time. This delay in the nowcasting of the geomagnetic indices makes their forecasting even more necessary, since some SW systems would benefit from having the indices in real-time to drive theirs operation. For this purpose, real time solar wind parameters measured at the L1 point are used to feed the systems that forecast these indices. The forecasting methods based on physical or empirical relationships between the solar wind and the indices start with the works by Burton et al. [60] and O'Brien et al. [61], which achieve a reasonable level of accuracy, about 30% better than the persistence model. Rastatter et al. [62] compares the performance of 30 models forecasting a 1-min *Dst* index (equivalent to SYM-H) during four events: two events representing highly disturbed situations, and the

[1] http://wdc.kugi.kyoto-u.ac.jp/

other two events representing quieter times. None of the models consistently performed best for all the events.

Initially, due to the large availability of historical data, the work was centered around the forecasting of the *Kp* index [63], [64]. Zhelavskaya et al. [65] made a systematic analysis of the different methods and ML techniques tailored towards the *Kp* index prediction: they tested how different algorithms perform on nowcasting and forecasting the *Kp* for prediction horizons of up to 12 hours. The models were trained using solar wind and IMF data from NASA OMNIWeb data service during the time period of 1998-2017. The best model achieved a RMSE of 1.0632 and a Correlation coefficient of 0.6477. Other authors such as Hernandez et al. [66] focused on other indices; they forecast the *AL* index using ANNs with non-linear approaches, given the solar wind and the IMF data.

Some models are even being used in real-time by different countries, such as Sweden's Regional Warning Center [2], USA's Rice Space Institute [3], Brazil's Instituto Nacional de Pesquisas Espaciais [4], China's Space environment Prediction Center [5]. Finally, the US Space Weather Prediction Center (SWPC/ NOAA), from 2010 to 2018 provided real-time *Kp* forecasts using the model from Wing et al. [63] for 1 and 4 hours in advance until it was replaced by a physics based model developed at the University of Michigan [67].

However, the objective of some models is not the forecast of the indices' values; they predict the occurrence of a storm. Since this approach is a classification problem, instead of a regression one, different metrics needed to be employed. For example, Tan et al. [68] and Maimaiti et al. [69] used the F1 score as a metric for their classification task. The F1 score is the harmonic mean of the precision and recall, where the precision is the number of true positives divided by the total number of positives (including false positives), and the recall is the number of true positives divided by the total samples that should have been classified as positives. On the one hand, Tan et al. used it as a metric, along with RMSE and MAE, to train a LSTM based ANN to forecast the occurrence of geomagnetic storms (when the *Kp* index reaches values higher than 5). On the other hand, Maimaiti et al. developed a classification model to predict the occurrence probability of the onset of a magnetic substorm over the next hour. Their model was based on CNNs, they used the solar wind speed, proton density and the IMF as input variables.

Nevertheless, the index that has traditionally received most of the attention has been the *Dst* index. Due to its hourly nature it is preferred over the *AL* and *KP* indices. It is also used to drive operational forecasts of various geospace systems, including the thermosphere for mass density and satellite drag. Considering that, Licata et al. [70] benchmarked several forecasting models used to drive the High Accuracy Satellite Drag Model, employed by the USAF. Hu et al. [71] used the forecasting of the *Dst* index to monitor the commencement of ionospheric storms, since they have severe effect on high-frequency communications, reducing the maximum usable frequency. Similarly, several works have been developed in order to forecast the *Dst* index. For example, Wei et al. [72] developed a forecasting system using both active and non-active periods, whereas Ji et al. [73] focused on the forecasting of intense geomagnetic storms. Despite this index being used to drive real-time operations, its forecast has considerable problems when used that way, since plasma values, which have an important correlation to the index, are often not

[2] http://www.lund.irf.se/
[3] http://mms.rice.edu/
[4] http://www2.inpe.br/climaespacial/portal/swd-forecast/
[5] http://eng.sepc.ac.cn/Kp3HPred.php

available in real-time; this is caused by the saturation of the solar wind plasma instruments when exposed to large emissions of particles and radiation [74].

If we focus on ANN related works, we can also highlight the work by Gruet et al. [75]. They combine LSTM layers with a Gaussian process to provide a forecast up to 6 hours ahead of the *Dst*. This model used hourly data from the OMNIWeb and GPS databases. Their model obtained great accuracy, with a correlation coefficient higher than 0.873 and an RMSE lower than 9.86 nT. However, despite the overall performance, it is unable to obtain accurate predictions of intense storms where the *Dst* reaches values lower than -250 nT.

Lazzús et al. [76] explored and compared several ML techniques for the *Dst* index forecast problem. In their work, several ANN are studied, as well as its combination with bio-inspired algorithms, such as particle swarm optimization, genetic algorithms and a hybridization of both to improve the system's accuracy. In order to evaluate the models' performance, the RMSE and the correlation coefficient metrics were computed, as recommended in the previously mentioned surveys. Their model achieved a RMSE lower than 5 nT when forecasting up to 3 hours in advance, and a RMSE lower than 7 nT when forecasting up to 6 hours ahead.

However, the hourly nature of the *Dst* can lead to a loss of relevant information. Since ML algorithms can take advantage of high resolution data, the latest research is focused on forecasting the SYM-H and ASY-H indices. Both indices have one minute-resolution, so the amount of information that may be lost due to the averaging is almost negligible. Regarding these indices, SYM-H is considered a high resolution *Dst* index; both are proxies of the symmetric ring current. Therefore, the SYM-H index has been increasingly used over the *Dst* index. For example, the SYM-H index is used by Forsyth et al. [77] to construct forecasts of the >2 MeV flux. Tshisaphungo et al. [78] used the SYM-H index in their model to forecast the ionospheric F2 layer (foF2) changes during geomagnetic storms. Similarly, the ASY-H index measures the longitudinally asymmetric part of the geomagnetic disturbance field at low- to mid-latitude [33], [79].

For the forecast of these indices, we can highlight the work of Bhaskar and Vichare [80] using a Nonlinear Autoregressive Network with exogenous inputs to provide the SYM-H and ASY-H prediction. Their model considers an input history of 30 minutes and an output feedback of 120 minutes. As for the inputs, they consider velocity, density and IMF. For the training data, SYM-H and ASY-H indices during geomagnetic storms from 1998 to 2013 are used. However, only storms when the SYM-H reaches values below -85 nT are used as the target for training the networks. Predictions for both indices during nine geomagnetic storms of the 24th solar cycle are showcased, including the large storm that occurred on St. Patrick's Day in 2015, presenting the model capabilities for predicting these indices about an hour before the storm reaches the Earth. Regarding the data gaps that can be found in the plasma variables, the authors opted to linearly interpolate the missing values. However, using the interpolation method makes this approach unfeasible in a real-time environment.

Siciliano et al. [36] made a comparison of two ANNs for the SYM-H index forecasting in the next hour: one based on LSTM layers and another one based on CNN layers. With the objective of developing operational models, they only use the IMF measured by ACE's magnetometer (they discard variables such as speed or density due to the data gaps) to forecast storms in which the SYM-H index achieved values lower than -100 nT. They also

Table 2.6: Metrics for the SYM-H prediction made by the models of Siciliano et al. [36] and Collado-Villaverde et al. [37] over the test storms set (Table 2.4).

#	SYM-H 1 hour prediction				SYM-H 2 hours prediction			
	RMSE		R^2		RMSE		R^2	
	Siciliano	Collado	Siciliano	Collado	Baseline	Collado	Baseline	Collado
26	6.7	6.630	0.89	0.870	12.297	8.989	0.555	0.766
27	8.9	8.913	0.94	0.939	15.050	13.418	0.826	0.862
28	5.4	5.858	0.95	0.936	9.332	5.877	0.838	0.936
29	7.2	6.683	0.93	0.922	11.305	9.314	0.777	0.848
30	5.6	5.200	0.95	0.946	7.363	7.288	0.891	0.894
31	10.7	8.584	0.96	0.971	17.116	12.436	0.885	0.939
32	8.3	7.259	0.95	0.953	15.170	8.937	0.795	0.929
33	16.3	13.340	0.96	0.965	29.703	18.481	0.825	0.932
34	11.3	10.034	0.75	0.798	15.048	13.941	0.548	0.612
35	8.5	7.693	0.90	0.907	11.137	9.932	0.805	0.845
36	8.7	9.525	0.89	0.864	14.669	12.058	0.677	0.782
37	17.5	15.184	0.96	0.966	30.190	21.084	0.865	0.934
38	4.2	4.080	0.94	0.939	7.346	5.213	0.802	0.900
39	5.6	6.431	0.96	0.932	12.235	6.798	0.754	0.924
40	5.5	4.673	0.95	0.966	6.340	5.281	0.937	0.957
41	9.0	7.882	0.96	0.969	15.269	11.707	0.882	0.931
42	5.9	5.669	0.97	0.968	10.120	8.273	0.898	0.932
Avg:	8.547	7.861	0.93	0.930	14.099	10.527	0.798	0.878

compare the performance of the networks, both with and without the index itself as an input parameter. The model based on LSTM layers achieved a RMSE of 23.6 nT without using the index as an input and 9.0 nT using it. By its side, the CNN model achieved a RMSE of 21.4 nT without using the index and 9.7 nT using it.

Later on, Collado-Villaverde et al. [37] combined both the LSTM and CNN layers into a single network for providing the forecast of the SYM-H and ASY-H indices for the next one and two hours simultaneously. The storms used to train, validate and test the model were the same as those used by Siciliano et al. [36] (presented in Table 2.4), to facilitate model comparison. Since the objective of the work was the development of a real-time forecasting system, they only use the IMF and the indices themselves. The plasma variables are discarded due to the large amount of missing values.

Considering the objectives of recent works [36], [37], the main research focus has shifted towards real-time forecasting of the SYM-H and ASY-H indices. Taking into account the constraints imposed by the real-team time system, only the IMF and the index itself are used as an input to predict the indices values in the two subsequent hours. The computed metrics from the two mentioned works for the SYM-H indices are shown in Table 2.6; one work is limited to one hour prediction (Sicliano et al. [36]) whereas the other does it over one and two hours (Collado-Villaverde et al. [37]).

For the ASY-H index forecasting, the comparison is harder than for the SYM-H. The last two developed models, made by Collado-Villaverde et al. [37] and Bhaskar and Vichare [80] used different training and testing storms. Thus, only the storms that both models used for testing can be compared. In this case, only one test storm is shared, the one that occurred in March 2015. However, both works used different time frames for the storm, Collado-Villaverde et al. [37] reported the metrics for the 10 days around the peak whereas the metrics of Bhaskar and Vichare were calculated around 23 days, making the

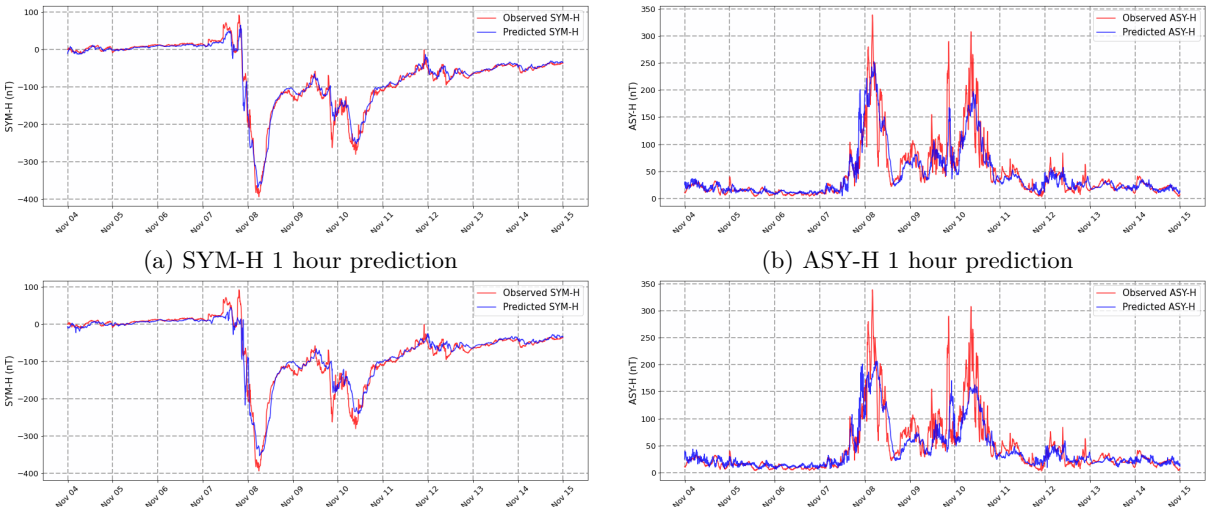
Table 2.7: Metrics for the ASY-H prediction of Collado-Villaverde et al. [37] over the test storms set (Table 2.4), comparing with the baseline for the 1 and 2 hours prediction.

#	ASY-H 1 hour prediction				ASY-H 2 hours prediction			
	RMSE		R^2		RMSE		R^2	
	Persistence	Collado	Persistence	Collado	Persistence	Collado	Persistence	Collado
26	9.104	8.688	0.685	0.713	13.682	10.298	0.289	0.597
27	13.331	12.522	0.788	0.813	18.177	14.921	0.606	0.735
28	12.876	10.442	0.621	0.751	16.764	11.623	0.359	0.692
29	9.356	9.089	0.791	0.803	13.122	10.843	0.592	0.722
30	8.015	8.917	0.753	0.695	11.106	9.897	0.528	0.625
31	23.347	21.174	0.677	0.735	30.737	26.128	0.442	0.597
32	14.742	11.264	0.604	0.769	17.993	13.285	0.412	0.679
33	19.438	18.860	0.741	0.756	25.525	20.219	0.554	0.720
34	24.616	18.932	0.266	0.566	29.080	21.359	-0.021	0.449
35	17.054	13.979	0.494	0.660	21.331	16.273	0.209	0.540
36	16.590	14.116	0.344	0.525	20.247	16.274	0.026	0.371
37	29.021	23.115	0.692	0.805	33.708	27.533	0.586	0.724
38	7.386	7.335	0.623	0.628	9.892	8.203	0.324	0.535
39	12.126	10.639	0.655	0.734	15.143	10.412	0.464	0.747
40	10.536	9.163	0.779	0.833	11.943	9.826	0.718	0.809
41	17.818	17.924	0.731	0.728	21.462	20.380	0.611	0.649
42	11.126	9.944	0.802	0.842	13.962	11.144	0.689	0.802
Avg:	15.087	13.300	0.650	0.727	19.051	15.213	0.435	0.647

comparison not strictly fair, since the longer the period, more non-disturbed time will be predicted which will decrease the overall error metrics. Nevertheless, the model by Collado-Villaverde et al. [37] achieved a lower RMSE of 17.924 nT, below the 20.43 nT achieved by Bhaskar and Vichare. Due to the difficulty in making any comparison, both forecasts compare the obtained metrics against a persistence model. The results of the model by Collado-Villaverde et al. [37] are shown in Table 2.7.

An example of the forecasting made by the model developed by Collado-Villaverde et al. [37] is shown in Figure 2.13. It is a very intense storm in which the SYM-H index reaches values of almost -400 nT. As we can see, the one-hour prediction for both indices is considerably better than the two hours one. Additionally, SYM-H forecasting is significantly more accurate than ASY-H. Despite the great performance of the models, there are some intervals where the predictions deviate from the real values. As discussed earlier, the plasma parameters, which are closely related to the indices and are essential for accurate forecasts were not included in the input features being a potentil reason for the loss of performance. Most notably the previously mentioned solar wind dynamic pressure, derived from those plasma parameters, is playing a major role in the solar wind-magnetosphere interaction.

As mentioned earlier, plasma variables are closely related to geomagnetic indices, and incorporating them into ML models can significantly improve forecasting accuracy. However, the plasma variables, particularly those from instruments like ACE's SWEPAM, are often incomplete, especially during intense storms when measurements can be lost due to instrument saturation [20]. Incomplete or erroneous data can significantly harm the model's performance, presenting a substantial challenge in real-time forecasting where high-quality, continuous data are crucial [81]. This issue of missing or corrupted data is one of the main obstacles in the ML field, especially for time series forecasting, where



(a) SYM-H 1 hour prediction

(b) ASY-H 1 hour prediction

(c) SYM-H 2 hours prediction

(d) ASY-H 2 hours prediction

Figure 2.13: Predictions for the storm of November 2004. Extracted from Collado-Villaverde et al. [37].

data gaps typically occur at critical moments, due to instrumentation or transmission issues [20].

The problem of missing data is well-documented in both ML and SW research, with various imputation techniques proposed to address it. In the time series forecasting field, traditional imputation methods such as linear or spline interpolation, forward filling, and more advanced ML approaches like KNN and Random Forests have been widely used [82], [83]. Particularly, in SW related reserach, there are different approaches. For instance, Bhaskar and Vichare [80] used a piecewise cubic Hermite polynomial to interpolate missing values in their input variables, including IMF components and solar wind parameters, while Temerin and Li [84] used linear interpolation to standardize data intervals for *Dst* forecasting. In addition, Keesee et al.[85] applied interpolation within limits and discarded data with large gaps, while Iong et al.[38] used linear interpolation to fill gaps in SYM-H forecasting.

However, interpolation methods have limitations in real-time applications, as they require both preceding and following data points to estimate missing values. In operational settings, where future values are unknown, this makes real-time interpolation impractical. Consequently, recent real-time forecasting efforts, such as those by Siciliano et al.[36] and Collado-Villaverde et al.[37], have limited their models to use only magnetic field variables, which have near-complete availability, instead of plasma data. In the broader field of space weather, innovative imputation methods are also being developed to address data gaps in various contexts; for example, Wang et al.[86] used a GAN model to supplement 2-D soft X-ray magnetosphere images, and Yang and Shen[87], [88] employed an ANN to fill missing solar wind speed data from interplanetary scintillation observations.

In summary, while using plasma variables has the potential to improve forecasting accuracy for geomagnetic indices, their inconsistent availability remains a challenge in real-time forecasting. Developing robust imputation methods that can handle missing data dynamically in operational environments is essential for advancing SW forecasting systems.

Another current trend that is gaining importance is the interpretable ML, that is, models than can be understood in human terms. In that regard, the work of Iong et al. [38] developed a model using Gradient Boosting Machines (GBM) to predict the SYM-H index 1 and 2 hours ahead using the previous SYM-H values, IMF, solar wind and derived parameters. Since the plasma features can present missing values they linearly interpolated the missing plasma values. For the derived parameters, they used the solar wind dynamic pressure, defined as $P = \rho V_x^2 (\text{nPa})$ and the rectified electric field, defined as $E_s = \max(0, -|V_x| B_z)\,(mV/m)$. Additionally, since they worked with trees, the contribution for each feature could be calculated using Shapley additive explanation (SHAP) [89]. Applying that method, they concluded that the density and velocity of the solar wind had a larger independent contribution over the dynamic pressures. They also considered the rectified electric field as an input feature but had a lesser predictive value. Their model obtained a lower RMSE on the SYM-H index on 1 hour prediction but their performance on the 2 hours forecast was worse than previous works.

2.5 Conclusions

The primary objective of this PhD is to develop a real-time forecasting system for specific geomagnetic indices using ML techniques. Traditionally, this problem has been addressed through physics-based models that aim to mathematically explain and predict geomagnetic indices. However, with the increasing availability of data and advancements in ML, the current trend has shifted toward using ML models for forecasting.

In particular, ANNs have emerged as the dominant technique in recent works, focusing on the forecasting of SYM-H and ASY-H indices due to their higher temporal resolution, which complements the capabilities of ANNs. Research efforts are primarily directed towards predicting the indices during intense geomagnetic storms for short-term forecasts (one to two hours ahead), providing sufficient time for preventive actions

Nevertheless, the final objective is the development of a real-time system, which raises some additional considerations to be taken into account and creates additional problems. Not all necessary features, especially plasma-related data, are fully available in real-time. Most current models therefore rely solely on IMF variables and the indices themselves. The main drawback is that plasma features, which are highly correlated with magnetospheric disturbances, especially during storms, are not used, leading to a loss of accuracy. Another challenge is the potential delay in data transmission, whether from spacecraft reception or processing of incoming data.

Taking the previous issues into account, we will focus on properly deal with the missing values, caused both by the saturation of the instruments and delays in the transmission. Using the extra information provided by the plasma variables, the forecasting accuracy will be increased.

Chapter 3

Neural Network architecture for SYM-H and ASY-H forecasting

"The Sun, with all those planets revolving around it and dependent on it, can still ripen a bunch of grapes as if it had nothing else in the universe to do."
— Galileo Galilei

This chapter details the technical implementation of the forecasting model developed for this dissertation, emphasizing the importance of designing DNN models capable of accurate, real-time forecasting of SYM-H and ASY-H indices in operational settings. Special attention is given to the operational evaluation and the unique requirements of deployment in real-world environments. The models are designed to predict the geomagnetic indices one and two hours ahead, addressing the demands of both laboratory testing and the challenges anticipated in their operational use.

For a fair and comprehensive comparison, we utilized the same set of geomagnetic storms documented in previous studies [36], [38]. However, we enhanced the quality of the training data by incorporating SWICS data to address missing values in the SWEPAM dataset. This integration of additional data sources improves the training set, enhancing the model's forecasting capabilities. As a result, this approach yields a more robust and reliable forecasting tool, capable of improved performance in predicting geomagnetic storm activities, even in situations when SWICS data is not available.

Additionally, a preliminary version of the SYM-H forecasting model was deployed for real-time operation and is accessible at http://www.senmes.es/pub/ISG/lastSYMforUAH.png. This model successfully predicted the moderate geomagnetic storm on November 5th 2023, as detailed in Section 3.6.

3.1 Database

Following the current literature, we have used the 42 geomagnetic storms that occurred between 1998 and 2018 for which the SYM-H reached values lower than -100 nT, shown in Tables 3.1–3.3. Following the best practices of ML projects, the data needs to be split into three subsets: train, validation and test. To properly split the storms into the three subsets, we follow the proposal of Siciliano et al. [36] which has also been followed by subsequent works [37], [38]: the training set consists of 20 geomagnetic storms (48% of the storms, Table 3.1), 5 storms are used for the validation set (12%, Table 3.2), and the remaining 17 storms are reserved for testing the network (40%, Table 3.3). This split has been done so every subset is evenly populated regarding the geomagnetic storm complexity and intensity. Following the same split as previous works also eases comparison of the different models.

We use the following input features for the network to forecast the SYM-H and ASY-H indices:

- The Interplanetary Magnetic Field (IMF): the magnitude of the Magnetic Field and its X, Y and Z components in Geocentric Solar Magnetospheric (GSM) coordinates, all expressed in nT.
- Plasma features: the proton density (ρ) expressed in cm^{-3}, the X component of the proton speed (Vx) expressed in km s^{-1}, and the proton temperature (T) expressed in Kelvin.
- Derived parameters: we also use derived features: the solar wind dynamic pressure and the electric field. They are relevant features to forecast the geomagnetic index as showcased by Iong et al. [38].
- Geomagnetic index: the SYM-H or the ASY-H index, depending on which one we are forecasting, expressed in nT. The indices are the variables that we are forecasting, but we will also use the indices' values up to the time of the prediction as another input feature.

The use of those features is motivated by their relationship to the geomagnetic indices, considering the solar wind magnetosphere-coupling [90]. Particularly, the Bz component of the IMF plays a major role in the development of the geomagnetic storms, showcasing a positive correlation with the SYM-H index [91]. Some plasma features also have a close relationship to the geomagnetic indices [92]–[94] or can give useful information for the prediction. For example, the solar wind speed is useful to estimate the time when the storm would hit the Earth, despite not having a very strong direct correlation with the index itself [95]. Moreover, other studies [93] show a close relationship between the derived parameters, particularly the electric field with the Dst index, which is considered similar to the SYM-H, albeit with a lower time resolution.

Following the current research in the field, we work with 5-minute averages of both the input data and the forecasting index. For the solar wind data, we use the measurements taken by the ACE spacecraft because it measured of all the selected storms. Newer spacecrafts, such as DSCOVR (launched on 2015), have not yet gathered enough data to be used in ML projects. All the averages are done within a closed right approach to

Table 3.1: Storms used to train the DNN models. From left to right: number used to identify the storm, start and end days, occurrence (Y) or not (N) of a multi-dip (MP) storm, SYM-H index minimum value and % of missing plasma values in the SWEPAM dataset and correlation between the SWEPAM and SWICS datasets.

#	Date start (DD/MM/YYYY)	Date end (DD/MM/YYYY)	MP	Min SYM-H (nT)	Plasma features					
					Density		Speed		Temperature	
					% m	Corr	% m	Corr	% m	Corr
1	14/02/1998	22/02/1998	Y	-119	24.73	0.926	24.73	0.967	24.73	0.798
2	02/08/1998	08/08/1998	Y	-168	0.15	0.907	0.15	0.956	0.15	0.952
3	19/09/1998	29/09/1998	N	-213	0.32	0.907	0.32	0.974	0.03	0.921
4	16/02/1999	24/02/1999	Y	-127	0.35	0.908	0.39	0.993	0.35	0.939
5	15/10/1999	25/10/1999	N	-218	8.11	0.881	8.11	0.997	0.09	0.940
6	09/07/2000	19/07/2000	N	-335	14.80	0.925	14.80	0.993	14.80	0.941
7	06/08/2000	16/08/2000	Y	-235	2.65	0.892	2.65	0.992	2.65	0.906
8	15/09/2000	25/09/2000	Y	-196	1.70	0.881	1.74	0.996	0.07	0.908
9	01/11/2000	15/11/2000	Y	-174	11.60	0.891	11.60	0.995	9.42	0.934
10	14/03/2001	24/03/2001	Y	-165	44.88	0.854	44.88	0.993	0.29	0.913
11	06/04/2001	16/04/2001	N	-275	0.35	0.881	0.35	0.992	0.32	0.928
12	17/10/2001	22/10/2001	N	-210	43.63	0.862	43.63	0.996	0.35	0.891
13	31/10/2001	10/11/2001	N	-313	35.39	0.946	35.39	0.996	10.35	0.862
14	17/05/2002	27/05/2002	Y	-113	2.24	0.968	2.24	0.988	0.22	0.840
15	15/11/2003	25/11/2003	N	-488	1.01	0.861	1.01	0.994	1.01	0.921
16	20/07/2004	30/07/2004	Y	-208	3.69	0.894	0.16	0.992	0.13	0.910
17	10/05/2005	20/05/2005	N	-302	0.76	0.906	0.82	0.992	0.76	0.731
18	09/04/2006	19/04/2006	N	-110	2.27	0.949	1.80	0.997	0.32	0.962
19	09/10/2006	19/12/2006	N	-206	7.32	0.893	7.32	0.994	6.98	0.941
20	01/03/2012	11/03/2012	Y	-149	83.43	0.974	26.01	0.989	20.90	0.838

ensure that no information flows from the future to the past. That is, the average for the minute 5 will be calculated using the data between after minute 0 until exactly minute 5. The values after 5:00 will be used in the average for minute 10.

All the datasets for the different features are downloaded from NASA's Coordinated Data Analysis Web (CDAWeb). For the index, we retrieve the data from the `OMNI_HRO_5MIN` dataset. Meanwhile the SYM-H can be obtained from OMNI for the training and testing of the model, for its operational deployment we compute it in real time following the procedure proposed by Nahayo et al. [96]. For the IMF data we use the level-2 data measured by the Magnetic Field Experiment instrument on ACE. We will use the 16 second averages, specifically the `AC_H0_MFI` dataset. Later on, we group the measurements into 5 minutes averages, the same resolution as the index. For the plasma features, we retrieve the data from both the SWEPAM and SWICS instruments. The level-2 `AC_H0_SWE` SWEPAM dataset provides the proton density, speed and temperature on 64-second averages. For the SWICS measurements, we use the level-2 `AC_H6_SWI` dataset which contains measurements of the solar wind proton density, the solar wind proton speed and the solar wind proton thermal speed in 12-minutes averages. The SWICS dataset will be used to fill the missing values in the SWEPAM dataset as described in Section 3.2. For the derived parameters, following Iong et al. [38] findings, we use y component of the electric field: $E_y = V_x B_z$, where V_x is the x component of the bulk proton speed and B_z is the z component of the IMF. This derived parameter is related to the magnetic flux in the north-south direction; as evidenced by Iong et al. [38] is one of the most useful features to forecast the SYM-H index. The other derived parameter is the dynamic pressure: $P = \rho V_x^2$. The usage of the pressure is motivated because, on

Table 3.2: Storms used to validate the DNN models. From left to right: number used to identify the storm, start and end days, occurrence (Y) or not (N) of a multi-dip (MP) storm, SYM-H index minimum value and % of missing plasma values in the SWEPAM dataset and correlation between the SWEPAM and SWICS datasets.

#	Date start (DD/MM/YYYY)	Date end (DD/MM/YYYY)	MP	Min SYM-H (nT)	Plasma features					
					Density		Speed		Temperature	
					% m	Corr	% m	Corr	% m	Corr
21	28/04/1998	08/05/1998	N	-268	12.09	0.942	12.09	0.988	0.60	0.863
22	19/09/1999	26/09/1999	N	-160	0.00	0.910	0.00	0.985	0.00	0.897
23	25/10/2003	03/11/2003	Y	-427	32.01	0.961	32.01	0.991	32.01	0.491
24	18/06/2015	28/06/2015	Y	-207	36.99	0.890	6.60	0.983	2.78	0.821
25	01/09/2017	11/09/2017	Y	-144	25.92	0.870	15.47	0.978	12.82	0.871

Table 3.3: Storms used to test the DNN models. From left to right: number used to identify the storm, start and end days, occurrence (Y) or not (N) of a multi-dip (MP) storm, SYM-H index minimum value and % of missing plasma values in the SWEPAM dataset and correlation between the SWEPAM and SWICS datasets.

#	Date start (DD/MM/YYYY)	Date end (DD/MM/YYYY)	MP	Min SYM-H (nT)	Plasma features					
					Density		Speed		Temperature	
					% m	Corr	% m	Corr	% m	Corr
26	22/06/1998	30/06/1998	N	-120	0.54	0.893	0.54	0.992	0.54	0.913
27	02/11/1998	12/11/1998	Y	-179	0.57	0.765	0.57	0.951	0.57	0.893
28	09/01/1999	18/01/1999	N	-111	1.49	0.944	1.49	0.985	1.49	0.841
29	13/04/1999	19/04/1999	N	-122	0.25	0.893	0.25	0.960	0.25	0.888
30	16/01/2000	26/01/2000	Y	-101	12.66	0.941	12.66	0.976	0.13	0.829
31	02/04/2000	12/04/2000	N	-315	0.57	0.925	0.57	0.996	0.28	0.920
32	19/05/2000	28/05/2000	Y	-159	0.62	0.937	0.62	0.994	0.62	0.930
33	26/03/2001	04/04/2001	N	-434	4.17	0.826	4.17	0.995	0.22	0.894
34	26/05/2003	06/06/2003	Y	-162	0.09	0.903	0.09	0.994	0.09	0.905
35	08/07/2003	18/07/2003	Y	-125	1.48	0.863	1.48	0.995	0.09	0.923
36	18/01/2004	27/01/2004	Y	-137	0.03	0.913	0.03	0.991	0.03	0.941
37	04/11/2004	14/11/2004	Y	-393	4.70	0.955	4.70	0.997	0.03	0.956
38	10/09/2012	05/10/2012	N	-138	92.57	0.934	39.32	0.990	7.36	0.854
39	28/05/2013	04/06/2013	N	-134	4.30	0.962	0.48	0.996	0.48	0.926
40	26/06/2013	04/07/2013	N	-110	13.31	0.918	0.50	0.993	0.50	0.934
41	11/03/2015	21/03/2015	N	-233	59.00	0.910	7.26	0.994	0.35	0.877
42	22/08/2018	03/09/2018	N	-205	29.43	0.924	2.43	0.996	0.13	0.936

most storms, the dynamic pressure presents a substantial increase prior to the onset of storm, compressing the magnetosphere. This phenomenon is reflected on the SYM-H by a positive increase right at the beginning of the storm, known as sudden storm commencement [97]. Previous models that did not use the plasma features failed to forecast this sudden spike in the SYM-H [37].

3.2 Imputation of the plasma features

One of the most important issues regarding the usage of plasma variables in a forecasting system is that they are highly susceptible of being missing, especially during intense storms [98]. For instance, Figure 3.1 depicts the absence of data when the SWEPAM instrument stopped providing measurements on the onset of the storm. This issue is of great importance because it usually happens when the forecasting is of utmost importance [99].

Due to this issue, previous works [36], [37] have only used the IMF data and the previous values of the index to perform the forecast. Particularly, the proton temperature and density are the features that have the highest rate of being missing, as noted by Larrodera and Cid [20]. Both features showed an increased amount of invalid data as the detectors aged. Nevertheless, an operational improvement increased the amount of scientific-quality SWEPAM measurements since October 2012. Still, a lot of storms happened before the improvement: 11 storms occurred between 2003 and the operational improvement for which there are a significant amount of missing plasma values.

Nevertheless, some works used the plasma features by interpolating the missing values, such as Iong et al. [38] and Bhaskar and Vichare [80]. While interpolation is a valid solution, plasma features during an intense storm present a high variability and methods such as linear interpolation would overlook the sudden increase in the proton speed and the high variability in the temperature, not properly filling the gaps. An example is shown in Figure 3.1 for the geomagnetic storm of November 2001 where we depicted the two most common imputation approaches: forwarding the last valid value until a new valid value is available and linearly interpolating the missing values.

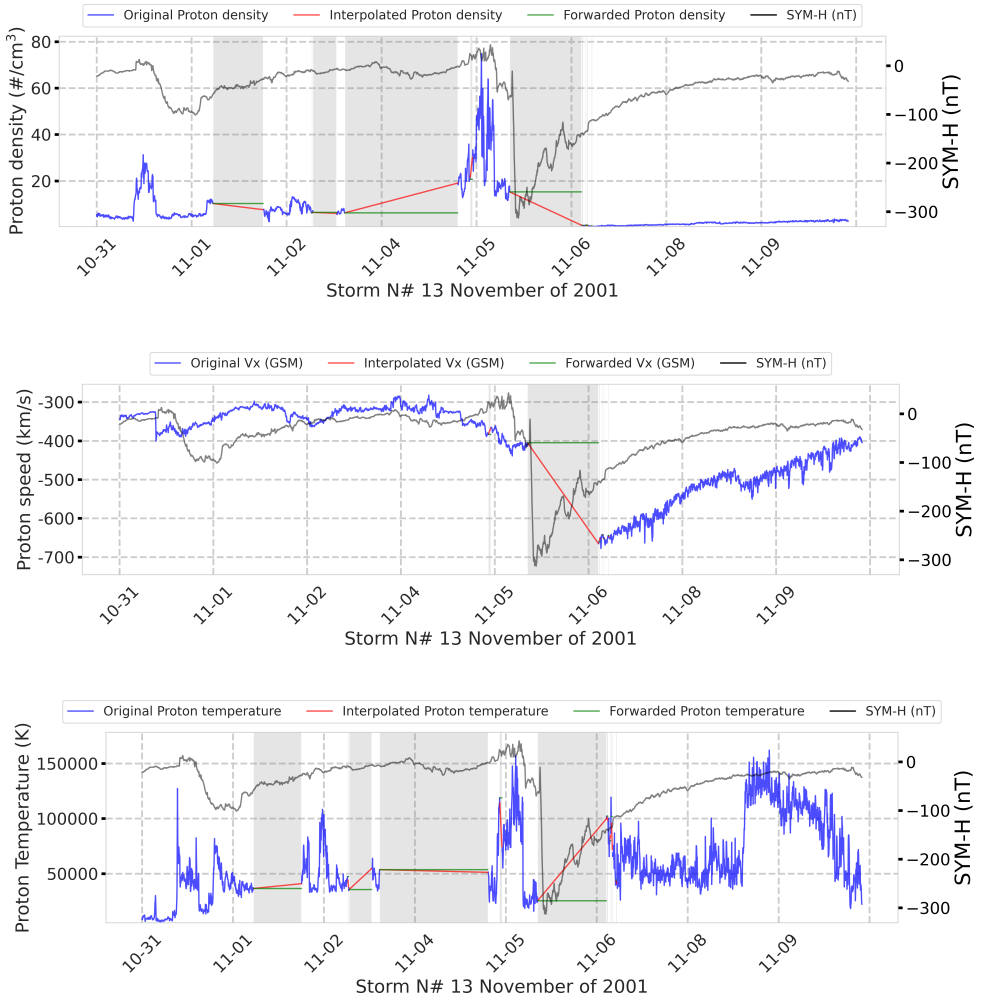


Figure 3.1: Proton density (top), proton speed (mid) and proton temperature (bottom) measured by ACE's SWEPAM for the geomagnetic storm 13, November 2001. The SYM-H index is represented in black. The time when the plasma values are missing is shaded in grey. Two imputation methods have been applied to fill the gaps: forwarding the last observed value (green) and linear interpolation (red).

To overcome that, we propose the usage of the data from another instrument onboard ACE, the SWICS instrument to fill the missing values in the SWEPAM dataset. The SWICS instrument measures the chemical and ionic-charge composition of the solar wind, the temperatures and mean speeds of all major solar-wind ions at all solar wind speeds. This dataset contains the solar wind proton number density expressed in cm^{-3}, the solar wind proton speed in km s^{-1} and the solar wind proton thermal speed in km s^{-1}. All the features are measured with a 12 minutes resolution. The proton density is measured using the same units as SWEPAM so it can be used directly to fill the missing values. Then, despite the proton temperature not being available by default, it can be calculated using the thermal speed, v_{th}, as in Equation 3.1, where T represents the proton temperature in Kelvin, k_B represents the Boltzmann constant and m represents the mass of the proton. The other difference is that SWICS only measures the bulk speed of the protons, not the individual components. Nevertheless, the negative proton bulk speed has almost perfect correlation with the x component of the bulk speed using GSM coordinates, which is the component that we use to perform the forecast, so it can be directly replaced after inverting the sign. To summarize, using the SWICS measurements we can fill the missing values on all the SWEPAM plasma features, albeit with a lower resolution. We can resample the SWICS data to 5-minute resolution and perform linear interpolation. Then, the missing values in the SWEPAM dataset can be replaced with the valid SWICS measurements.

$$T = \frac{v_{th}^2 * m}{2k_{\mathrm{B}}} \tag{3.1}$$

Using SWICS data to fill missing values can also solve some sensitivity issues in SWEPAM. According to the ACE Science Center team [100], for reasons related to accommodating the Ulysses flight spare on ACE at 1 AU, SWEPAM has reduced sensitivity at the lowest energies, and at times it underestimates the proton density, especially for very slow wind. Additionally, the solar wind velocity may also be culled, mostly during periods of both high solar energetic particle background and high wind speeds. Despite solar wind proton studies are not part of the primary science goals for SWICS, they can be used to fill in gaps in the SWEPAM data. A comparison of the data measured by SWEPAM and SWICS for the geomagnetic storm of November 2001 is depicted in Figure 3.2.

The measurements show very high correlation metrics for the all the features, being the mean correlation of the proton temperature 88.8%, 90.7% for the proton density and 98.8% for the proton speed. The individual correlations between the SWEPAM and SWICS instruments for each individual storm is shown in Tables 3.1–3.3. This merging has been done previously by the ACE Science Center in a level 3 product: the SWEPAM/SWICS Level 3 Merged Solar Wind Proton Dataset. It consists of 12-minute averages of solar wind proton data from the SWEPAM instrument, filling the gaps in the proton density, speed, and temperature using the data from SWICS. However, not all the storms used in this study are covered in that product.

The main drawback of using the SWICS data to fill SWEPAM's missing values is the different time resolution of the instruments. SWICS provides values in 12 minutes averages while SWEPAM provides them in 64 seconds averages. Resampling the 64 seconds SWEPAM measurements to the 5 minutes averages that will be used to perform the forecast can be done directly but when it is done over the SWICS measurements we need to

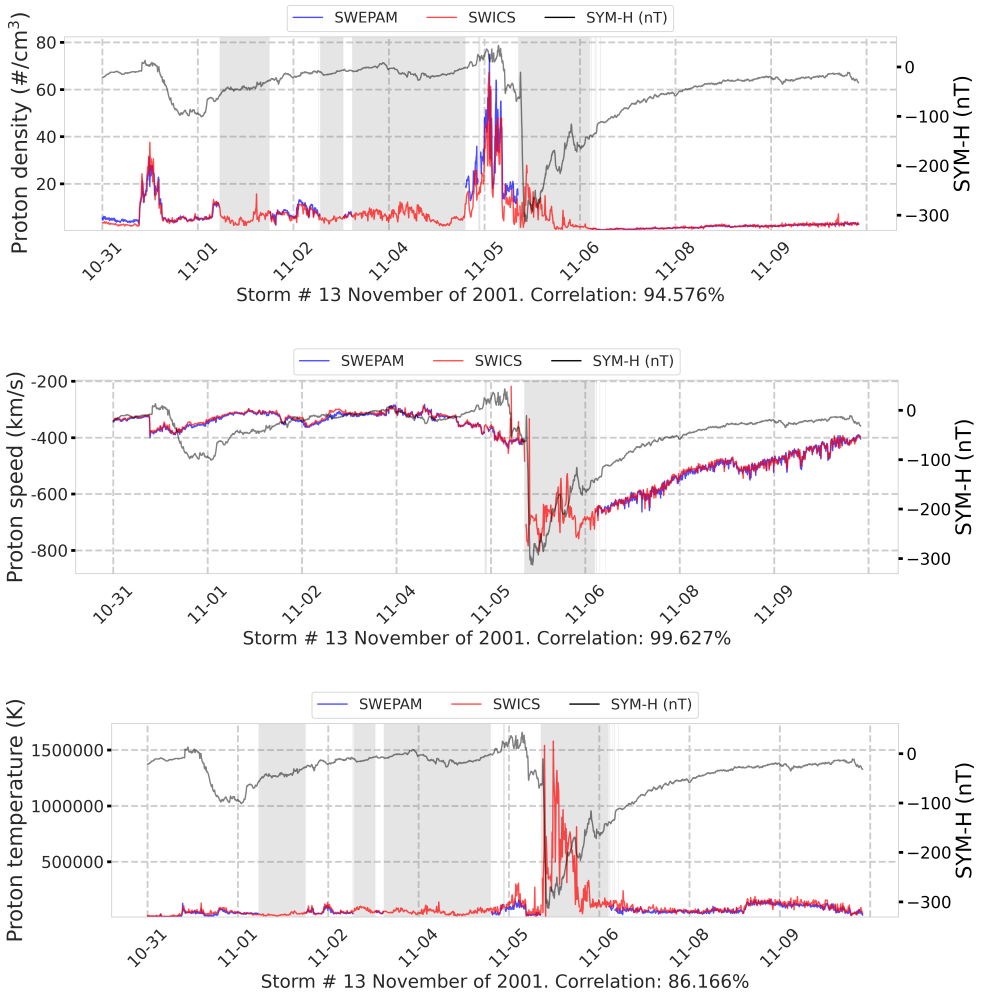


Figure 3.2: Comparison of the proton density (top), proton speed (mid) and proton temperature (bottom) measured by ACE's SWEPAM (blue) and SWICS (red) instruments for the storm 13, November 2001. In the areas shaded in gray only SWICS data is available, which will be used to fill SWEPAM's data.

perform an interpolation, incurring in some inevitable data loss. When we perform the resample operation to calculate the 5 minute averages for the SWICS instrument, similarly to the resample on the SWEPAM data, we use right-closed intervals. Nevertheless, it is still preferable over performing an interpolation between valid SWEPAM measurements. This can be seen in the proton speed and temperature of the Storm 13 in Figure 3.2. Nevertheless, there are still some storms in which the SWICS also presents missing values. In those situations we perform interpolation over the SWEPAM data if the value to interpolate to is available, otherwise we propagate the last valid measurement.

Additionally, to facilitate the convergence of the network, each variable has been standardized in a two step process:

1. The mean of each variable is subtracted, so the data is centered around 0.
2. The data is divided by the standard deviation, so the variance of each input feature is close to 1.

The process is depicted in Equation 3.2, where x are the original values of the feature, μ is the mean of the feature and σ is its standard deviation. This approach is preferred

over scaling the values between a fixed range using the minimum and maximum values. Features that have outliers with extreme values, such as the proton temperature (see Figure 3.2), would have a great impact on the global scaling of the data over using the standard deviation. Moreover, standardization is usually more practical for ML since the weights of the Neural Network (NN) layers are initialized to small values, close to 0. Furthermore, standardization can maintain useful information about outliers, being this scaling process less sensitive to them in contrast to normalization [101]. It is important to note that the plasma features have been standardized after the missing values on the SWEPAM measurements have been filled using the ones from SWICS.

$$Z = \frac{(x - \mu_x)}{\sigma_x} \tag{3.2}$$

3.3 Methodology

This section describes how we have tackled the indices forecasting problem. First, the problem is defined and then we describe the DNN architecture used to forecast the SYM-H and ASY-H indices along with the description of the training process.

3.3.1 Problem statement

The forecasting of the geomagnetic indices index is stated as a time-series forecasting problem. As described earlier, we consider the last 32 time-steps of the input features presented in Section 3.1. As each time-step aggregates 5 minutes of data, we use 160 minutes of historical data to feed the network. This history length choice is consistent with Siciliano et al. [36] findings, who stated that the R^2 and RMSE metrics started to worsen when the history length was greater than 180 minutes if the SYM-H was among the input parameters, particularly on ANNs based on CNNs and RNNs. We forecast the SYM-H and the ASY-H in the following one and two hours, using two instances of the the same DNN architecture for each time horizon. Both instances use the same input data but the outputs are the corresponding index values for each time horizon.

The decision to train different networks for each time horizon instead of only one that forecast both time horizons simultaneously is to maximize the performance on each time horizon separately. Using the same architecture to forecast both time horizons at the same time yielded lower performance on each of them and longer training time. During operational time, once both models are loaded, the forecast of a single time-step is not computationally expensive, moreover since both models have the same inputs, thus, the data preprocessing only needs to be done once.

3.3.2 Deep neural network architecture

To forecast the indices we have implemented a DNN that combines a MultiLayer Perceptron [102] (from now on will be referenced as Dense layers), single-dimensional CNN layers [103], LSTM layers [50] and Multi-Head Attention layers [51]. The individual layers are described below:

- Dense layers, also known as fully-connected layers, consist of an activation function, a weight matrix and a bias vector. They implement the operation described in Equation 3.3, where Y is the output of the layer, f is the element-wise activation function, W is the weight matrix, X is the input of the layer and b the bias vector. $\cdot$ represents the dot-product operation.

$$Y = f[(X \cdot W) + b] \tag{3.3}$$

- The single-dimensional CNNs layers are designed to process temporal data and to capture the relationships of the features between different time-steps. They achieve this by processing several time-steps at the same time, being the number of time-steps processed defined by the *kernel* parameter. Although its most common usage is the two-dimensional CNN variant applied to image processing, it has been successfully applied to time-series processing [104].
- The LSTM layers are a specialized RNN layer. In general, RNN layers try to capture temporal relationships by recursively processing each time-step, saving the relevant information from each time-step and combining them with the previous ones. LSTMs are variation of the RNN layers that aim to learn long-time dependencies over the time-series. The LSTM layers have been used in a wide variety of time-series processing tasks, such as forecasting [105], classification [106] and language processing [107]. The LSTM and CNN layers have also been used in previous works to forecast geomagnetic indices [36], [37], where their computations have already been explained. We encourage the reader to consult those articles for further information.
- The MultiHead Attention Layer was proposed by Vaswani et al. [51] and originally applied for the Neural Machine Translation problem. It is an evolution of the original attention mechanism and the Dot-product attention described on the work of Bahdanau et al. [108]. Since then, it has been widely used in sequence processing problems. The main idea is to mimic human's cognitive attention. To do that the layer re-weights the input data, increasing the values of the important elements and decreasing the irrelevant ones.

  The Attention Layer has three inputs: a Query Q, Key K and Value V vectors. First, it applies the equation shown in Eq 3.4. Then, the dot product is applied on the query and key projections, creating a context vector. Then, the result is scaled down by $\sqrt{d_k}$ and the softmax function is applied to obtain the attention probabilities. Finally, the value vector is multiplied by the attention probabilities obtaining the final re-weighted value vector. Most of the time the key and value vectors are the same.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{3.4}$$

This process allows the re-weight the value vector with respect to the context calculated using the query and key vectors. An evolution of this idea is the concept of Self-Attention; in this case all the three input vectors are the same $Q = K = V = X$ where X is the input of the attention layer, so the sequence can be re-weighted with respect to itself. Therefore, the intrinsically important parts are enhanced and the less important ones, diminished.

In addition to the concept of Self-Attention, Vaswani et al. [51] also introduce the concept of Multi-Head Attention. The idea is to perform the Attention operation h times in parallel over the sequence (where h is the number of heads) instead of performing only one Attention operation. In this case, the three input vectors are projected using dense layers, the Query vector will be projected to a vector with dimensionality of d_q, the key vector to a dimensionality of d_k and the value vector to a dimensionality of d_v. Once all the heads have performed the re-weight, the obtained values are concatenated and projected one final time. This approach enables the model to attend to information from different subspaces located at different positions at the same time. Multi-Head Attention is described in Equation 3.5. The projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

$$\text{MultiHeadAttention}(Q, K, V) = \text{Concat}(head_1, head_2, ..., head_h) \cdot W^O \tag{3.5}$$

$$\text{where } head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.6}$$

$$\text{for } i \in 1, 2, ..., h$$

Focusing on our proposed architecture, depicted in Figure 3.3, the initial input time-series is processed through 3 single-dimensional CNN layers, all of them are configured with a kernel size of 7, causal padding and a stride of 1. Then, the resulting series is re-weighted using a MultiHead Attention layer and added to the previous result. This operation of adding the result of a layer with its input is a common practice known as "residual connections" and is commonly used with convolutional layers for image processing [109] and for time-series processing with Attention layers [51], [110]. However, it is important to address a key difference between our model and other works with similar architectures: after the MultiHead Attention layers and the residual addition, a Layer Normalization operation is usually applied. This operation holds the mean of the features close to 0 and the standard deviation close to 1. For our particular case it actually harms the performance of the network. The key aspect of forecasting the geomagnetic indices is the relevance of the outliers compared to other cases, making the usage of the Layer Normalization not advisable.

After that, we make use of a Bidirectional LSTM layer. In this case, the sequence is processed in both directions and the output will be all the hidden states of the layer instead of only the last one. We decided to add the backward processing because of one particular behaviour of the geomagnetic indices during the storms: there is usually a sudden increase in the dynamic pressure and, most notably, on the SYM-H before it starts to rapidly decrease. Since these phenomenons happen just before the sudden fall of the index, we believe that the backward processing can help the network to identify them and take them into account in the final prediction.

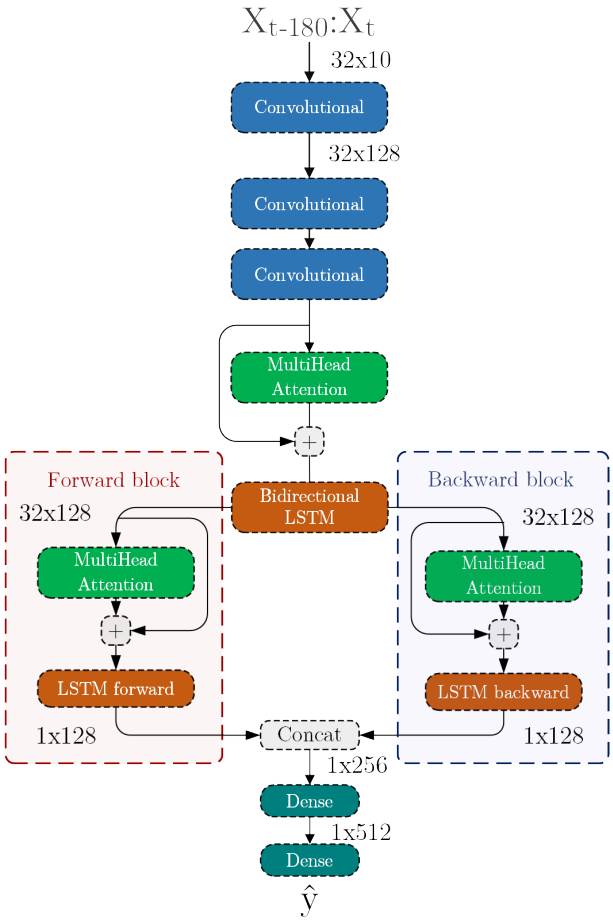


Figure 3.3: DNN architecture for the SYM-H and ASY-H forecast. The input shape is expressed in t ± minutes. The shapes between layers are expressed in time-steps × features.

After the Bidirectional LSTM, the network is split into two similar branches, one for each of the resulting series, represented as the Forward (red) and Backward (blue) blocks depicted in Figure 3.3. Each block implements a MultiHead Attention layer followed by a residual connection. With the MultiHead Attention layer the output of each direction from the previous LSTM layer can be re-weighted, eliminating the unnecessary information. Then, we use another LSTM layer, this time in a single direction. Using several LSTM layers stacked is a common practice that has been widely applied to several fields, such as traffic [111] or stock market [112] prediction. From these two LSTM layers, named LSTM forward and backward, we only keep the last state. That is, the resulting vector from processing the whole sequence.

The result of both blocks will be then concatenated and processed by two final dense layers that will produce the final forecast of the index. This approach of separating the forward and backward processing of the time series to, later on, use them together has also been used in time-series with missing values using attention, particularly in imputation tasks [113].

The hyperparameters have been selected as the best performing ones from the following spaces:

- A dimensionality of 128 for all the layers except the last two dense layers; the first one has 512 units and the last one that produces the final forecast of the index has only one. The search space for the dimensionality of all the layers was [32, 64, 128, 256] and [128, 256, 512, 1024] for the second-to-last dense layer. Which is still big considering the available training data points.
- The kernel size of all the convolutional layers is set to 7 and the chosen activation function is the ELU function, described as in Equation 3.7. We have tested a kernel size of [5, 7, 9] and the ReLU and ELU activation functions.
- The number of heads for all the MultiHead Attention layers is set to 4. We have tested for 2, 4 and 8 heads.
- The activation function for the first dense layer is also ELU and the last one has no activation function. We have tested the ReLU and ELU activation functions for the first dense layer.

$$ELU(z) = \left\{ \begin{array}{cc} z & z > 0 \\ (e^z - 1) & z <= 0 \end{array} \right\} \tag{3.7}$$

Considering the selected hyperparameters, the network consists of 1.688.577 trainable parameters.

The main difference between our proposed architecture and previous works based on NNs is the usage of the Attention mechanism and the plasma features. Previous models did not use plasma features due to their risk of being missing, which compromises the performance of the network due to their reliance on them. Real-time measuring devices often contain noise, especially during extreme events, or missing values, particularly the plasma, that can negatively impact the performance of RNNs. Attention mechanisms enable the model to focus on the most relevant and reliable parts of the sequence, mitigating the impact of missing data assigning higher weights to the more complete data points. Consequently, our attention-based model demonstrates improved robustness and achieves better predictive capabilities, even in a real-time scenario when there are missing values. Moreover, since the Attention computes the time-series in parallel, they can attend to multiple parts of the sequence simultaneously, being able to better capture long-term dependencies, whereas RNNs process the time series data sequentially. This difference also helps dealing with the vanishing gradient problem [114] which RNNs suffer making it difficult for them to capture long-term dependencies.

3.3.3 Training and validation

We have trained two instances of the proposed architecture on each of the time horizons, 1 and 2 hours ahead. One has only been trained using SWEPAM's data and the missing values have been interpolated. The other one has been trained using SWICS to fill SWEPAM's missing values; if there are missing values for both instruments, SWEPAM's data is interpolated.

The trained models are tested in a simulated operational environment. The operational simulation consist of only using IMF and SWEPAM data, since those are the only

instruments that provide the measurements in real-time. Moreover, the missing values in the SWEPAM dataset can only be interpolated if they are surrounded by valid values. If there is no later valid value at the time of the forecast, the last valid measurement will be propagated forward.

Additionally, we have considered two extra scenarios, which represent laboratory conditions where all the data is available at the time of the forecast, filling SWEPAM missing values with SWICS in the test storms. Summarizing, we have the following six scenarios for each index:

- SWEPAM operational 1h: it forecasts the index in the following hour. Trained only with SWEPAM data and evaluated on the operational scenario.
- SWICS operational 1h: it forecasts the index in the following hour. Trained with SWEPAM data filled using SWICS and evaluated on the operational scenario.
- SWICS laboratory 1h: it forecasts the index in the following hour. Same model as the previous case but evaluated on laboratory conditions.
- SWEPAM operational 2h: it forecasts the index two hours ahead. Trained only with SWEPAM data and evaluated on the operational scenario.
- SWICS operational 2h: it forecasts the index two hours ahead. Trained with SWEPAM data filled using SWICS and evaluated on the operational scenario.
- SWICS laboratory 2h: it forecasts the index two hours ahead. Same model as the previous case but evaluated on laboratory conditions.

As we pursue an operational forecast, we have evaluated the model trained with both imputation procedures in a simulated operational settings, based on the hypothesis that having more instances of complete training data can help the model to perform better even when it is used with non-complete real time data. This evaluation is relevant because despite SWICS data is not available in real-time, it can improve the final performance on the model when used during the training process. Moreover, the extra scenarios are evaluated to present the performance that could be achieved under ideal circumstances.

For the training and validation process we have followed the procedure done in previous works: the DNN is trained on the 20 training storms and after each epoch the performance is evaluated on the validation set. To avoid overfitting, the weights that achieved the best performance on the validation set are saved when the performance is improved. The training is then stopped if the performance has not improved over the last 15 epochs, restoring the weights that obtained the best validation loss. To optimize the weights of the DNN we have chosen the AdaBelief optimizer [115]. It is a variant of the Adam optimizer used in previous works [36], [37]. AdaBelief scales the learning rate by the difference between the predicted and observed gradients and it has achieved a better generalization than the base one, which is the main obstacle when the forecasting has a great amount of outliers.

The hyperparameters selected for the optimizer are the same for both models and set as follows:

- The learning rate at which the weights are optimized is fixed to 10^{-3}.
- The minimum admissible learning rate has been set to 10^{-5}.
- The weight decay for each parameter is set to 10^{-4}.
- The exponential decay rate for the first moment estimates, β_1, has been set to 0.9.
- The exponential decay rate for the second moment estimates, β_2, has been set to 0.999.
- ϵ a small constant to provide numerical stability and to avoid divisions by zero is set to 10^{-14}.

The DNNs have been implemented using Keras [116], an open source library that provides a Python interface for the development of ANNs. It also acts as an interface for the TensorFlow library [45].

3.4 Model evaluation for the SYM-H index

In this section we present a comparison of the predictive capabilities for the proposed model on the defined scenarios for the SYM-H index. Also, we compare our forecasts to previous works, particularly, the GBMs made by Iong et al. [38], the LSTM-CNN of our previous work [37], the LSTM based model by Siciliano et al. [36] and the Burton equation [117]. The metrics for those previous works are extracted from the papers and we also compare the predictions to the persistence model as a baseline. The persistence model uses the last value of the index at time t as the prediction for the next one and two hours. To compare the performance of the models we provide the Root RMSE and the Forecast Skill Score (FSS). We have chosen these metrics because they are commonly used in time-series forecasting and are the metrics reported by the most recent SYM-H index forecasting works, so we can fairly compare different models. They were also recommended in Camporeale's survey [57].

The RMSE is defined in Equation 2.1, where y is the observed SYM-H index and $\hat{y}$ is the forecasted value. The FSS is a complementary metric proposed by Murphy [118] that compares the performance of a model against another one selected as the baseline. Similarly to Iong et al. [38] we will use the MSE as the comparison metric against the baseline as defined in Equation 3.8, where $\hat{y}_b$ represents the predictions from the model considered as the baseline. The FSS can yield values between 1 and $-\infty$. If the metric is positive, it means that the selected model performs better than the baseline, being higher values better. Negative values mean that the selected model performs worse than the baseline, with lower values being worse. We have selected as the baseline the Burton equation, described in O'Brien and McPherron [117]. The Burton equation forecasts the Dst index in real time using the Quicklook Dst and the ACE real time solar wind measurements. This comparison was also included in the GBM work by Iong et al. [38].

$$\mathrm{FSS}(y, \hat{y}, \hat{y}_b) = 1 - \frac{\mathrm{MSE}(y, \hat{y})}{\mathrm{MSE}(y, \hat{y}_b)} \tag{3.8}$$

Tables 3.4 and 3.5 present the RMSE and the FSS values compared to the Burton equation for the 1 hour SYM-H predictions on the selected test storms (see Table 3.3). For this forecast, the proposed model evaluated under laboratory conditions, using SWICS to fill the SWEPAM missing values, achieves the best performance on all the test storms except storms 28, 29, 33 and 36, being the second best on storms 28, 29 and 36 and the third best on storm 33. The same model evaluated on the operational scenario overall performs the second best, even performing better than the model evaluated with SWICS data on storm 33. In that particular storm the SWICS has a slightly lower correlation compared to the other storms. The same model trained using only SWEPAM data performs worse than the one trained using SWICS, but still better than the other previous models, it is even the best model for storms 29 and 36, and the second best for storms 27, 30 and 38. Notwithstanding, all those storms are not very intense, such as storms 31, 33 and 37, in which the models trained with SWICS perform considerably better.

Table 3.4: RMSEs for 1-hour forecast over the test storms for the SYM-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	GBM	LSTM-CNN	LSTM	Persistence	Burton
26	5.676	<u>5.355</u>	**5.171**	5.863	6.630	6.700	7.631	6.839
27	<u>7.652</u>	7.655	**7.455**	7.729	8.913	8.900	9.623	7.955
28	5.209	4.945	<u>4.697</u>	**4.281**	5.858	5.400	5.814	5.697
29	**4.796**	5.218	<u>4.941</u>	5.833	6.683	7.200	7.174	6.511
30	<u>3.934</u>	4.156	**3.919**	4.927	5.200	5.600	4.810	4.614
31	6.778	<u>6.707</u>	**6.700**	8.277	8.584	10.700	10.429	8.838
32	7.076	<u>6.654</u>	**6.357**	6.841	7.259	8.300	10.528	9.487
33	14.810	<u>14.211</u>	14.217	14.492	**13.340**	16.300	21.167	16.630
34	10.511	10.215	**9.777**	10.190	<u>10.034</u>	11.300	10.913	10.888
35	6.353	<u>6.197</u>	**6.089**	7.154	7.693	8.500	8.011	7.918
36	**7.522**	7.737	<u>7.594</u>	8.512	9.525	8.700	9.708	9.082
37	13.295	<u>11.528</u>	**11.282**	14.548	15.184	17.500	19.698	15.713
38	<u>3.864</u>	4.054	**3.634**	3.886	4.080	4.200	4.482	4.572
39	5.321	<u>4.696</u>	**4.598**	5.901	6.431	5.600	7.597	6.663
40	4.333	<u>4.216</u>	**4.165**	4.976	4.673	5.500	5.057	5.371
41	7.585	7.877	**7.513**	<u>7.558</u>	7.882	9.000	9.984	8.358
42	4.947	<u>4.873</u>	**4.867**	5.030	5.669	5.900	6.036	5.549
Mean	7.039	<u>6.841</u>	**6.646**	7.412	7.861	8.547	9.354	8.276

On average, the model evaluated under laboratory conditions performs 2.93% better than the same model evaluated on the operational scenario. However, the SWICS operational model improves the performance of the SWEPAM operational model by 2.81%, supporting our hypothesis that using SWICS data during training enhances the learning process of the model. In this regard, the extra training samples filled using SWICS improve the capabilities of the model, even if it will be evaluated on incomplete data.

Comparing the SWICS operational model to the GBM [38], which is the only ML model that uses plasma among the input features, our best operational model performs 7.70% better than GBM, reducing the average RMSE to 6.841 nT and obtaining better predictions in 13 out of the 17 test storms. It is important to remark that in our operational simulation, we only perform interpolation in the missing plasma data if the value to interpolate to is known at the time of the prediction, otherwise we propagate the last valid observation. Instead, the GBM model is evaluated using interpolation, without considering the operational restrictions.

Table 3.5: Forecast Skill Scores (Compared to the Burton Equation) as the Baseline for 1-hour forecast over the test storms for the SYM-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	GBM	LSTM-CNN	LSTM
26	0.170	<u>0.217</u>	**0.244**	0.143	0.031	0.020
27	<u>0.038</u>	0.038	**0.063**	0.028	-0.120	-0.119
28	0.086	0.132	<u>0.175</u>	**0.249**	-0.028	0.052
29	**0.263**	0.199	<u>0.241</u>	0.104	-0.026	-0.106
30	<u>0.147</u>	0.099	**0.151**	-0.068	-0.127	-0.214
31	0.233	<u>0.241</u>	**0.242**	0.063	0.029	-0.211
32	0.254	<u>0.299</u>	**0.330**	0.279	0.235	0.125
33	0.109	<u>0.145</u>	0.145	0.129	**0.198**	0.020
34	0.035	0.062	**0.102**	0.064	<u>0.078</u>	-0.038
35	0.198	<u>0.217</u>	**0.231**	0.096	0.028	-0.074
36	**0.172**	0.148	<u>0.164</u>	0.063	-0.049	0.042
37	0.154	<u>0.266</u>	**0.282**	0.074	0.034	-0.114
38	<u>0.155</u>	0.113	**0.205**	0.150	0.108	0.081
39	0.201	<u>0.295</u>	**0.310**	0.114	0.035	0.160
40	0.193	<u>0.215</u>	**0.225**	0.074	0.130	-0.024
41	0.092	0.058	**0.101**	<u>0.096</u>	0.057	-0.077
42	0.108	<u>0.122</u>	**0.123**	0.094	-0.022	-0.063
Mean	0.149	<u>0.173</u>	**0.197**	0.104	0.050	-0.033

We can highlight that using the SWICS during training data greatly reduces the RMSE on the storm 37, the second most intense in which the SYM-H reached values of almost -400 nT, reducing the forecasting error over the previous best model, the GBM, by 22.45%. Another storm worth reviewing is the storm 32. This is the storm that presents the highest improvement over the Burton equation, achieving 0.330 in the FSS metric. Despite not being a particularly intense storm (the SYM-H reached a minimum value of -159), it is still a fairly complex storm due to its multi-dip nature.

Tables 3.6 and 3.7 report the RMSE and FSS values for the 2 hours SYM-H forecasts. This problem is more complex than forecasting the next hour. This is reflected by the higher error metrics across the storms. In this case, the proposed model using SWICS data and evaluated under laboratory conditions performs the best on 12 storms, the second best on 3 of the 5 remaining storms, and the third best on the last two. Nevertheless, it still performs the best in the overall score.

Comparing the model evaluated under laboratory conditions to the operational scenarios, it performs 2.58% better than the SWICS operational model and 6.41% better than the SWEPAM based one. Nevertheless, both models operational perform better than previous ones in the overall computation of the metrics.Once again, the usage of SWICS data during training improves the overall performance of the model. In this case, since the complexity of the problem is greater than the one hour ahead forecast, the additional training samples have a larger impact, improving more the overall performance by 3.73%.

For this forecast, generally ANN-based models perform better than other approaches, like the GBM model and the Burton equation. This is best reflected on the FSS presented in Table 3.7; all the best improvements over the Burton equation are achieved by ANN-based models, and the only 2 storms in which the GBM model is best performing one (storms 27 and 34) are not very intense, reaching minimum values no less than -180 nT.

Table 3.6: RMSEs for 2-hours forecast over the test storms for the SYM-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	GBM	LSTM-CNN	Persistence	Burton
26	8.525	8.144	**7.937**	8.285	8.899	12.374	10.690
27	12.285	12.144	11.765	**11.585**	13.418	15.387	12.465
28	5.400	5.491	**5.236**	5.650	5.877	9.331	8.858
29	8.268	8.541	**7.953**	8.826	9.314	11.415	9.776
30	6.326	6.040	**5.695**	7.280	7.288	7.416	6.266
31	9.131	8.738	**8.738**	12.613	12.436	17.193	13.604
32	9.343	9.708	9.457	9.927	**8.937**	15.282	13.766
33	22.837	18.930	18.926	24.519	**18.481**	33.927	25.729
34	**13.683**	14.945	13.800	13.736	13.941	15.109	14.695
35	9.095	8.834	**8.725**	9.504	9.932	11.211	10.586
36	10.789	10.697	**10.571**	12.068	12.058	14.687	13.117
37	19.807	18.788	**18.778**	22.327	21.084	30.582	24.446
38	5.201	5.281	**4.955**	5.153	5.213	7.353	6.546
39	6.691	6.076	**6.053**	7.391	6.798	12.322	10.159
40	5.541	5.458	5.378	5.633	**5.281**	6.373	6.032
41	11.997	10.939	**10.597**	12.121	11.707	15.437	12.622
42	7.579	7.543	**7.540**	7.976	8.273	10.130	8.877
Mean:	10.147	9.782	**9.536**	10.858	10.530	14.443	12.249

For reference, Figure 3.4 shows the predictions and measured error of storm 37, one of the most intense geomagnetic storms of the test. For clarity and conciseness we provide the 1 and 2 hours forecast of the laboratory conditions model and the best operational model.

Table 3.7: Forecast Skill Scores (Compared to the Burton Equation) as the Baseline for 2-hours forecast over the test storms for the SYM-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	GBM	LSTM-CNN
26	0.203	<u>0.238</u>	**0.258**	0.225	0.168
27	0.014	0.026	<u>0.056</u>	**0.071**	-0.076
28	<u>0.390</u>	0.380	**0.409**	0.362	0.337
29	<u>0.154</u>	0.126	**0.187**	0.097	0.047
30	-0.010	<u>0.036</u>	**0.091**	-0.162	-0.163
31	0.329	<u>0.358</u>	**0.358**	0.073	0.086
32	<u>0.321</u>	0.295	0.313	0.279	**0.351**
33	0.112	0.264	<u>0.264</u>	0.047	**0.282**
34	**0.069**	-0.017	0.061	<u>0.065</u>	0.051
35	0.141	<u>0.165</u>	**0.176**	0.102	0.062
36	0.177	<u>0.185</u>	**0.194**	0.080	0.081
37	0.190	<u>0.231</u>	**0.232**	0.087	0.138
38	0.205	0.193	**0.243**	<u>0.213</u>	0.204
39	0.341	<u>0.402</u>	**0.404**	0.272	0.331
40	0.081	0.095	<u>0.108</u>	0.066	**0.125**
41	0.049	<u>0.133</u>	**0.160**	0.040	0.072
42	0.146	<u>0.150</u>	**0.151**	0.101	0.068
Mean:	0.172	<u>0.201</u>	**0.222**	0.114	0.140

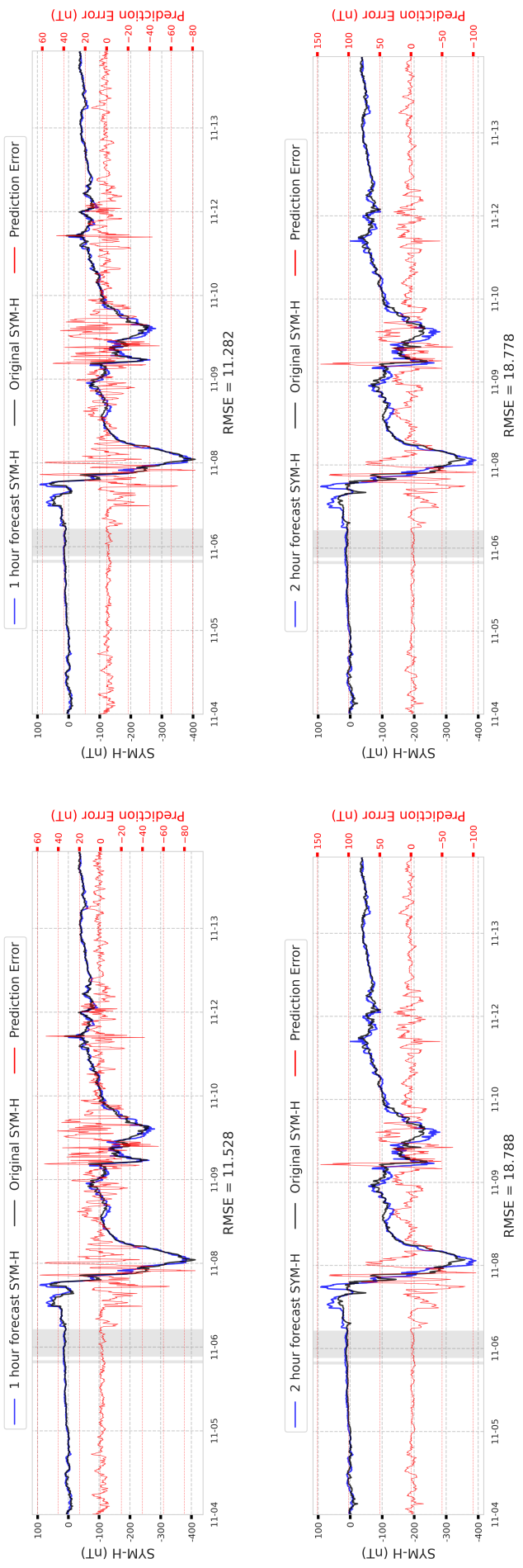


Figure 3.4: Predictions for Storm 37, November of 2004 made by the model trained with SWEPAM and SWICS data. Top: 1-h (left) and 2-h (right) predictions for the model evaluated in the operational scenario. Bottom: 1-h (left) and 2-h (right) predictions for the model evaluated in laboratory conditions. The black line represents the original SYM-H values, the blue line the model's predictions and the red line the prediction error. The shaded areas represent the times when SWEPAM's data is missing.

3.5 Model evaluation for the ASY-H index

Table 3.8 present the RMSE values compared to the persistence model for the 1 hour ASY-H predictions on the selected test storms (see Table 3.3). For this forecast, the proposed model evaluated under laboratory conditions, using SWICS to fill the SWEPAM missing values, achieves the best overall performance. However, the difference is not that big compared to the previous case. This is caused by the inherently higher difficulty of forecasting the ASY-H index compared to the SYM-H. Moreover, the considered split following the original one made by Siciliano et al. [36] is specifically tailored for the SYM-H index, leading to imbalanced sets if the same strategy is applied to the ASY-H index. Nevertheless, the strategy of using the SWICS data to train the model proves again to be successful, improving the performance of the model in the operational scenario where only the SWEPAM data is used and the missing values are filled similar to how they can be filled in an operational scenario.

Table 3.8: RMSEs for 1-hour forecast over the test storms for the ASY-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	Persistence
26	7.921	7.828	**7.431**	9.148
27	11.197	11.259	**11.062**	13.328
28	9.833	9.735	**9.299**	12.849
29	**8.096**	8.140	8.314	9.428
30	**7.061**	7.137	7.127	8.011
31	20.739	20.611	**19.244**	23.303
32	10.523	10.508	**10.454**	14.714
33	19.148	**19.031**	19.705	21.877
34	18.166	18.062	**17.953**	24.576
35	13.540	**13.482**	13.613	17.028
36	**13.336**	13.415	13.379	16.558
37	20.595	**19.746**	21.277	28.968
38	6.390	6.391	**6.367**	7.383
39	10.327	10.099	**9.677**	12.125
40	9.422	9.310	**8.964**	10.552
41	**16.438**	16.763	17.419	17.797
42	9.512	**9.382**	9.387	11.110
Mean:	12.485	12.406	**12.393**	15.2208

Table 3.9 presents the RMSE values for the 2-hour ASY-H predictions, compared to the persistence model, for the selected test storms (as listed in Table 3.3). In this evaluation, the model's performance is measured on the three previous configurations. The results reveal that forecasting the ASY-H index for a 2-hour horizon is a very complex task, as evidenced by the closely clustered RMSE values across the different scenarios. Notably, the SWICS Laboratory setup shows a marginally better performance, yet the differences among the scenarios are not as pronounced as one might expect. This observation points to the intrinsic challenge in accurately forecasting the ASY-H index. For this case, the same problem as with the 1h forecast is present, the split is not ideally suited for the ASY-H index. Despite these challenges, the use of SWICS data in model training demonstrates its efficacy, particularly in the operational scenario where it augments SWEPAM data. This approach not only compensates for the missing data in an operational setting but also

underscores the potential of SWICS data in enhancing forecast accuracy for the ASY-H index.

Table 3.9: RMSEs for 2-hour forecast over the test storms for the ASY-H index. Best prediction for each storm is highlighted in bold and the second best is underlined.

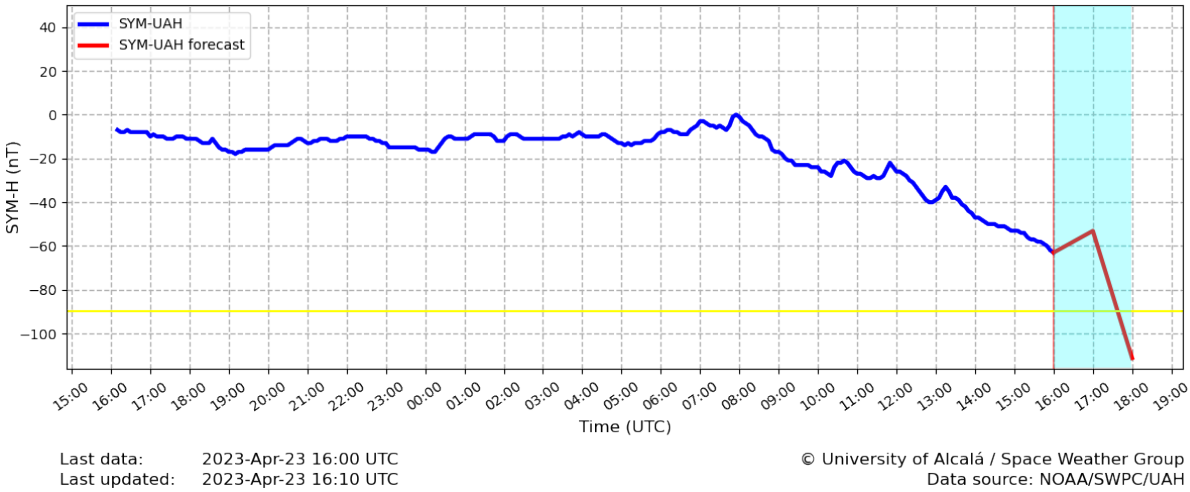
Storm #	SWEPAM Operational	SWICS Operational	SWICS Laboratory	Persistence
26	8.948	9.055	**9.055**	13.666
27	**13.545**	14.104	14.101	18.164
28	**10.658**	10.790	10.790	16.698
29	10.727	**10.527**	10.528	13.227
30	**8.462**	8.670	8.660	11.084
31	26.745	**25.217**	25.218	30.622
32	13.696	13.068	**13.061**	17.942
33	21.837	20.809	**20.806**	28.732
34	21.590	21.174	**21.173**	29.050
35	**16.316**	16.339	16.341	21.275
36	16.058	15.975	**15.973**	20.746
37	**26.464**	24.968	24.966	33.586
38	**7.299**	7.406	7.328	9.894
39	10.454	10.213	**10.103**	15.084
40	9.704	9.874	**9.866**	11.999
41	**19.807**	20.198	20.228	21.405
42	11.139	10.963	**10.959**	13.919
Mean:	14.909	14.665	**14.656**	19.241

3.6 Operational deployment

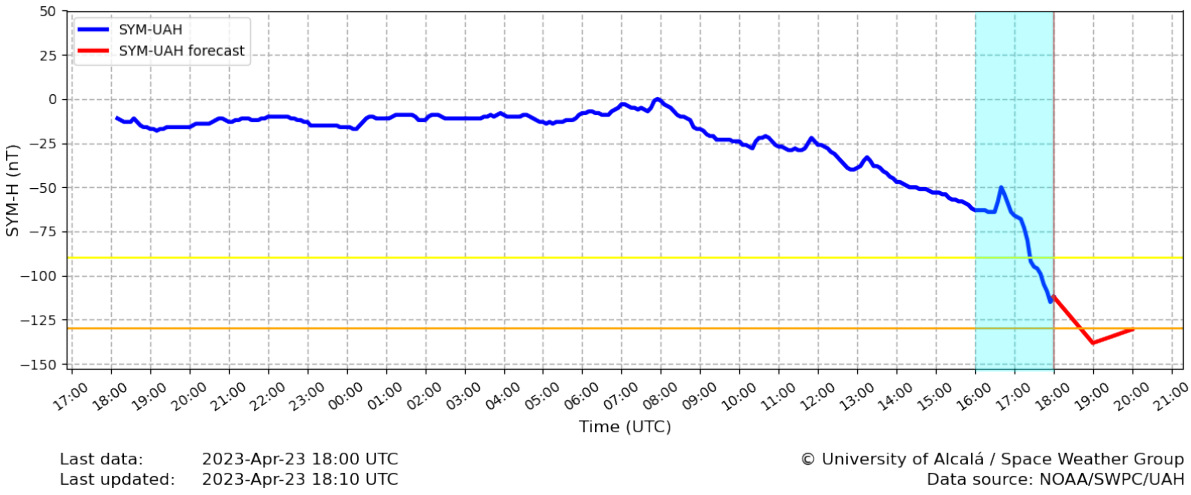
The presented model developed for forecasting the SYM-H index has been successfully operationalized and is currently making real-time forecasts at http://www.senmes.es/pub/ISG/lastSYMforUAH.png. This deployment marks a significant milestone in our work, showcasing the practical applicability of our research in real-world scenarios.

Since then, the model has forecasted two notable storms. The first one occurred on April 23rd, 2023. This storm was of moderate intensity and we can highlight the 2 hours forecast, shaded in blue in Figure 3.5 (a), which was exceptionally close to the real value shown in Figure 3.5 (b).

The second significant event was during the geomagnetic storm that occurred on November 5th, 2023. This storm, characterized by its moderate intensity, was forecasted reliably, showcasing the model's proficiency in predicting the SYM-H index. Figure 3.6 illustrates the model's forecasting prowess during the November 5th storm event.

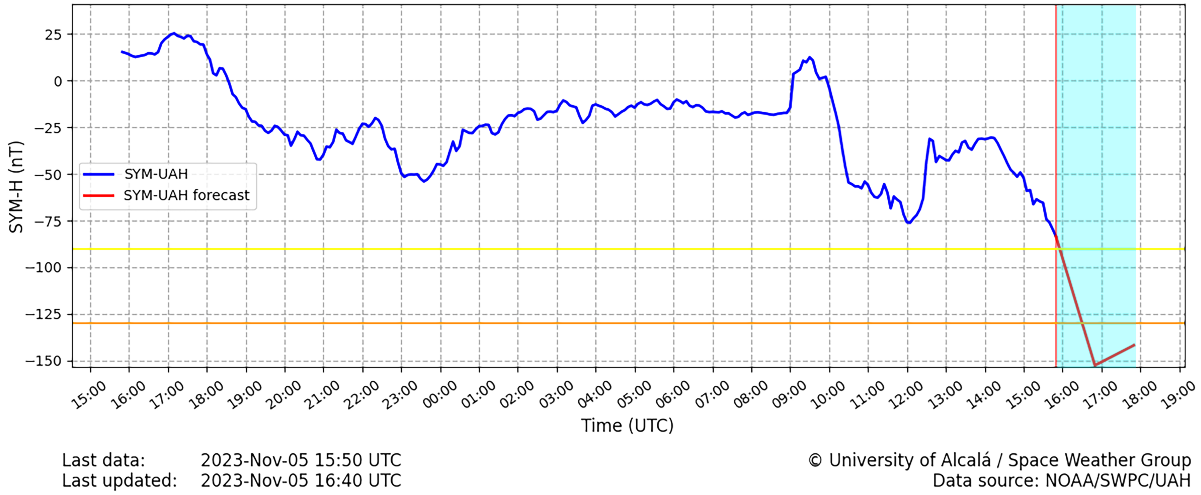


(a) The blue shaded area marks the peak of the storm, highlighting the forecast.

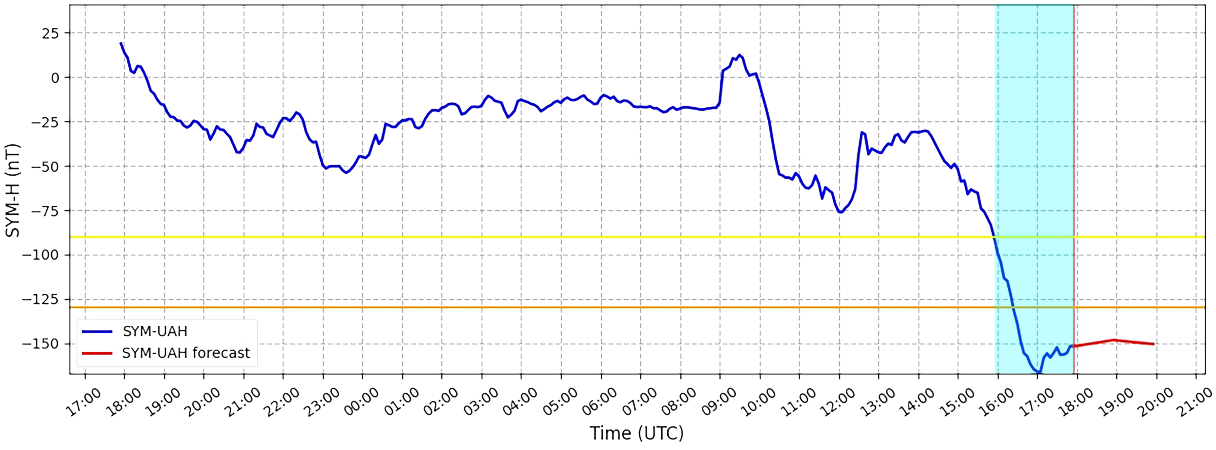


(b) The blue shaded area are the observed values that were forecasted in top previous image.

Figure 3.5: Screen shot of the model operating in real-time. The observed values for the SYM-H index are in blue and the forecasted ones in red at the onset of the April 23rd storm. Being the forecast remarkably close to the observed values.



(a) The blue shaded area marks the peak of the storm, highlighting the forecast.



(b) The blue shaded area are the observed values that were forecasted in top previous image.

Figure 3.6: Screen shot of the model operating in real-time. The observed values for the SYM-H index are in blue and the forecasted ones in red at the onset of the November 5th storm. Being the forecast remarkably close to the observed values.

3.7 Conclusions

One of the most challenging problems in any ML related project is the quality of the data used to train the models and to make the predictions in a real-time operational scenario. In the particular case of forecasting geomagnetic indices, some of the input features present a considerably high missing rate, namely, the plasma features. They are directly measured by ACE's SWEPAM instrument but the instrument can be saturated during intense storms [119], limiting the performance of the models that rely on those features to make the prediction if they are missing.

Previous works decided to either not use the plasma features and rely only on the features that have close to 100% availability or linearly interpolate the missing values. Notwithstanding, both approaches have limitations. On the one hand, if the plasma features are not used, many relevant information is lost. On the other hand, performing a linear interpolation over the missing values may lead to inaccurate results if the feature is being missing during the intense periods of the storm.

To minimize the impact of the missing plasma features we propose the usage of the data measured by ACE's SWICS instrument, which provides the plasma bulk speed, thermal velocity and temperature to fill the missing values of the SWEPAM instrument. Nevertheless, the original data measured by SWEPAM is preferred because SWICS measures the data with a 12-minute resolution, much lower than the usual 5-minute aggregation used in the forecast.

Supported on that data pre-processing, we propose a DNN architecture for operational real-time forecasting of the SYM-H and ASY-H indices 1 and 2 hours in advance. We have trained two instances of the proposed model, one using only SWEPAM data, filling the missing data using interpolation and another one filling the missing measurements using SWICS data. Then, both models have been evaluated on a simulated operational scenario, where SWICS data is not available in real time, and, therefore, it is not used to fill missing values in the SWEPAM data. Under those circumstances, interpolation is only performed if the value to interpolate to is available at the time of the prediction; otherwise, the last valid measurement is propagated forward. Under that evaluation, the model trained with SWICS performs better than the one trained only with SWEPAM. This supports the idea that improving the training data will improve the model's performance during operation, even if the plasma data presents missing values.

We have also evaluated the model under laboratory conditions, where all the data is available at the time of the forecast, to quantify the performance loss when operational constraints are applied. Results are, on average, significantly better for all the cases when compared to previous models, with the models trained with SWICS the ones that perform the best. In this regard, having access to SWICS quality data during operation will allow us to further improve the forecast.

In addition to our comprehensive analysis on the SYM-H index, we have also extended our approach to the forecasting of the ASY-H index. Unlike the SYM-H index, for ASY-H, there are no existing models to serve as a comparative benchmark, thereby restricting our comparison solely to the persistence model. Despite this limitation, the conclusions drawn from the SYM-H index studies are equally applicable to ASY-H. We found that incorporating SWICS data into the training process is beneficial even in operational scenarios. However, it is important to note that forecasting the ASY-H index presents a

significantly higher challenge. The current subset of data used for SYM-H is not entirely suitable for ASY-H, indicating a need for a revised and specifically tailored dataset for the ASY-H index. This refinement is essential to enhance the forecast accuracy and reliability for the ASY-H index, highlighting the critical importance of data quality in ML projects, particularly in real-time operational scenarios.

Chapter 4

Improving the Framework

Everyone wants to be the Sun to lighten up someone life
but why not be the Moon to brighten in the darkest hour.

Revising the initial selection of geomagnetic storms is essential for developing a more accurate and comprehensive forecasting model. The current selection focuses on the SYM-H index, which fundamentally differs from the ASY-H index and is primarily centered on CMEs, with limited coverage of other space weather phenomena, such as High-Speed Streams (HSSs). Additionally, the absence of a universally accepted threshold in the scientific community for defining the intensity at which a disturbance qualifies as a storm presents a significant challenge, particularly in the absence of a statistically validated classification based on intensity.

To address these challenges, we have undertaken a thorough study of all available data for both SYM-H and ASY-H indices in Section 4.1. This study aims to classify storms based on their intensity, with the final objective of ultimately refining and expanding upon the initial storm selection established by Siciliano et al. [36]. This expansion is vital for the comprehensive training and validation of ML models. Additionally, given the distinct nature and behavior of the ASY-H index, our study develops a separate, tailored split of storms for this index, ensuring effective training and development of ML models specific to ASY-H.

Recent advances in ML have yielded promising results in the realm of geomagnetic storm forecasting. However, the evaluation of ML models in this field has typically relied on generic regression metrics, such as RMSE and the coefficient of determination (R^2). These metrics fail to fully capture the unique challenges associated with forecasting geomagnetic storms, particularly during periods of high geomagnetic activity. To address this limitation, we introduce the Binned Forecasting Error (BFE) metric, which offers a more nuanced evaluation of model performance across varying storm intensity levels. The BFE metric enables a robust comparison of different forecasting models, even when storm events vary in timing or intensity.

To operationalize the evaluation framework, we conduct a comparative study between a baseline NN model and a persistence model, demonstrating the effectiveness of the BFE metric during periods of intense geomagnetic activity. Moreover, we incorporate real-time considerations into our evaluation by utilizing preliminary solar wind and IMF

measurements from the ACE satellite, simulating operational conditions where forecasting models are expected to function. This approach enhances the applicability of our methods for real-world SW forecasting scenarios.

4.1 Classifying and bounding geomagnetic storms based on the SYM-H and ASY-H indices

Geomagnetic storms are extraordinary disturbances. However, not all storms have the same intensity, leading to different consequences. Since the intensity of storms is measured using the disturbance caused in the geomagnetic indices, they are often classified into different classes depending on the maximum disturbance caused in the indices. Consequently, thresholds are used to determine the different classes of storms, based on their intensity and probability of occurrence.

An accurate classification of the intensity of geomagnetic storms serves two main purposes. First, from an academic perspective, understanding the underlying mechanisms that generate the storms. Several studies analyze geomagnetic storms focusing on those of similar intensity [120]–[123]. Second, there is a natural concern for mankind [124] due to the varying intensity of geomagnetic storms and their potential consequences. These consequences range from damaging satellites [15] and increasing their orbital drag [125], [126] to disrupting GNSS signals [127], [128] and even damaging the power grid infrastructure [129].

Despite the great importance of having an accurate classification of geomagnetic storms based on their intensity, there is no consensus among authors. Nevertheless, the most used index to determine the intensity of geomagnetic storms, at least in scientific literature, is the Dst index. Initially, Gonzalez et al. [124] classified the geomagnetic storms depending on the minimum value reached by the Dst values from 1976 to 1986 (solar cycle 21). They set the thresholds shown in Table 4.1 to classify the geomagnetic storms into weak, moderate and intense groups, using the distribution of the observed hourly values as a guideline. Weak storms are those when the minimum value reached by the Dst is between -30 and -50 nT, corresponding to 25% of the hourly distribution of the observed values; moderate storms encompass those with a minimum Dst value between -50 and -100 nT, corresponding to 8% of the values; finally, storms when the index falls below -100 nT are considered intense, corresponding to 1% of the observed values. This classification was also used by other authors, such as Kamide et al. [120], to analyze storms of similar intensity.

Later, Gonzalez et al.[130] and Echer et al. [131] proposed a new category of superintense storms for those storms in which the Dst index reached values lower than -250 nT. This classification is the most widely used in the literature [128]. Nevertheless, authors

Table 4.1: Geomagnetic storms classification by Gonzalez et al. [124].

Level	Dst (nT)	Percentile (%)
Weak	[-30, -50)	25
Moderate	[-50, -100)	8
Intense	[-100, -∞)	1

like Rawat et al. [132] and Li et al. [133] considered the superintense storms those where the index reached values lower than -300 nT. A different classification was proposed by Loewe and Prölss [134]. They considered storms whenever a depression in the Dst index exceeds -30 nT. Then, they qualified the storms into four classes with different thresholds. They considered storms in which the Dst peak is between -30 nT to -50 nT as weak storms, corresponding to 44% of the identified storms then, between -50 nT and -100 nT to be moderate storms corresponding to 32% of their identified storms, between -100 nT and -200 nT are considered strong, corresponding to 19% of the detected storms, from -200 nT until -350 nT are considered severe storms, corresponding to 4% of the detected storms. Finally, storms more intense than -350 nT are considered exceptional, corresponding to the 1% most intense storms. However, they performed the percentile calculations to determine each class over the detected storms, as opposed to considering the whole time-series as other authors. This classification has been used by other authors, such as Uwamahoro and Habarulema [123], to model the total electron content during storms. In Loewe's work they also performed a superposed epoch analysis on the different classes, using the Dst peak as the common epoch time.

Nonetheless, the Dst is not the only index that has been used to classify geomagnetic storms. For instance, Gosling et al. [135] used the Kp index to classify the storms into four groups: major, large, medium and small. For example, this classification has been used by Richardson and Cane [122] but the applications are limited due to the 3-hour time resolution of the Kp index as some SW impacts are driven by short-time (∼1.5–2 h) disturbances [136].

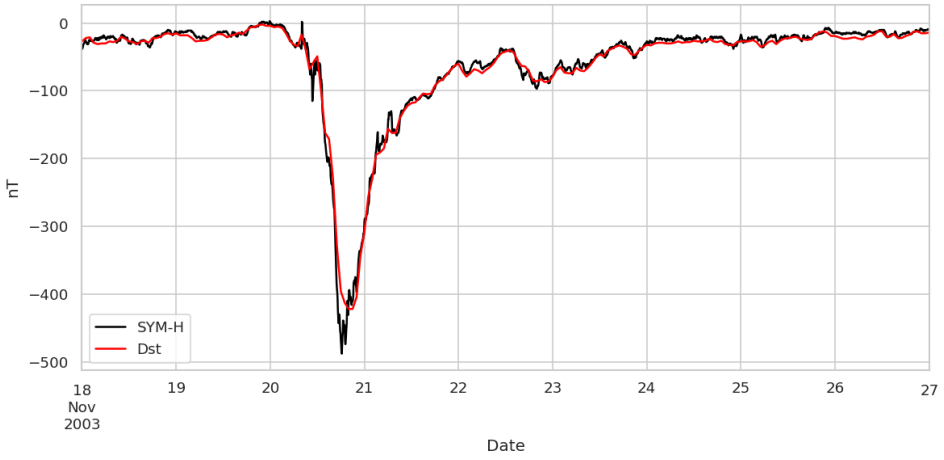


Figure 4.1: Geomagnetic storm of November 2003. The SYM-H index deviates from the Dst index on very high intensities, below -300 nT.

One of the indices that is being increasingly used by the community is the SYM-H index. Wanliss and Showalter [137] proved that the SYM-H index is comparable to the Dst index until -300 nT. For quiet times and for small storms the deviations are typically no more than 10 nT between both indices. Moderate storms feature deviations typically only slightly more than 10 nT, while intense storms have deviations that are usually less than 20 nT. Nevertheless, in those storms where the SYM-H index reaches values below -300, the indices are no longer comparable. For example, in the geomagnetic storm of November 2003, depicted in Figure 4.1, the peak value of the Dst index is -422 nT, while the SYM-H peak is -488 nT, 1.15 times lower.

Therefore, the previous classification based on the Dst index cannot be used as a direct proxy for the SYM-H index. Additionally, the thresholds set by Gonzalez et al. [124] only used the Dst data from 1976 to 1986; for half of the temporal frame used to determine the thresholds the SYM-H is not even available, since it has been calculated from 1981 onward. Since then, we have had more than 40 years' worth of SYM-H measurements with new storms that are more intense than any of the storms which happened during that period. Taking that into account, the distribution of the index calculated by Gonzalez et al. [124] has changed considerably when considering all the available data for the SYM-H index. In this regard, Hutchinson et al.[138] separated the storms into three categories based on the minimum value reached by the SYM-H: weak (-150 < SYM-H < -80) nT, moderate (-300 < SYM-H < -150) nT and intense (SYM-H < -300) nT. However, the threshold values were selected without offering explicit reasoning.

Currently, there are several applications and users that can take advantage of a proper storm classification. One of those is the forecasting of the geomagnetic indices. This is an ongoing problem that is constantly gaining more importance: by forecasting the indices it could be possible to predict a geomagnetic storm, and, therefore, to take the appropriate preventive containment measures. Particularly, most forecasting systems are focused on geomagnetic storms, when the consequences are more severe. Several techniques have been used to forecast the various geomagnetic indices, ranging from mathematical models to ML based solutions. Zhelavskaya et al. [65], Shprits et al [139] and Wintoft et al. [64] forecast the Kp index. O'Brien and McPherron [117], Gruet et al. [75] and Lazzús et al. [76] forecast the Dst index. More recently, attention has shifted to forecast the SYM-H and ASY-H indices [36]–[38], [80], [140] due to their high temporal resolution compared to the previous indices.

The works that forecast the SYM-H and ASY-H indices focus only on the prediction of the index during intense geomagnetic storms. However, it is not clearly defined which storms are considered intense and, therefore, used to train and evaluate the forecasting models. For instance, Cai et al. [140] considered the storms in which SYM-H peak (the minimum value reached by the SYM-H) was less than -60 nT, while Bhaskar and Vichare [80] considered the storms in which the SYM-H peak was less than -85 nT. By its side, Siciliano et al. [36] considered the storms with a SYM-H peak of less than -100 nT. Moreover, for other indices, such as the ASY-H, currently there is no classification of the geomagnetic storm.

Trying to provide a comprehensive and objective storm classification based on the SYM-H and ASY-H indices, we propose a classification method based on the cumulative distribution function of the indices. We apply the industry-wide percentiles used in risk assessment to the distribution to separate the different classes of geomagnetic storms according to the peak values of the indices. Using the calculated percentiles, the storms can be classified according to their intensity and asymmetry, based on the probability of occurrence. Then we set the boundaries of the storms using a superposed epoch analysis. This approach allows us to arrange an "intensity-classified storms set" useful for later studies.

4.1.1 Classifying geomagnetic storms

The procedure to classify the geomagnetic storms based on their intensity and asymmetry according to the SYM-H and ASY-H indices is the same for both indices: the only difference between the SYM-H and the ASY-H indices is that for the SYM-H intense values are represented with increasingly negative values whereas they are represented with positive values for the ASY-H.

If we consider the geomagnetic storms as infrequent events that pose a risk, the severity of the risk can be quantified using the peak value reached by the evaluated geomagnetic index (maximum in the case of the ASY-H index or minimum in the case of the SYM-H). Additionally, the time interval of similar events can be estimated as the inverse of the probability, which is indicated by the Cummulative Distribution Function (CuDF). For this purpose, the time series data of indices have to be made of random variables, that is, their values have to be independent. Notwithstanding, both the 1-minute and 5-minutes aggregated series of the indices are not random variables as the observations are highly correlated with the previous value, having an autocorrelation greater than 0.95. Additionally, the great amount of quiet time skews the distribution function of the original time series towards values corresponding to nominal behavior.

To make the time series useful for a probability analysis, we have resampled the series of SYM-H and ASY-H indices so subsequent data points do not have a causal relationship, which is indicated by a low autocorrelation value. For this purpose, we have aggregated the data in 27 days intervals. The choice of 27 days is motivated to avoid the persistence in the indices caused by the disturbances associated with fast streams from coronal holes lasting several solar rotations. Then, we have chosen the peak value of the index of each aggregation period to identify the maximum disturbance in each period. The objective is to make the resampled series have a low autocorrelation value, making the time series statistically independent.

The approach of resampling the time series using the minimum value brings forth a reduction in the number of data points, effectively eliminating periods of relative calm and thus magnifying the significance of active periods. We have also tested using other periods to perform the resample of the time-series. Periods shorter than 27 days have considerable autocorrelation values. Moreover, opting for shorter intervals yields a surplus of data points corresponding to quieter times, resulting in smaller values within the CuDF. Nevertheless the autocorrelation of the values due to the rotation of the Sun invalidates the CuDF calculations. Conversely, the longer the aggregation interval, the greater prominence intense storms acquire. However, longer periods have an asymptotic behaviour.

We have also tested the option of performing the statistical calculations using the bootstrap method to eliminate the autocorrelation. However, the great imbalance of inactive time compared to the active time is still present, making the average values of the percentiles in the CuDF almost similar to the original time series.

To have statistical reasoning behind the selection of the thresholds to classify the storms, we decided to use the percentiles of the distribution of the intensity of the indices' time series, following Cid et al. [35]. We have chosen the 60^{th} percentile as the cutoff to determine when the disturbance is strong enough to be considered a geomagnetic storm. Values less intense than the 60^{th} percentile can be considered inactive or quiet time. Once

the starting point of the 60th percentile has been set we have chosen to differentiate the remaining CuDF into four different classes, starting with the geomagnetic storms of low intensity for those storms in which the peak value for the index is more intense than the 60th percentile.

Then, we chose the 80th percentile as the upper boundary for the low intensity storms and the start for the moderate storms. That particular percentile is in line with the IFRS 17 [141], in which industry-wide consensus for risk assessment is between the 70th to the 80th percentiles. Additionally, storms with an intensity between these thresholds already can harm technological systems in space such as satellites or disrupt GNSS signals.

After that, we considered the 95th percentile as the upper boundary for moderate storms. Storms in which the index reaches a peak value more intense than the 95th percentile will be considered as intense. Finally, we have chosen to further classify the remaining distribution in two classes: the intense and superintense storms, using the 99th percentile as the boundary. The reasoning behind that is because above the 95th percentile the data points are very different: for both indices the range of values over the 99th percentile is greater than all the values under it, making the extra class necessary to keep the groups as consistent as possible.

Note that in the following the values for the selected percentiles are rounded to the nearest ten.

4.1.1.1 Analysis of the SYM-H index

To perform the classification, we consider the historical records of the SYM-H index from 1981 to 2022. To resample the SYM-H we have chosen the minimum value in each time period of 27 days; the resulting series is depicted in Figure 4.2. Then, to test the independence of the time series we calculate the autocorrelation value for the 9 subsequent time-steps, as shown in Figure 4.3. After performing the resample, we obtain that the autocorrelation value is lower than 0.3 for any time lag.

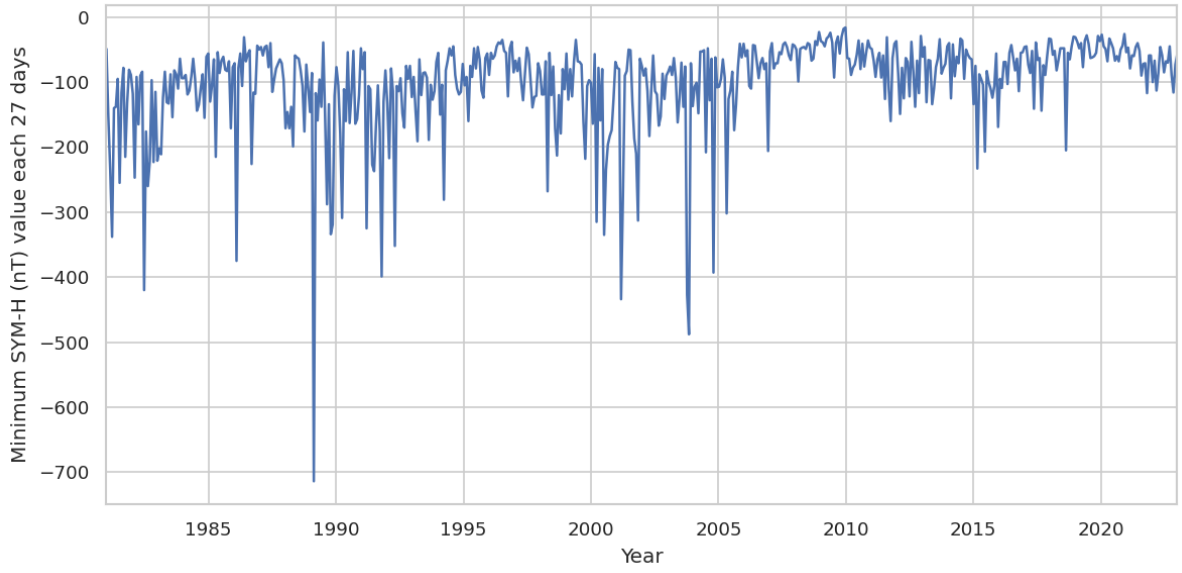


Figure 4.2: Time series of the resampled SYM-H index.

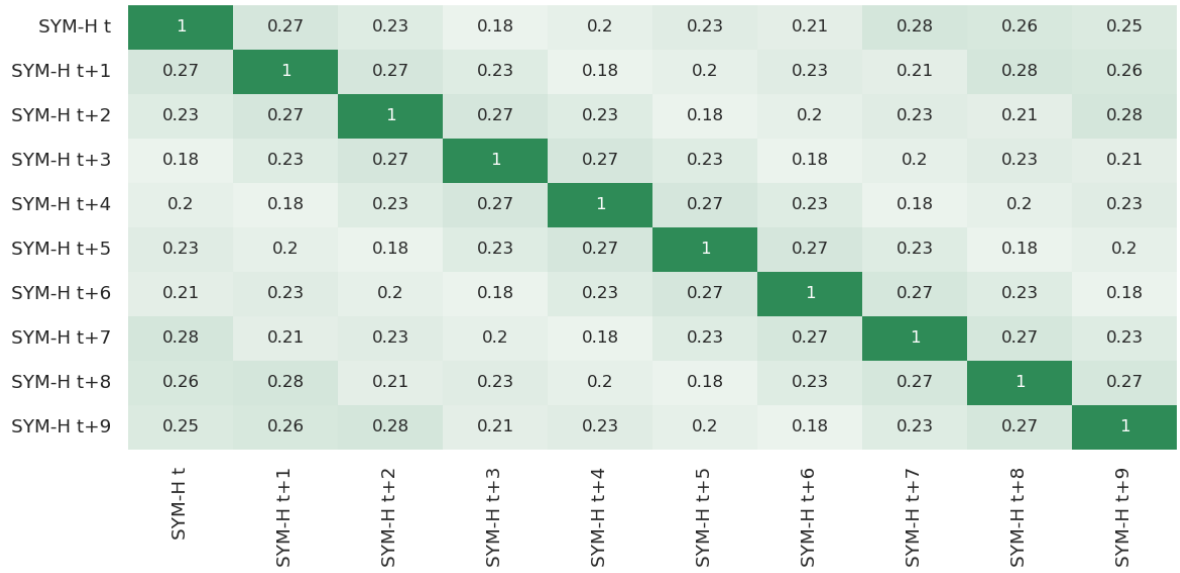


Figure 4.3: Heatmap of the autocorrelation for the resampled series of the SYM-H index.

Once we have a resampled series that is statistically independent, we can perform the CuDF to estimate the thresholds for each of the identified classes. Figure 4.4 depicts the CuDF of the SYM-H index, being the vertical dashed lines the selected percentiles for the different storm categories as stated below:

- The green area represents the inactive distribution, it extends until the 60^{th} percentile, corresponding to -90 nT.
- The yellow area represents the distribution for the low intensity storms, extending from the 60^{th} percentile to the 80^{th}, corresponding to -130 nT.
- The orange area represents the distribution for the moderate storms, starting on the 80^{th} percentile and extending until the 95^{th}, corresponding to -230 nT.
- The red area represents the distribution of the intense storms, encompassing from the 95^{th} percentile to the 99^{th}, corresponding to -390 nT.

The starting point of the 60^{th} percentile on the resampled time series corresponds to -90 nT. This threshold is a bit higher than the classification made by Hutchinson et al. [138], which considered -80 nT to be the lower bound for weak storms. It is also higher than the threshold for low storms considered in the earlier forecasting works [80], [140] where they used -60 nT and -85 nT as the threshold for low intensity storms. However, is a bit lower than the threshold used in the later forecasting works [36]. Compared to the Dst index, our proposed threshold is higher than the established classifications for the Dst. Nevertheless, this is related to the fact that Dst values are generally lower than SYM-H due to the hourly average nature of the index. We decided to set the upper bound for this classification at the 80^{th} percentile, corresponding to -130 nT.

Then, we consider the storms with a SYM-H peak lower than -130 nT to be of moderate intensity; it extends until the 95^{th} percentile, corresponding to -230 nT. Those values are in line with what have been considered moderate storms in the literature: a bit lower

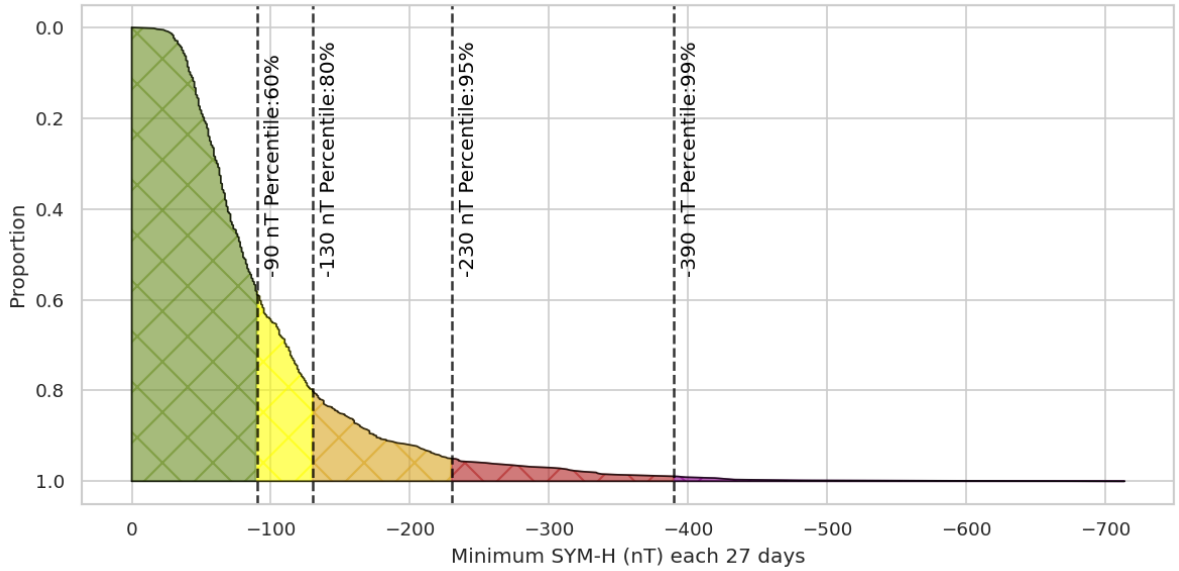


Figure 4.4: Cumulative Distribution Function of the minimum SYM-H (nT) every 27 days.

than the threshold considered by Hutchinson et al. [138] of -150 nT for the SYM-H for the moderate storms, but also higher than the threshold commonly used for moderate storms in the Dst index. Storms in this range are considered moderate since they can harm satellites and disturb GNSS signals [15], [127], but they are supposed to be not intense enough to cause black-outs and damage the electrical infrastructure.

After that, the intense storms range from a SYM-H peak of -230 nT, corresponding to the 95th percentile, to -390 nT, corresponding to the 99th percentile. Finally, we have chosen to create a fourth class for the remaining distribution to differentiate storms in which the SYM-H peak surpasses the 99th percentile. This distinction is needed because the data points in the resampled time series outside the 95th percentile are very different. As there are 29 samples with values lower than -230 nT, if they were grouped into the same class, the standard deviation would be 100 nT with a mean of -335 nT, which make the distribution of the group too sparse. Instead, we propose a superintense class for the storms with a SYM-H peak more intense than the 99th percentile, which corresponds to -390 nT.

This split separates the 29 data points into 22 intense and 7 superintense storms. The resulting intense group is much more consistent, having a mean of the SYM-H of -290 nT and a standard deviation of 43 nT. Despite that, the superintense group has one outlier caused by the storm of 1989 with a SYM-H peak of -714 nT, which in conjunction with the small amount of data greatly pollutes the mean and standard deviation of the group. In case of removing that data point, the remaining 6 superintense storms would have a mean of -430 nT and a standard deviation of 31 nT. However, keeping it in the sample, the resulting group has a mean of -470 nT and a standard deviation of 113 nT, evidencing the relevance of rare events in the analysis. Table 4.2 summarizes the classification for the SYM-H index.

Table 4.2: Geomagnetic storms classification using the SYM-H.

Intensity	SYM-H (nT)	Percentile (%)	Data points
Quiet	[-90, ∞)	[60, 0)	330
Low	[-130, -90)	[80 , 60)	123
Moderate	[-230, -130)	[95 , 80)	87
Intense	[-390, -230)	[99 , 95)	22
Superintense	(-∞, -390)	(100, 99)	7

4.1.1.2 Analysis of the ASY-H index

Similar to the SYM-H index, for performing the resample and CuDF computation, we consider the ASY-H records from 1981 to 2022. The resampled time series is depicted in Figure 4.5. The main difference between the indices' values of the SYM-H and ASY-H is that the disturbances are measured with positive values in the ASY-H. Thus, we select the maximum value of each 27 days group instead of the minimum. Considering that the index measures asymmetric disturbances, its behavior is more chaotic. This is reflected by a low autocorrelation value in the resampled time series. For this index, the autocorrelation is lower than 0.2 in the subsequent 9 time-steps as shown in Figure 4.6, which is significantly lower than the autocorrelation for the SYM-H.

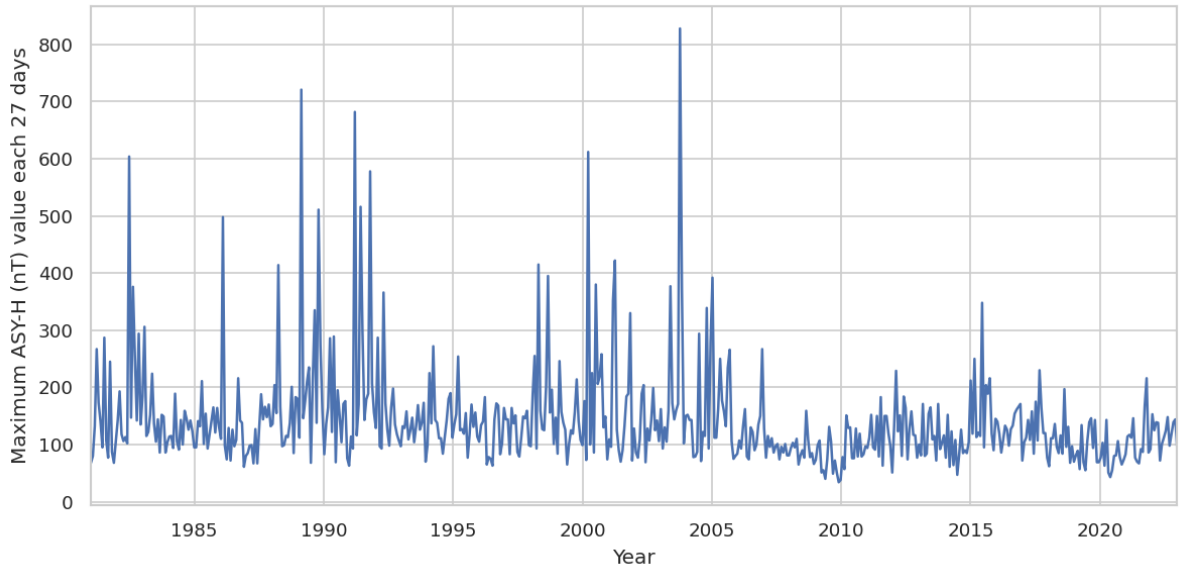


Figure 4.5: Time series of the resampled ASY-H index.

After confirming the independence of the time series we have calculated the CuDF, presented in Figure 4.7. The vertical dashed lines depict the percentiles that define the thresholds between classes. Since there is no previous classification of geomagnetic storms based on the ASY-H index we cannot compare the proposed thresholds to other authors. If we compare the shape of CuDF between the ASY-H and the SYM-H, they are fairly similar, with the main difference being that the ASY-H has, generally, higher absolute values. For instance, in the resampled time series, the average ratio between SYM/ASY is around 1.4, being the greatest in the most intense times. These values are in line with Echer et al. [131] findings, indicating a higher degree of asymmetry for the superintense storms. Considering that, despite not having other works to compare the thresholds for

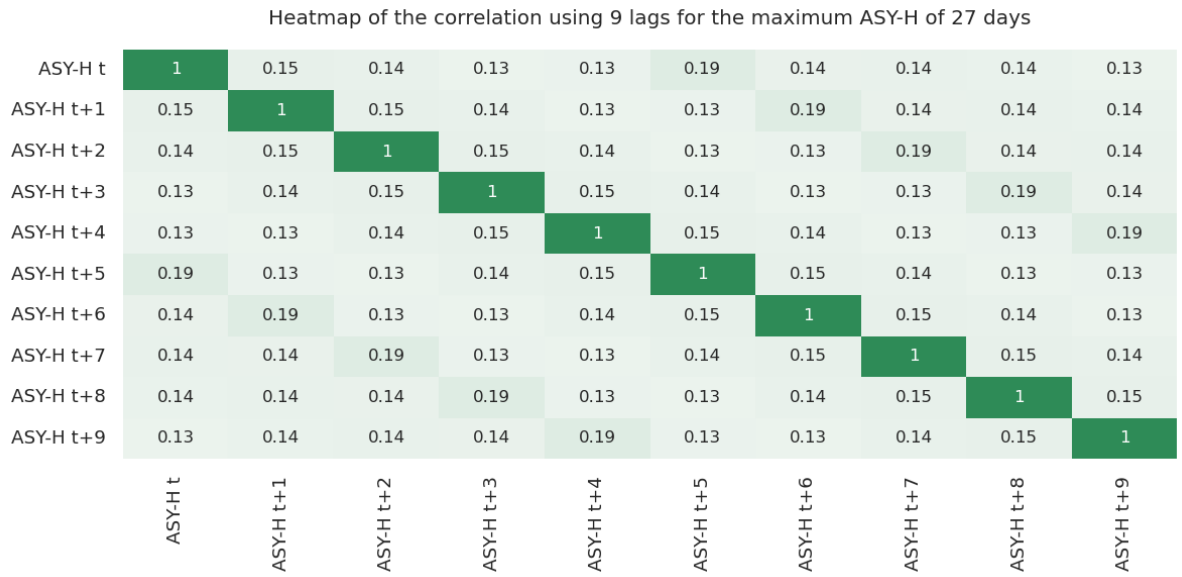


Figure 4.6: Heatmap of the autocorrelation for the resampled series of the ASY-H index.

the different classes of geomagnetic storms for the ASY-H, we will use the same percentiles as the selected for the SYM-H as guidelines.

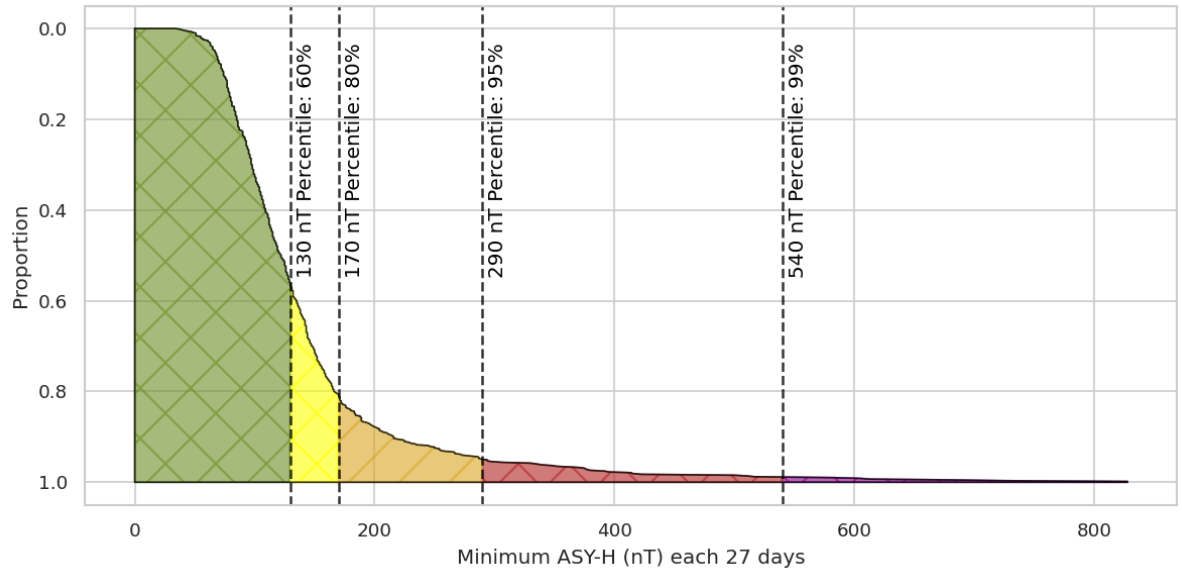


Figure 4.7: Complementary Cumulative Distribution Function of the maximum ASY-H (nT) every 27 days.

We consider the 60^{th} percentile as the threshold for differentiating between inactive time and the start of low asymmetric disturbance. Then, storms are considered as of low asymmetry when the ASY-H peak is between the 60^{th} and 80^{th} percentile, corresponding to 130 nT and 170 nT. This group is the largest. In this case, the average ratio between the SYM/ASY is 1.35. A lot of similar data points are condensed in this group, as the standard deviation of low asymmetric data points is less than 11 nT, lower than the standard deviation for the same class in the SYM-H index.

Next, moderate asymmetric storms range from those with a ASY-H peak of 170 nT to 290 nT, which corresponds to the 80^{th} and 95^{th} percentiles, respectively. In this case, the disparity in the data points increases, having a standard deviation of 35 nT, which is higher than the standard deviation for the SYM-H in the same group (29 nT). This suggests that the asymmetrical component of the storms is accentuated by the intensity of the disturbances.

Finally, we have differentiated between intense and superintense asymmetric storms for the remaining distribution. We use the 99^{th} percentile as the threshold value for separating both classes. There are 29 data points in which the peak value is above the 95^{th} percentile. If no distinction is made, the standard deviation of the ASY-H peaks for those data points is almost 140 nT. Considering that the starting threshold value for that category is 290 nT, 140 nT is a relevant disparity in the remaining data points. Thus, we separate the storms creating the category of superintense with the 99^{th} percentile, corresponding to 540 nT as the thresholds between both groups. The 29 previous data points are separated into 23 intense ones, with a standard deviation around 65 nT, and 6 superintense storms with a standard deviation of 90 nT. In this case, since there is no extreme outliers like in the SYM-H case, even the superintense group has similar ASY-H peak values. This makes this group have a smaller standard deviation than the superintense group of the SYM-H index, but the mean is considerably higher, as the ratio between the indices is around 1.5 for this class. Table 4.3 summarizes the classification for the ASY-H index.

Table 4.3: Geomagnetic storms classification using the ASY-H.

Asymmetry	ASY-H (nT)	Percentile (%)	Data points
Quiet	[130, -∞)	[60, 0)	319
Low	[170, 130)	[80 , 60)	140
Moderate	[290, 170)	[95 , 80)	81
Intense	[540, 290)	[99 , 95)	23
Superintense	(∞, 540)	(100, 99)	6

4.1.1.3 Analysis of the Dst index

We have performed the same analysis for the Dst index, resampling the data using the same time interval (27 days) as in the cases of SYM-H and ASY-H. The covered time range includes data from 1981 to 2002, i.e., the same time-range as for the SYM-H and ASY-H analysis. The results are very similar compared to those obtained with the SYM-H, albeit the autocorrelation values are slightly higher.

Figure 4.8 depicts the CuDF of the Dst index. Meanwhile the general shape of the CuDF is similar to the SYM-H, the values corresponding to the selected percentiles are 10 to 30 nT higher, except the value for the superintense percentile which is 60 nT higher, as shown in Table 4.4. The obtained thresholds for the different storm classes are in line with previous works.

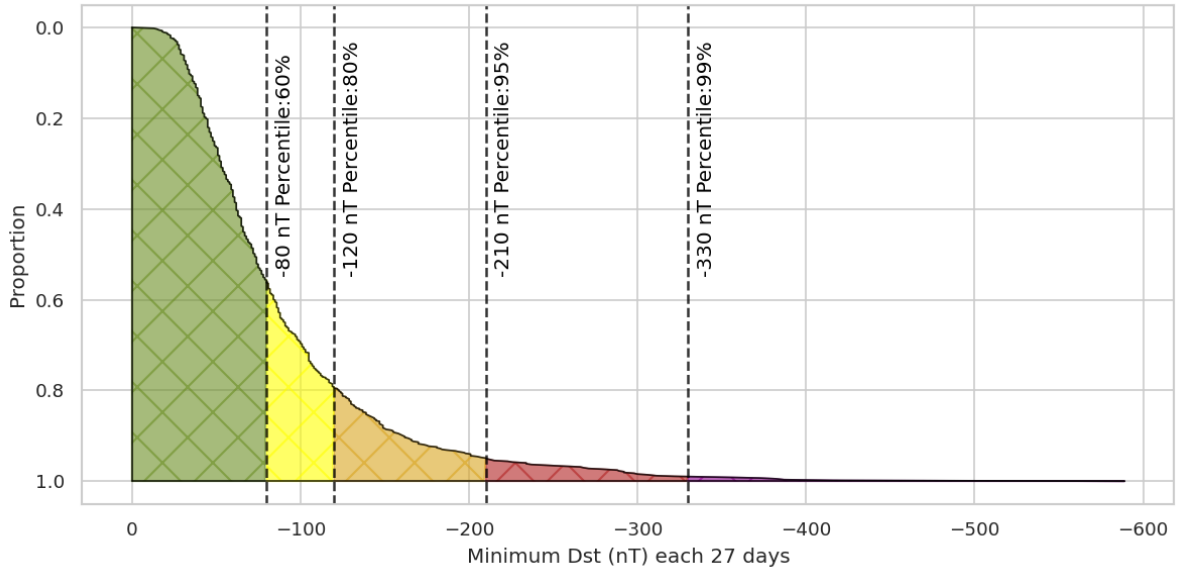


Figure 4.8: Cumulative Distribution Function of the minimum Dst (nT) every 27 days.

Table 4.4: Geomagnetic storms classification using the Dst.

Intensity	Dst (nT)	Percentile (%)	Data points
Quiet	[-80, ∞)	[60, 0)	316
Low	[-120, -80)	[80 , 60)	134
Moderate	[-210, -120)	[95 , 80)	89
Intense	[-330, -210)	[99 , 95)	24
Superintense	(-∞, -330)	(100, 99)	6

4.1.2 Setting the boundaries of a storm

To properly characterize the storms, aside from classifying them based on the disturbance caused on a geomagnetic index, we also need to timely determine the different parts of a geomagnetic storm. Different authors have defined the different parts of the storm using different approaches. For instance, Murphy et al. [142] defined each storm using three distinct times: the storm start, the epoch (defined as the time of minimum Dst during the storm) and the storm end. The storm's beginning is marked by increased solar wind activity, while its conclusion is determined by the recovery of the Dst after the solar wind activity diminishes. These phases correspond to the main phase and recovery phase of the storm but may vary in duration for each event. To conduct a superposed epoch analysis. Murphy et al. [142] normalized the initial phase of each storm is normalized to 30 hours, and the subsequent phase is normalized to 120 hours. On the contrary, Echer et al. [131] stated that the storm's main phase had a duration from 3 to 33 hours, averaging around 11 hours. Aguado et al. [143] analyzed the recovery duration of geomagnetic storms up to 48 hours after the Dst peak. Hutchinson et al. [138] considered that the recovery phase lasted until the index reached a "quiet" condition of -15 nT. Mannucci et al. [144] performed a superposed epoch analysis of the four most intense geomagnetic storms of 2003 and 2004, extending the recovery phase up to 25 hours after the start of the storm.

Wharton et al. [145] used four key dates to characterize the geomagnetic storms: the initial, main, and recovery phases of geomagnetic storms. The algorithm utilized the SYM-H index and a threshold of -80 nT to detect the storm minima. The end of the

main phase was defined by the SYM-H minimum, while the beginning of the main phase was determined by the time when the SYM-H reaches -15 nT. Before that, the initial phase ranges from the last time the SYM-H index had values greater than -15 nT. Finally, the recovery phase ranges from the SYM-H minimum until the index has recovered to values higher than -15 nT. However, there are cases in which relying on a specific value to set the bounds of the storm yields either too short or too long initial times. For example, in the storm of April, 1994, depicted in Figure 4.13, the first value over -15 nT before the SYM-H peak happened 15 days before.

Like the classification of storms based on their intensity, the duration of the storms is also a discussed topic and the duration varies among authors. In general, it is considered that the storm begins when the index starts taking values outside of the nominal values before the index peak, and is finished once the index has recovered to the nominal values. Some authors even separate the time between the nominal phase and the index peak in two different phases.

However, our focus is not on the morphology of the storm but the identification and selection of the time interval covering the whole storm. In this regard we have chosen to perform a superposed epoch analysis for the different identified classes of geomagnetic storms. This analysis can help to identify patterns in the temporal evolution of the storms. It has also been used in previous works to determine recovery approximation functions, such as Aguado et al. [143] and Hutchinson et al. [138], or to study storms of similar intensity [144].

Since we are not trying to define the morphology of the storm, we have not divided the storm into the initial, main and recovery phases similar to other authors. We have set the duration of the storm based on how the time series for the indices behaves surrounding the peak in the superposed epoch plot for each of the different classes.

The analysis consists of aligning the storms of the same class at a particular time, then all the storms are averaged and plotted to study the evolution of similar storms. We have chosen to align the storm on the indices peaks, that is, the minimum SYM-H and the maximum ASY-H. Then, we select the 120 hours surrounding the peak, average all the storms of every class, and plot the "average storms" together.

To determine the bounds of the storm we use the Pruned Exact Linear Time (PELT) [146] algorithm to identify change points in the superposed epoch plot, which indicates where significant shifts in data behavior occur. The PELT algorithm is widely used in environmental monitoring, allowing to detect locations where data significantly changes, pinpointing when a time series departs from its usual patterns.

Figure 4.9 depicts the superposed epoch plot for the SYM-H index, while Figure 4.10 presents the superposed epoch plot for the ASY-H index along with the detected change points for the different classes. In both cases, we can note that the scarce amount of superintense storms makes the superposed epoch plot a bit unstable for that category.

Regarding the SYM-H index, we have considered two days before the peak as the start of the storm. For all the classes, the disturbances start around one day and a half before the peak, being two days as the safe threshold to consider. All the change points detected with PELT are condensed during that time. Particularly, in the intense and super-intense classes, the disturbances are greater before the peak and start earlier. While considering

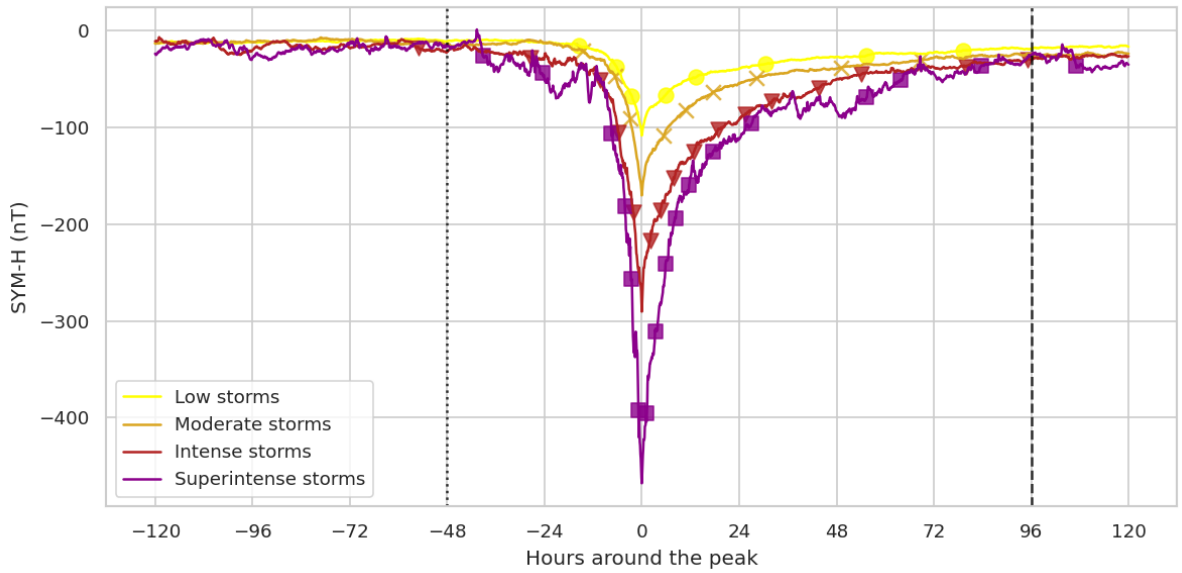


Figure 4.9: SYM-H superposed epoch plot. The bounds of the storm are represented by the vertical lines. The dotted line depicts the start of the storm, 2 days before the index peak and the dashed line depicts the end of the storm, 4 days after the peak.

only one day would be enough for the low and moderate intense classes, it is not enough for the other two, making the two days before the peak the appropriate bound.

For all the classes, the index starts to rapidly recover after the peak to its nominal values. It takes around three to four days to reach nominal values of around -20 nT, being four days as the safe threshold to determine the recovery period. Therefore, we have chosen to set the end of the recovery phase to 4 days after the peak of the index. In the superposed epoch plot, all the average storms converge to similar values after such time following the peak.

The ASY-H index presents a much higher variation one day around the peak but is more stable outside that range. In the same was as the SYM-H index, one day before the peak is not enough time to properly capture all the disturbances, as in some storms the disturbances start earlier, being two days again the appropriate bound for the start of the storm. In this case, the recovery is much faster; the disturbances after the second day are minor for all the classes but it has not completely recuperated until the fourth day after the peak when it finally stabilizes at around 20 nT.

Considering the superposed epoch analysis for both indices we have set two days before the index peak as the starting bound and four days after as the end bound for the storms.

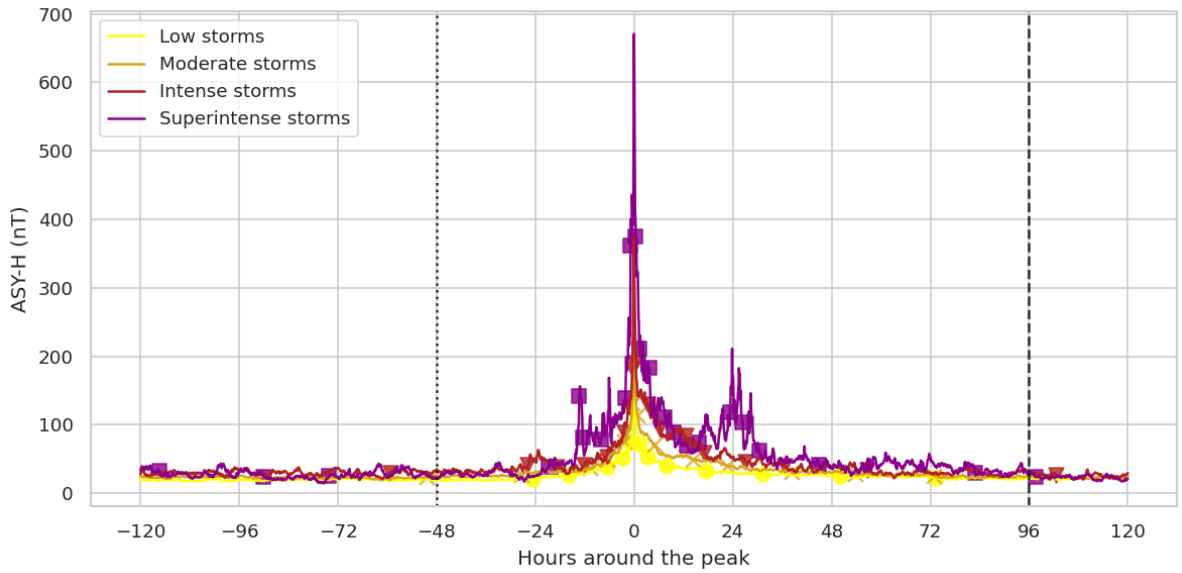


Figure 4.10: ASY-H superposed epoch plot. The bounds of the storm are represented by the vertical lines. The dotted line depicts the start of the storm, 2 days before the index peak and the dashed line depicts the end of the storm, 4 days after the peak.

4.1.3 Identification of geomagnetic storms

Once the threshold has been determined for what intensity is considered a storm of a given class and how long before and after the peak it extends, storms can be identified and classified in the time series for each index following the next procedure, which is depicted in Figure 4.11:

1. Identify a time period in which the index falls below the "low" threshold category.
2. Extend the selection two days before the first value in which the index falls below the "low" threshold category, rounding to the start of the day.
3. Extend the selection four days after the last value in which the index falls below the "low" threshold category. If in that section a new date in which the index is below the "low" threshold category is found, there is another storm before the current one has completely recovered, so we select four days after the new last "low" value, rounding to the end of the day.

Once all the storms have been identified they can be categorized into their respective category based on the peak value of the SYM-H or ASY-H index value. Summarizing, for the SYM-H index we have identified 166 storms of low intensity, 90 of moderate intensity, 23 intense storms, and 7 superintense ones. For the ASY-H index, we have identified 164 storms of low asymmetry, 91 of moderate asymmetry, 23 intense asymmetric storms, and 6 superintense asymmetric storms. The number of storms of each type for each index by year is depicted in Table 4.5, along with the corresponding solar cycle.

Figures 4.12 and 4.13 depict two storms identified using the previous procedure for the SYM-H index. In the first storm the recovery period is considerably large because the index falls below the low intensity threshold multiple times after the first index peak, to

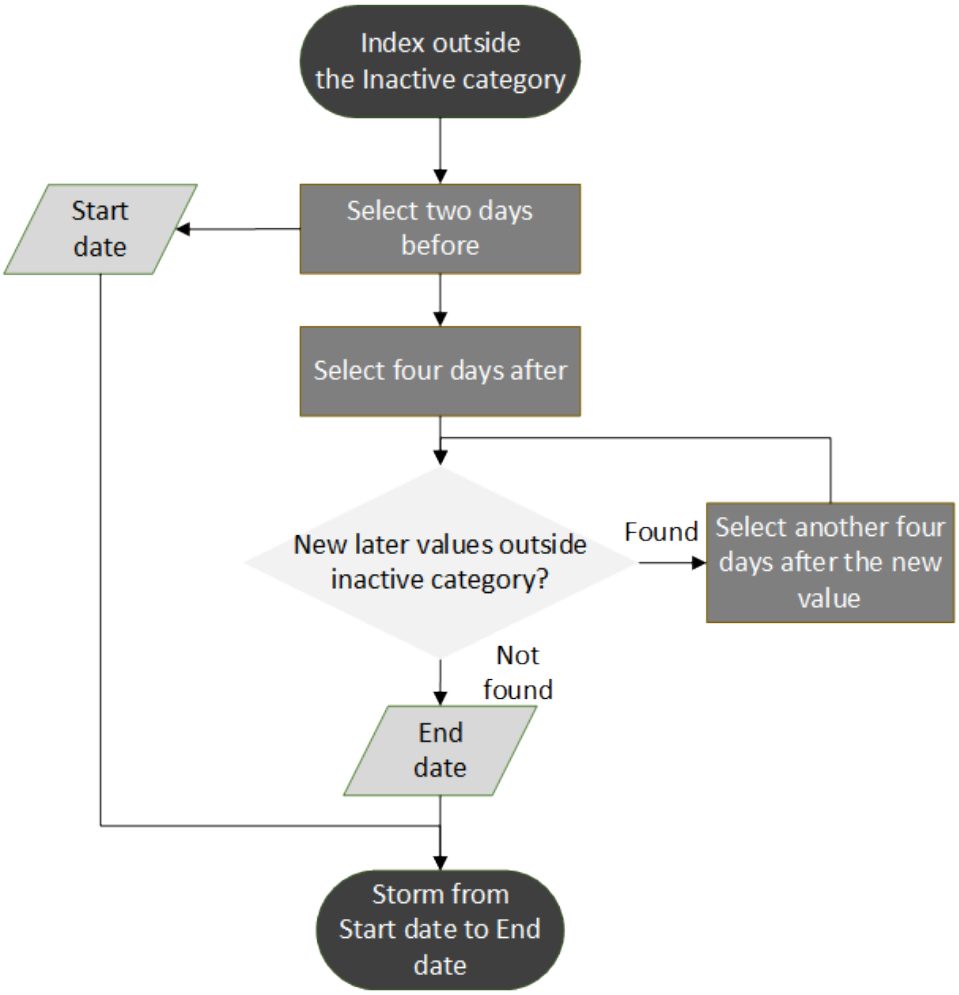


Figure 4.11: Flowchart for the identification of geomagnetic storms.

finally stabilize four days after the last one. The second storm is a regular storm in which there are no previous or subsequent peaks in the SYM-H so the initial and recovery period are not extended. For this particular storm, other approaches to set the storm phases, such as Wharton et al. [145], which rely on the index crossing the threshold of -15 nT (marked with the dotted horizontal line in the plot) would mark a very long initial phase, since the last time the SYM-H was above -15 nT was around 15 days earlier.

Figures 4.14 and 4.15 show two example storms for the ASY-H index. In the first one the recovery period is considerably long, due to the multiple disturbances that caused a complex geomagnetic storm, while in the second figure there are no multiple peaks, leading to the minimal initial and recovery times.

Table 4.5: Number of geomagnetic storms for the SYM-H and ASY-H index per year. The number of storms for each category is represented as the number of storms for the SYM-H | number of storms for the ASY-H index.

Cycle	Year	Low	Moderate	Intense	Superintense
21	1981	3 | 1	4 | 4	3 | 0	0 | 0
	1982	5 | 4	4 | 2	2 | 2	1 | 1
	1983	4 | 5	6 | 2	0 | 1	0 | 0
	1984	8 | 5	3 | 1	0 | 0	0 | 0
	1985	1 | 7	3 | 1	0 | 0	0 | 0
22	1986	4 | 2	1 | 1	1 | 1	0 | 0
	1987	3 | 4	0 | 2	0 | 0	0 | 0
	1988	4 | 7	5 | 3	0 | 1	0 | 0
	1989	6 | 1	4 | 9	3 | 2	1 | 1
	1990	3 | 4	6 | 7	1 | 0	0 | 0
	1991	5 | 2	4 | 3	2 | 2	1 | 2
	1992	9 | 2	5 | 5	1 | 1	0 | 0
	1993	8 | 8	3 | 1	0 | 0	0 | 0
	1994	6 | 5	1 | 5	1 | 0	0 | 0
	1995	7 | 6	1 | 2	0 | 0	0 | 0
23	1996	1 | 4	0 | 2	0 | 0	0 | 0
	1997	7 | 10	1 | 0	0 | 0	0 | 0
	1998	4 | 5	4 | 4	1 | 2	0 | 0
	1999	6 | 7	2 | 2	0 | 0	0 | 0
	2000	6 | 4	5 | 5	3 | 1	0 | 1
	2001	1 | 2	3 | 3	3 | 4	1 | 0
	2002	8 | 3	3 | 3	0 | 0	0 | 0
	2003	4 | 8	3 | 3	0 | 1	2 | 1
	2004	3 | 5	3 | 1	0 | 2	1 | 0
	2005	8 | 3	2 | 4	1 | 1	0 | 0
	2006	2 | 5	1 | 1	0 | 0	0 | 0
	2007	0 | 0	0 | 0	0 | 0	0 | 0
24	2008	1 | 1	0 | 0	0 | 0	0 | 0
	2009	1 | 1	0 | 0	0 | 0	0 | 0
	2010	0 | 2	0 | 0	0 | 0	0 | 0
	2011	4 | 5	1 | 1	0 | 0	0 | 0
	2012	3 | 4	2 | 2	0 | 0	0 | 0
	2013	1 | 3	2 | 2	0 | 0	0 | 0
	2014	3 | 1	0 | 0	0 | 0	0 | 0
	2015	8 | 3	3 | 5	1 | 1	0 | 0
	2016	4 | 7	0 | 1	0 | 0	0 | 0
	2017	0 | 3	2 | 2	0 | 0	0 | 0
	2018	0 | 3	1 | 1	0 | 0	0 | 0
25	2019	0 | 4	0 | 0	0 | 0	0 | 0
	2020	0 | 1	0 | 0	0 | 0	0 | 0
	2021	2 | 2	0 | 1	0 | 0	0 | 0
	2022	3 | 6	0 | 0	0 | 0	0 | 0

Figure 4.12: Example of an identified storm for the SYM-H index, the green shaded area corresponds to the initial phase and the blue shaded area to the recovery phase. The horizontal dashed lines are the thresholds for the different classes.

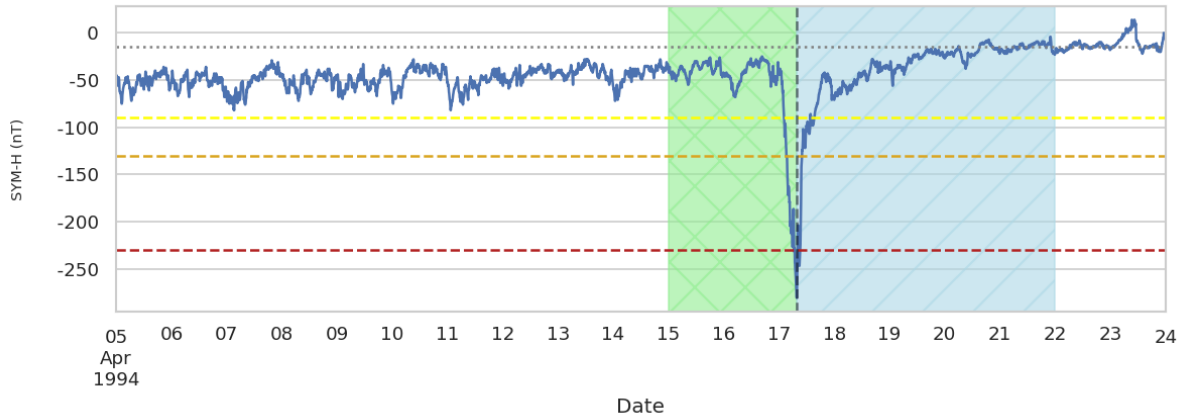


Figure 4.13: Example of an identified storm for the SYM-H index, the green shaded area corresponds to the initial phase and the blue shaded area to the recovery phase. The horizontal dashed lines are the thresholds for the different classes. The horizontal dotted line is the -15 nT mark that has been used by other authors as a threshold to mark the initial phase.

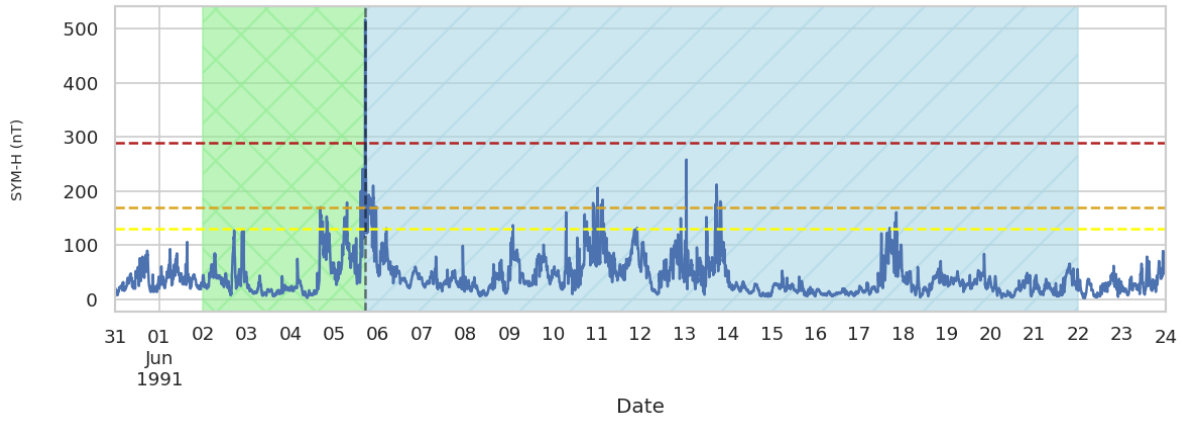


Figure 4.14: Example of an identified storm for the ASY-H index, the green shaded area corresponds to the initial phase and the blue shaded area to the recovery phase. The horizontal dashed lines are the thresholds for the different classes.

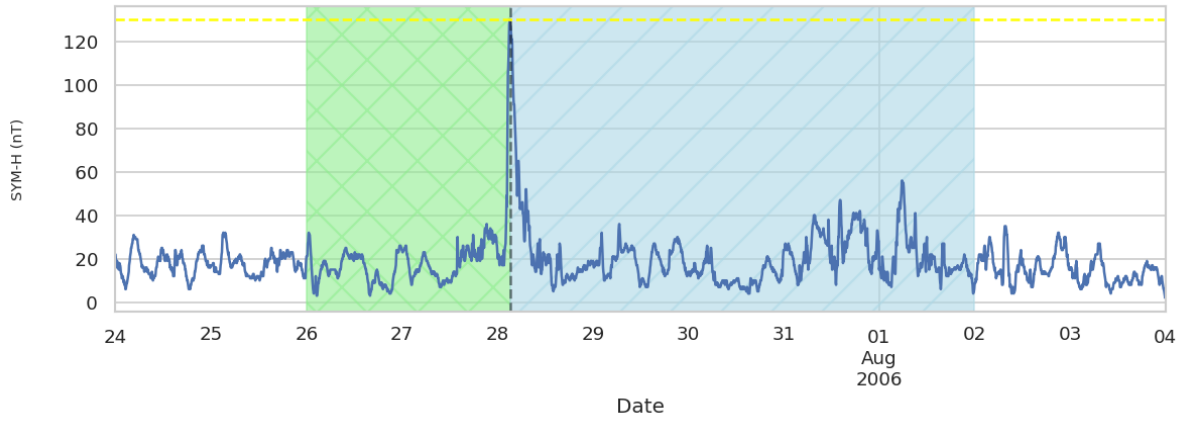


Figure 4.15: Example of an identified storm for the ASY-H index, the green shaded area corresponds to the initial phase and the blue shaded area to the recovery phase. The horizontal dashed lines are the thresholds for the different classes.

4.2 Evaluation framework

Subsequent to Siciliano et al. [36], there has been a consistent practice of using the same selection of geomagnetic storms for model training, validation and testing for the SYM-H index. This approach has been instrumental in facilitating model comparisons. However, it presents an important limitation, as it is mainly focused on storms caused by CMEs, excluding most of the disturbances caused by HSSs, since the disturbance caused on the index is generally minor, but can still have severe consequences [147]. In an effort to overcome this limitation and expand the models' capacity to generalize across a broader spectrum of geomagnetic conditions, we have incorporated additional geomagnetic storms identified in Section 4.1. By doing so, we have substantially increased the number of storms available for model training and validation, enhancing the models' learning capabilities. Notably, we have carefully ensured that the storms included in the previous test set are not used in the updated training set, preserving the ability to make meaningful comparisons between new models and those established in previous research. This expanded dataset enables us to explore the models' performance under various geomagnetic conditions, contributing to a deeper understanding of their predictive capabilities.

While we can use Siciliano's separation strategy for the SYM-H index as a base for the extended dataset, it is important to note that the original separation can not be used for the ASY-H index, as this index exhibits distinct behavior compared to the SYM-H index during geomagnetic storms. Given these considerable differences, it becomes imperative to reconsider the partitioning of geomagnetic storms and to create a tailored separation strategy specifically suited to the ASY-H index, since maintaining the separation of the SYM-H index would yield imbalanced sets.

Finally, it is common that those forecasting models are evaluated over corrected data, being tested in ideal conditions. However, deploying them in operational scenarios, where they are suppose to provide meaningful forecasts, requires to also test the models with incomplete and/or not corrected input data. For such evaluation, we propose the use of a second test set, using ACE's preliminary parameters. These values will be closer to the real-time parameters that the model will process in operational settings, giving a closer evaluation of how the model will perform in the real-time scenario.

To evaluate the forecasting accuracy, previous works have mainly used some of the accuracy metrics collected in Liemohn et al. [148], most notably the RMSE and the R^2. For instance, Tan et al. [68] used the MAE, RMSE and R^2 for assessing the performance of a model that forecasts the Kp index. For the Dst index, Gruet et al. [75] analysed the RMSE and R^2. In the same way, later works that forecast the SYM-H and ASY-H indices evaluated the accuracy using the same two metrics [38], [68], [80]. However, those metrics treat all points in the evaluation equally, calculating the average error across all predictions. This approach, while useful for a general assessment, overlooks a critical aspect of forecasting geomagnetic indices: the varying significance of errors at different activity levels. The error made in forecasting the peak of the storm carries far more importance than errors made during inactive periods, as such, should not be treated equally to the error made during inactive times.

Traditional regression metrics like RMSE and R^2 do not differentiate between errors made during these crucial high-activity periods and less critical low-activity periods. As a result, they fail to adequately reflect the specific challenges and priorities inherent to

geomagnetic index forecasting. To address this gap, we propose the Binned Forecasting Error (BFE) metric, which is described in detail in Section 4.2.3. Before that, section 4.2.1 presents the revised storm sets for the SYM-H and ASY-H indices. Then, section 4.2.2 discusses the traditional metrics used in the field and the limitations they present, motivating the need for the BFE metric. Finally, section 4.2.4 presents a case study comparing the performance of a baseline NN model and a persistence model using the discussed metrics.

4.2.1 Storms sets

The expansion and revision of the initial geomagnetic storm sets for model training has been carefully considered to address specific needs in the predictive accuracy of geomagnetic indices. Recognizing that NNs typically benefit from a larger and more diverse training dataset [149], our approach includes not only high-intensity storms but also a significant number of low and moderate-intensity events. The rationale behind this strategy is twofold: first, to leverage the capabilities of NNs to discern patterns from a broad range of data inputs, and second, to ensure that the predictive model is not biased towards high-intensity geomagnetic disturbances.

While the initial dataset primarily comprised storms driven by CMEs, the extended dataset now encompasses storms induced by solar HSSs as well. The inclusion of these additional events is crucial because HSSs can lead to geomagnetic activity distinct from that produced by CMEs [150]. Low-intensity storms, although less disruptive than their high-intensity counterparts, are still significant for a holistic understanding of SW impacts. Training models predominantly on severe storm data and not testing them on those additional events, might inadvertently skew their predictive capabilities, potentially causing them to overestimate the intensity of lower-scale storms. This is a particular concern in operational settings, when the lower intensity storms are more common and overestimation could lead to unnecessary mitigation measures, resulting in economic and logistical inefficiencies [151].

Moreover, although the end-goal of forecasting models is to achieve real-time operational forecasting, the data for the test storms might not completely replicate real-time conditions, since we are using level 2 corrected data. However, in an effort to replicate real-time scenarios, we are also using preliminary parameters for ACE's SWEPAM and MAG instruments as a second testing set. Particularly, we are using the `AC_K0_MFI` and `AC_K0_SWE` CDAWeb datasets. These measurements, known as "key parameters", are closer to the real-time data our model will eventually process, available through NOAA's Space Weather Prediction Center (SWPC). As of the time of writing, this data extends from 2017/07/01 until 5 days before the present. Considering that, evaluating the performance on this data will offer a closer approximation of real-time model performance. This set will be named "Test key parameters".

In order to systematically partition the available geomagnetic storms into distinct sets for our analysis, we performed a multistep approach. Initially, we categorized the storms into various categories based on the classification framework established in Section 4.1, thereby ensuring that each storm was grouped into four categories according to the index peak: superintense, intense, moderate and low. Subsequently, we further considered the solar cycle during which each storm occurred, recognizing the potential impact of solar cycle variations on geomagnetic storm behavior. To properly perform the separation

process and maintain a representative distribution of storm complexity, even inside each category, we adopted a stratified sampling technique. This technique uses the peak of the index as a complexity metric and takes it into account to partition the storms into three distinct sets: a training set comprising 60% of the storms, a validation set with 10%, and a test set encompassing 30%. This approach to partitioning ensured that each set contained a diverse representation of storms, thereby facilitating the comprehensive evaluation of forecasting models across a wide range of geomagnetic storm scenarios. Stratified sampling has been widely used on ML projects when randomly separating the available data can make sets to not have enough representative samples [152], [153]. Although the category of each storm is already considered, the range within each category is notably wide. Applying this technique ensures balanced intra-category separation.

4.2.1.1 Storm sets for SYM-H index

In an effort to maintain a coherent framework for comparative analysis, we have chosen to preserve the original storm separation as established by Siciliano et al. [36]. This approach allows the comparison with previous models, since the storms that the model is evaluated on have not been used for training in either separation. Building on this baseline, we have integrated all the additional storms that were not previously considered. This expansion is aligned with the original categorization criteria, ensuring that all the sets contain sufficient representative storms and the augmentation of the dataset complements the model training and validation sets. The methodology for adding the supplementary storms to the original sets are as follows:

- The only new intense storm is assigned to the training set.
- There are enough moderate storms in the solar cycle 23 to apply the stratified sampling as explained before.
- For the solar cycle 24, there are only 5 new storms, which are not enough samples to apply stratified sampling, so they are randomly split.
- Since there are sufficient storms of low intensity from both solar cycles, the storms for each cycle will be split using stratified sampling, keeping the distribution of each solar cycle consistent across the sets.
- For the test key parameters set, we will use all storms occurred after 2017-10-01, as this is the earliest date for when we have the key parameters available. By prioritizing the test key parameters set when available, we move the test storm from August, 2018 to the test key parameters set.

The storms assignment taking into account their category, solar cycle, set, and correspondence with the original Siciliano et al. [36] separation is presented in Table 4.6. Table 4.8 presents the relevant information of the selected storms for the SYM-H index. Particularly, the number assigned to the storm, the start and end dates for the storm (the selected time is from 00:00 of the start date until the start the end date at 00:00), the SYM-H peak, the category of the storm according to the SYM-H peak following Section 4.1's criteria, the solar cycle of the storm and whether the disturbance was caused by a CME or HSS, according to the CME catalog in Larrodera and Temmer [154].

Table 4.6: Storm sets for SYM-H index grouped by category and solar cycle.

Category	Solar cycle	Set	Storms
Low	23	Train	24
		Validation	4
		Test	14
	24	Train	15
		Validation	3
		Test	7
	25	Test Key Parameters	5
Moderate	23	Train	15
		Validation	3
		Test	8
	24	Train	4
		Validation	3
		Test	3
		Test Key Parameters	1
Intense	23	Train	6
		Validation	1
		Test	1
	24	Test	1
Superintense	23	Train	1
		Validation	1
		Test	2

To evaluate the consistency of the different sets we have performed a superposed epoch analysis for all the considered storms for the SYM-H index divided by category and set. A superposed epoch analysis is a technique that consists of averaging multiple time series, aligning them on a reference point or event, to enhance the visibility of patterns and variations around that reference point and ease the comparison. In our study, we have used the index peak as the reference point, and considered data 2 days before and 4 days after this peak, as shown in Figure 4.16. It is evident that the patterns in the low, moderate, and intense storm categories are somewhat consistent, even when the most recent storms in the test key parameters are considered. This consistency is also the result of having a sufficient number of samples in these categories. With the exception of the test key moderate storms, for which we only have one single storm. Nevertheless, that storm is fairly similar to the average storm of the other subsets, albeit a bit more intense than the average. However, in the superintense category, the plot shows significantly larger variation and inconsistency. This is caused by the very limited number of samples available. For instance, there is only one storm in the validation set, which is a multi-point storm, making the period before the peak different compared to the other two categories. Multi-point storms are characterized by having more than one significant depression in the geomagnetic index, indicating multiple distinct periods of increased geomagnetic activity within a single storm event, for example multiple CMEs or a CME followed by HSSs. However, the only training storm and the two test storms are fairly similar, both before and after the peak.

To ensure an objective assessment, we have calculated the Mean Absolute Deviation (MADev) for the average storm in each set and compared these values to the overall mean across all storms. These results are shown in Table 4.7. The consistency of MADev values across sets with sufficient samples indicates that our data splitting process was effective,

providing representative examples of each storm category in each set. This makes the separation suitable for training a forecasting model.

Table 4.7: SYM-H MADev for each storm set and category.

Category	Training MADev	Validation MADev	Test MADev	Test Key MADev
Superintense	21.269	61.737	29.274	NA
Intense	4.89	19.881	13.73	NA
Moderate	4.795	5.484	6.238	12.513
Low	2.573	5.693	3.755	4.379

Table 4.8: Details of the SYM-H index storms used to train, validate and test the model.

Storm index	Start date	End date	Solar cycle	Min SYM-H	Category	Source	Correspondence to Siciliano set
Train storms							
1	1998-02-15	1998-02-23	23	-119	Low	CME	Train 1
2	1998-08-04	1998-08-12	23	-168	Moderate	CME	Train 2
3	1998-09-23	1998-09-30	23	-213	Moderate	CME	Train 3
4	1999-02-16	1999-02-24	23	-127	Low	CME	Train 4
5	1999-10-20	1999-10-27	23	-218	Moderate	CME	Train 5
6	1999-11-11	1999-11-18	23	-106	Low	CME	New storm
7	1999-12-11	1999-12-18	23	-97	Low	CME	New storm
8	2000-02-10	2000-02-17	23	-164	Moderate	CME	New storm
9	2000-04-14	2000-04-21	23	-94	Low	CME	New storm
10	2000-06-06	2000-06-13	23	-90	Low	CME	New storm
11	2000-07-13	2000-07-25	23	-335	Intense	CME	Train 6
12	2000-08-09	2000-08-17	23	-235	Intense	CME	Train 7
13	2000-09-15	2000-09-23	23	-196	Moderate	CME	Train 8
14	2000-09-28	2000-10-11	23	-183	Moderate	CME	New storm
15	2000-10-12	2000-10-19	23	-100	Low	CME	New storm
16	2000-10-27	2000-11-03	23	-120	Low	CME	Train 9
17	2000-11-04	2000-11-15	23	-174	Moderate	CME	Train 9
18	2000-11-25	2000-12-04	23	-126	Low	CME	New storm
19	2001-03-17	2001-03-26	23	-165	Moderate	CME	Train 10
20	2001-04-09	2001-04-27	23	-275	Intense	CME	Train 11
21	2001-05-08	2001-05-15	23	-90	Low	CME	New storm
22	2001-10-19	2001-11-12	23	-313	Intense	CME	Train 12 & 13
23	2001-11-22	2001-11-30	23	-233	Intense	CME	New storm
24	2002-04-15	2002-04-25	23	-183	Moderate	CME	New storm
25	2002-05-09	2002-05-17	23	-109	Low	CME	New storm
26	2002-05-21	2002-05-28	23	-113	Low	CME	Train 14
27	2002-07-31	2002-08-07	23	-114	Low	CME	New storm

Continued on next page

Storm index	Start date	End date	Solar cycle	Min SYM-H	Category	Source	Correspondence to Siciliano set
28	2002-09-29	2002-10-13	23	-153	Moderate	CME	New storm
29	2002-11-18	2002-11-26	23	-126	Low	CME	New storm
30	2002-12-19	2002-12-26	23	-90	Low	CME	New storm
31	2003-04-29	2003-05-06	23	-93	Low	HSS	New storm
32	2003-05-08	2003-05-15	23	-91	Low	CME	New storm
33	2003-08-16	2003-08-24	23	-138	Moderate	CME	New storm
34	2003-11-18	2003-11-27	23	-488	Superintense	CME	Train 15
35	2004-03-08	2004-03-15	23	-101	Low	HSS	New storm
36	2004-07-15	2004-08-02	23	-208	Moderate	CME	Train 16
37	2004-08-28	2004-09-05	23	-128	Low	CME	New storm
38	2005-02-16	2005-02-23	23	-95	Low	CME	New storm
39	2005-04-03	2005-04-10	23	-93	Low	HSS	New storm
40	2005-05-06	2005-05-25	23	-302	Intense	CME	Train 17
41	2005-05-28	2005-06-05	23	-126	Low	CME	New storm
42	2005-07-08	2005-07-17	23	-113	Low	CME	New storm
43	2005-08-22	2005-09-05	23	-174	Moderate	CME	New storm
44	2005-09-09	2005-09-17	23	-135	Moderate	CME	New storm
45	2006-04-03	2006-04-19	23	-110	Low	CME	Train 18
46	2006-12-12	2006-12-20	23	-206	Moderate	CME	Train 19
47	2011-03-09	2011-03-16	24	-92	Low	CME	New storm
48	2011-05-26	2011-06-02	24	-93	Low	CME	New storm
49	2011-09-24	2011-10-02	24	-111	Low	CME	New storm
50	2012-03-05	2012-03-14	24	-149	Moderate	CME	Train 20
51	2012-04-21	2012-04-29	24	-125	Low	CME	New storm
52	2012-11-12	2012-11-19	24	-117	Low	CME	New storm
53	2013-03-15	2013-03-23	24	-131	Moderate	CME	New storm
54	2014-02-17	2014-03-05	24	-125	Low	CME	New storm
55	2014-04-10	2014-04-17	24	-91	Low	CME	New storm
56	2015-01-05	2015-01-12	24	-134	Moderate	CME	New storm
57	2015-05-11	2015-05-18	24	-95	Low	CME	New storm
58	2015-08-14	2015-08-21	24	-93	Low	CME	New storm
59	2015-08-25	2015-09-02	24	-100	Low	CME	New storm
60	2015-10-05	2015-10-13	24	-124	Low	CME	New storm
61	2015-11-05	2015-11-12	24	-106	Low	CME	New storm
62	2015-12-18	2015-12-26	24	-169	Moderate	CME	New storm
63	2015-12-29	2016-01-06	24	-117	Low	CME	New storm
64	2016-01-18	2016-01-25	24	-95	Low	CME	New storm

Continued on next page

Storm index	Start date	End date	Solar cycle	Min SYM-H	Category	Source	Correspondence to Siciliano set
65	2016-10-11	2016-10-19	24	-114	Low	CME	New storm
Validation storms							
66	1998-03-08	1998-03-16	23	-119	Low	HSS	New storm
67	1998-05-02	1998-05-10	23	-268	Intense	CME	Val 21
68	1999-09-20	1999-09-28	23	-160	Moderate	CME	Val 22
69	2001-09-21	2001-10-09	23	-187	Moderate	CME	New storm
70	2002-08-19	2002-08-26	23	-119	Low	CME	New storm
71	2003-10-12	2003-10-19	23	-102	Low	HSS	New storm
72	2003-10-27	2003-11-05	23	-427	Superintense	CME	Val 23
73	2004-04-01	2004-04-10	23	-148	Moderate	CME	New storm
74	2005-01-16	2005-01-27	23	-107	Low	CME	New storm
75	2009-07-20	2009-07-27	24	-93	Low	CME	New storm
76	2015-06-20	2015-06-30	24	-207	Moderate	CME	Val 24
77	2016-03-04	2016-03-12	24	-109	Low	CME	New storm
78	2016-05-06	2016-05-14	24	-103	Low	HSS	New storm
79	2017-05-26	2017-06-02	24	-141	Moderate	CME	New storm
80	2017-09-05	2017-09-13	24	-144	Moderate	CME	Val 25
Test storms							
81	1998-06-24	1998-07-01	23	-120	Low	CME	Test 26
82	1998-08-25	1998-09-01	23	-171	Moderate	CME	New storm
83	1998-10-17	1998-10-24	23	-120	Low	CME	New storm
84	1998-11-06	1998-11-19	23	-179	Moderate	CME	Test 27
85	1999-01-11	1999-01-19	23	-111	Low	CME	Test 28
86	1999-02-27	1999-03-06	23	-93	Low	HSS	New storm
87	1999-04-15	1999-04-22	23	-122	Low	CME	Test 29
88	2000-01-21	2000-01-28	23	-101	Low	CME	Test 30
89	2000-04-04	2000-04-12	23	-315	Intense	CME	Test 31
90	2000-05-15	2000-05-29	23	-159	Moderate	CME	Test 32
91	2001-03-29	2001-04-07	23	-434	Superintense	CME	Train 11 & Test 33
92	2001-08-15	2001-08-22	23	-130	Moderate	CME	New storm
93	2002-01-31	2002-02-07	23	-90	Low	HSS	New storm
94	2002-03-22	2002-03-29	23	-114	Low	CME	New storm
95	2002-09-02	2002-09-13	23	-167	Moderate	CME	New storm
96	2003-05-27	2003-06-07	23	-162	Moderate	CME	Test 34
97	2003-06-15	2003-06-23	23	-162	Moderate	CME	New storm
98	2003-07-10	2003-07-21	23	-125	Low	HSS	Test 35
99	2004-01-20	2004-01-27	23	-137	Moderate	CME	Test 36

Continued on next page

Storm index	Start date	End date	Solar cycle	Min SYM-H	Category	Source	Correspondence to Siciliano set
100	2004-02-09	2004-02-16	23	-107	Low	CME	New storm
101	2004-11-05	2004-11-17	23	-393	Superintense	CME	Test 37
102	2005-01-05	2005-01-13	23	-108	Low	CME	New storm
103	2005-06-10	2005-06-18	23	-112	Low	CME	New storm
104	2005-06-21	2005-06-28	23	-101	Low	HSS	New storm
105	2006-08-18	2006-08-25	23	-94	Low	CME	New storm
106	2008-03-07	2008-03-14	24	-99	Low	CME	New storm
107	2011-08-03	2011-08-11	24	-126	Low	CME	New storm
108	2011-10-22	2011-10-30	24	-160	Moderate	CME	New storm
109	2012-07-13	2012-07-21	24	-122	Low	CME	New storm
110	2012-09-29	2012-10-18	24	-138	Moderate	CME	Test 38
111	2013-05-30	2013-06-06	24	-134	Moderate	CME	Test 39
112	2013-06-27	2013-07-04	24	-110	Low	CME	Test 40
113	2014-09-10	2014-09-17	24	-95	Low	CME	New storm
114	2015-03-15	2015-03-23	24	-233	Intense	CME	Test 41
115	2015-06-06	2015-06-13	24	-104	Low	HSS	New storm
116	2015-09-06	2015-09-16	24	-110	Low	CME	New storm
Test key parameters							
117	2018-08-24	2018-08-31	24	-205	Moderate	CME	Test 42
118	2021-08-26	2021-09-02	25	-90	Low	CME	New storm
119	2021-11-02	2021-11-09	25	-117	Low	CME	New storm
120	2022-01-12	2022-01-19	25	-100	Low	HSS	New storm
121	2022-03-11	2022-03-19	25	-113	Low	CME	New storm
122	2022-11-05	2022-11-12	25	-116	Low	HSS	New storm

4.2.1.2 Storm sets for ASY-H index

For the definition of the ASY-H sets, it is important to note that we do not have a established separation to build upon. This necessitates a ground-up approach to separate the geomagnetic storms specifically tailored to the ASY-H index. To address this, we separate the storms in a decreasing manner considering the intensity. The methodology for creating the sets is as follows:

1. For the superintense storms in the ASY-H index only two samples are available on the solar cycle 23.
    - One storm is allocated to the training set.
    - The other storm is assigned to the testing set.

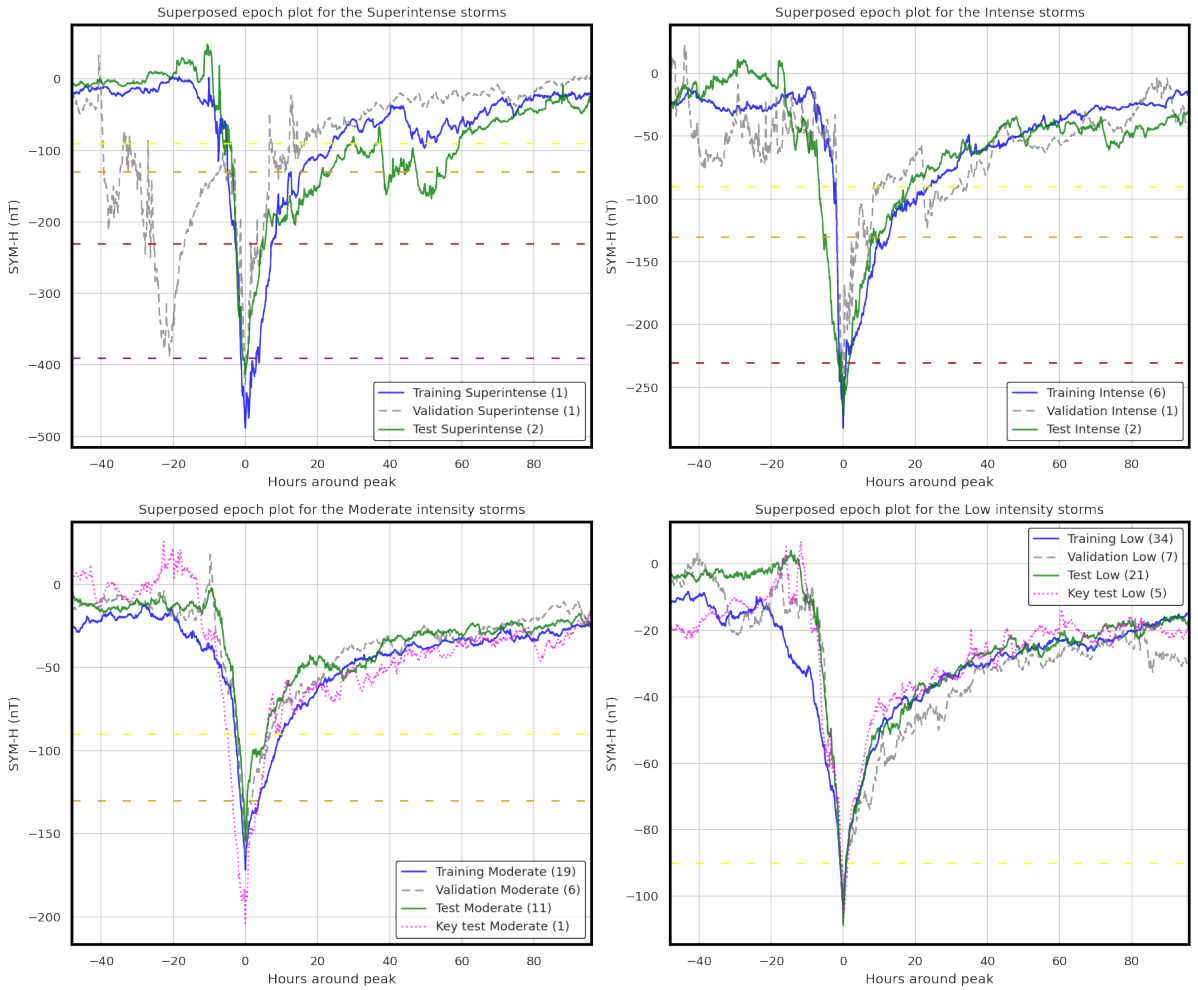


Figure 4.16: Superposed epoch plot centered on the SYM-H peak for all the storms divided by category and set. The number in parenthesis is the amount of storms in each set for that given category. The horizontal lines mark the separation of the storms classes established in Section 4.1.

2. For the intense storms:

   - In solar cycle 24, there is only one intense storm, designated for testing.
   - In solar cycle 23, the remaining intense storms are divided using stratified sampling. Using the maximum ASY-H peak as the complexity metric for the separation.

3. For moderate and low storms in both solar cycles there are enough samples available to apply the stratified sampling technique using the ASY-H peak as the complexity metric. This ensures a uniform distribution of each solar cycle across all sets.

To evaluate the model in the real-time scenario, we reserved all storms after October 2017, for which key parameters are available, for the second test set. This includes two moderate storms and several low-intensity ones, ensuring a thorough evaluation. This test set is crucial for assessing the model's performance in an operational environment using real-time data. The detailed split based on category, solar cycle, and set is presented in Table 4.9.

Table 4.9: Storm sets for ASY-H index grouped by category and solar cycle

Category	Solar cycle	Set	Storms
Low	23	Train	25
		Validation	6
		Test	11
	24	Train	19
		Validation	4
		Test	7
		Test Key Parameters	7
	25	Test Key Parameters	9
Moderate	23	Train	15
		Validation	4
		Test	7
	24	Train	8
		Validation	2
		Test	3
		Test Key Parameters	1
	25	Test Key Parameters	1
Intense	23	Train	6
		Validation	2
		Test	3
	24	Test	1
Superintense	23	Train	1
		Test	1

Table 4.10 details the MADev results for the ASY-H sets while Figure 4.17 depicts the superposed epoch plot for each intensity class. In most instances, the MADev values for ASY-H are lower than those for SYM-H. This indicates a successful data separation, especially considering the inherently more chaotic nature of the ASY-H index compared to SYM-H. Notably, high MADev values are observed in two specific cases: the superintense storms, which have only two extreme samples with unique characteristics, and the intense storm in the validation set, also limited to two samples, which difficulties the average comparison.

Table 4.11 presents the relevant information of the selected storms. The assigned number, the start and end dates for the storm, the ASY-H peak along with the category of the storm according to Section 4.1, the solar cycle of the storm, and source according to the CME catalog.

Table 4.10: ASY-H MADev for each storm set and category.

Category	Training MADev	Validation MADev	Test MADev	Test Key MADev
Superintense	23.377	NA	23.377	NA
Intense	7.548	19.807	10.785	NA
Moderate	2.047	5.257	4.101	10.905
Low	1.622	4.115	2.757	4.214

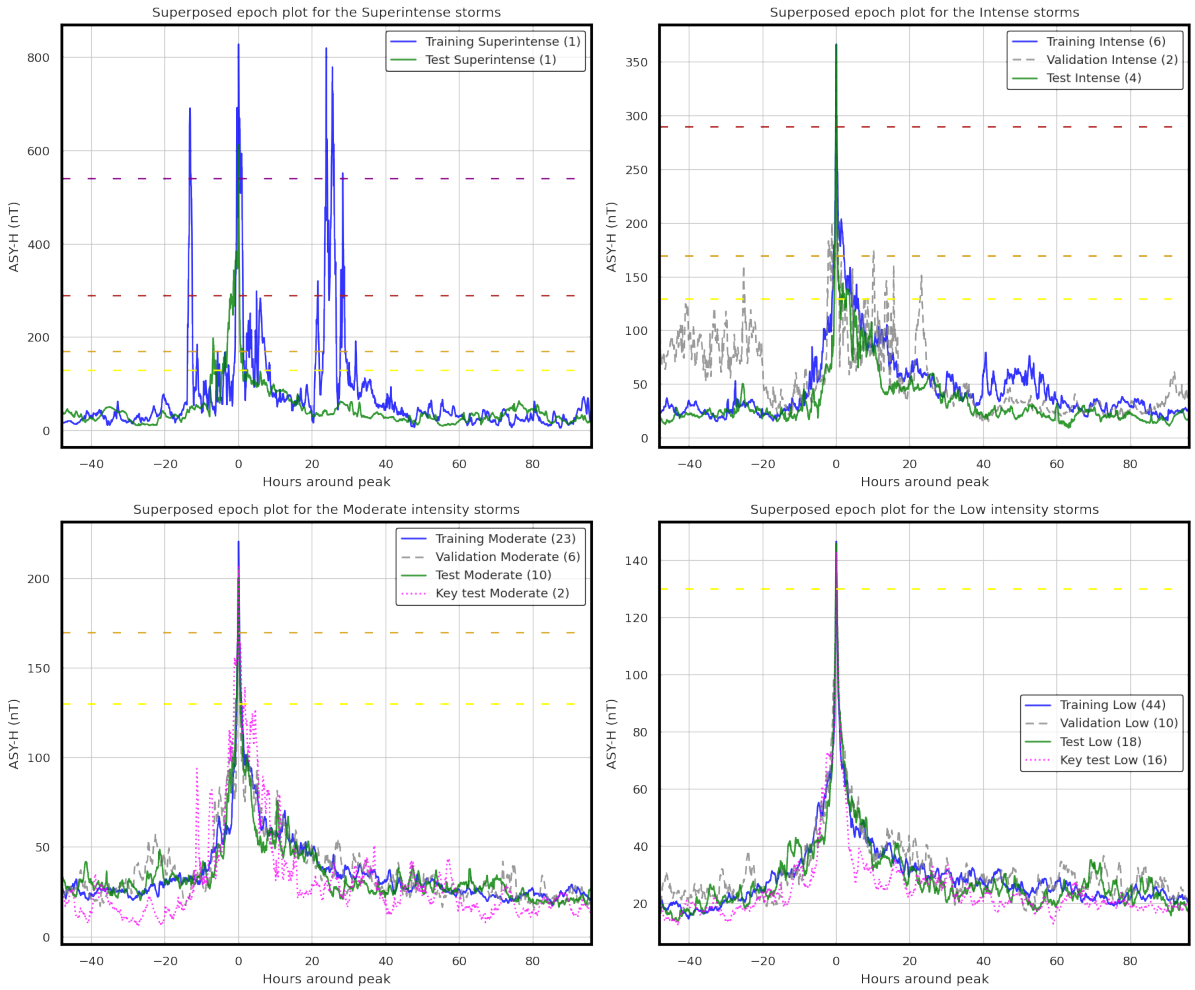


Figure 4.17: Superposed epoch plot centered on the ASY-H peak for all the storms divided by category and set. The number in parenthesis is the amount of storms in each set for that given category. The horizontal lines mark the separation of the storms classes established in Section 4.1.

Table 4.11: Details of the ASY-H index storms used to train, validate and test the model.

Storm index	Start date	End date	Solar cycle	Max ASY-H	Category	Source
Train storms						
1	1998-03-08	1998-03-15	23	255	Moderate	HSS
2	1998-03-19	1998-03-26	23	138	Low	CME
3	1998-05-27	1998-06-03	23	159	Low	CME
4	1998-08-04	1998-08-11	23	136	Low	CME
5	1998-08-25	1998-09-01	23	202	Moderate	CME
6	1998-10-05	1998-10-12	23	143	Low	HSS
7	1998-10-17	1998-10-24	23	156	Low	CME
8	1998-11-06	1998-11-19	23	196	Moderate	CME
9	1999-03-27	1999-04-03	23	157	Low	HSS
10	1999-09-20	1999-09-28	23	157	Low	CME

Continued on next page

Storm index	Start date	End date	Solar cycle	Max ASY-H	Category	Source
11	1999-10-08	1999-10-15	23	140	Low	HSS
12	2000-02-10	2000-02-19	23	176	Moderate	CME
13	2000-05-22	2000-05-29	23	225	Moderate	CME
14	2000-07-12	2000-07-21	23	380	Intense	CME
15	2000-09-15	2000-09-23	23	219	Moderate	CME
16	2000-10-03	2000-10-10	23	258	Moderate	CME
17	2000-11-25	2000-12-04	23	149	Low	CME
18	2001-03-18	2001-03-25	23	225	Moderate	CME
19	2001-03-29	2001-04-05	23	352	Intense	CME
20	2001-04-06	2001-04-23	23	422	Intense	CME
21	2001-08-15	2001-08-22	23	132	Low	CME
22	2001-10-10	2001-10-17	23	143	Low	CME
23	2001-10-19	2001-11-02	23	189	Moderate	CME
24	2001-11-04	2001-11-11	23	329	Intense	CME
25	2002-04-15	2002-04-25	23	189	Moderate	CME
26	2002-05-09	2002-05-16	23	154	Low	CME
27	2002-05-21	2002-05-28	23	204	Moderate	CME
28	2002-09-05	2002-09-13	23	143	Low	CME
29	2003-05-08	2003-05-15	23	160	Low	CME
30	2003-05-27	2003-06-05	23	377	Intense	CME
31	2003-06-26	2003-07-03	23	152	Low	HSS
32	2003-08-04	2003-08-11	23	144	Low	CME
33	2003-10-12	2003-10-19	23	141	Low	HSS
34	2003-10-22	2003-11-25	23	828	Superintense	CME
35	2004-01-04	2004-01-11	23	139	Low	CME
36	2004-02-09	2004-02-16	23	152	Low	CME
37	2004-03-08	2004-03-15	23	143	Low	HSS
38	2004-11-05	2004-11-15	23	339	Intense	CME
39	2004-12-31	2005-01-13	23	279	Moderate	CME
40	2005-05-28	2005-06-04	23	177	Moderate	CME
41	2005-07-16	2005-07-23	23	132	Low	CME
42	2005-09-09	2005-09-17	23	266	Moderate	CME
43	2006-03-17	2006-03-24	23	132	Low	HSS
44	2006-04-12	2006-04-19	23	162	Low	CME
45	2006-11-08	2006-11-15	23	135	Low	HSS
46	2006-11-28	2006-12-05	23	150	Low	CME
47	2006-12-12	2006-12-20	23	267	Moderate	CME
48	2010-05-27	2010-06-03	24	130	Low	CME

Continued on next page

Storm index	Start date	End date	Solar cycle	Max ASY-H	Category	Source
49	2011-02-27	2011-03-06	24	152	Low	CME
50	2011-03-09	2011-03-16	24	130	Low	CME
51	2011-05-27	2011-06-10	24	150	Low	CME
52	2011-08-03	2011-08-10	24	183	Moderate	CME
53	2011-09-24	2011-10-04	24	150	Low	CME
54	2012-04-21	2012-04-29	24	151	Low	CME
55	2012-06-15	2012-06-22	24	184	Moderate	CME
56	2012-09-29	2012-10-18	24	158	Low	CME
57	2013-02-27	2013-03-06	24	149	Low	CME
58	2013-03-15	2013-03-22	24	171	Moderate	CME
59	2013-05-30	2013-06-06	24	154	Low	CME
60	2013-06-27	2013-07-04	24	165	Low	CME
61	2014-02-17	2014-02-24	24	152	Low	CME
62	2015-01-05	2015-01-12	24	212	Moderate	CME
63	2015-03-15	2015-03-24	24	250	Moderate	CME
64	2015-08-13	2015-08-21	24	204	Moderate	CME
65	2015-08-25	2015-09-01	24	149	Low	CME
66	2015-12-18	2015-12-25	24	145	Low	CME
67	2016-05-07	2016-05-14	24	133	Low	HSS
68	2016-08-31	2016-09-07	24	133	Low	HSS
69	2016-10-23	2016-10-30	24	161	Low	CME
70	2016-11-23	2016-11-30	24	166	Low	HSS
71	2017-03-25	2017-04-01	24	143	Low	CME
72	2017-07-14	2017-07-21	24	175	Moderate	CME
73	2017-09-05	2017-09-13	24	230	Moderate	CME
74	2017-09-26	2017-10-03	24	163	Low	HSS
Validation storms						
75	1998-04-30	1998-05-10	23	415	Intense	CME
76	1999-02-16	1999-02-24	23	246	Moderate	CME
77	1999-10-20	1999-10-27	23	214	Moderate	CME
78	1999-11-05	1999-11-13	23	141	Low	CME
79	2000-08-10	2000-08-17	23	206	Moderate	CME
80	2000-11-04	2000-11-11	23	130	Low	CME
81	2002-11-19	2002-11-26	23	142	Low	CME
82	2003-01-20	2003-01-27	23	162	Low	CME
83	2003-09-15	2003-09-22	23	171	Moderate	HSS
84	2004-01-20	2004-01-27	23	150	Low	CME
85	2004-03-26	2004-04-10	23	143	Low	CME

Continued on next page

Storm index	Start date	End date	Solar cycle	Max ASY-H	Category	Source
86	2004-07-20	2004-08-01	23	294	Intense	CME
87	2012-03-05	2012-03-17	24	229	Moderate	CME
88	2012-07-13	2012-07-20	24	164	Low	CME
89	2015-09-18	2015-09-25	24	135	Low	CME
90	2016-01-18	2016-01-25	24	140	Low	CME
91	2016-12-19	2016-12-26	24	171	Moderate	HSS
92	2017-05-26	2017-06-02	24	158	Low	CME
Test storms						
93	1998-02-15	1998-02-23	23	188	Moderate	CME
94	1998-09-23	1998-09-30	23	395	Intense	CME
95	1999-01-11	1999-01-19	23	147	Low	CME
96	1999-02-27	1999-03-06	23	167	Low	HSS
97	1999-04-15	1999-04-22	23	138	Low	CME
98	2000-04-04	2000-04-12	23	612	Superintense	CME
99	2000-06-06	2000-06-13	23	133	Low	CME
100	2000-10-11	2000-10-19	23	142	Low	CME
101	2001-09-23	2001-10-08	23	184	Moderate	CME
102	2001-11-22	2001-11-29	23	330	Intense	CME
103	2002-09-28	2002-10-09	23	199	Moderate	CME
104	2003-03-15	2003-03-22	23	130	Low	CME
105	2003-04-28	2003-05-05	23	134	Low	HSS
106	2003-06-07	2003-06-21	23	198	Moderate	CME
107	2003-07-10	2003-07-17	23	172	Moderate	HSS
108	2003-08-16	2003-08-23	23	160	Low	CME
109	2005-01-15	2005-01-26	23	392	Intense	CME
110	2005-04-03	2005-04-10	23	163	Low	HSS
111	2005-05-06	2005-05-20	23	250	Moderate	CME
112	2005-07-08	2005-07-15	23	158	Low	CME
113	2005-08-22	2005-09-05	23	234	Moderate	CME
114	2006-07-26	2006-08-02	23	130	Low	HSS
115	2008-09-02	2008-09-09	24	159	Low	CME
116	2009-07-20	2009-07-27	24	131	Low	CME
117	2010-04-03	2010-04-10	24	151	Low	CME
118	2011-10-22	2011-10-30	24	150	Low	CME
119	2012-01-20	2012-01-27	24	132	Low	CME
120	2013-09-30	2013-10-07	24	171	Moderate	CME
121	2015-06-20	2015-06-28	24	348	Intense	CME
122	2015-09-07	2015-09-16	24	189	Moderate	CME

Continued on next page

Storm index	Start date	End date	Solar cycle	Max ASY-H	Category	Source
123	2015-10-05	2015-10-13	24	216	Moderate	CME
124	2016-09-27	2016-10-04	24	154	Low	HSS
125	2016-11-08	2016-11-15	24	148	Low	CME
Test key parameters						
126	2018-04-18	2018-04-25	24	136	Low	HSS
127	2018-08-24	2018-08-31	24	197	Moderate	CME
128	2018-09-09	2018-09-16	24	140	Low	HSS
129	2018-11-03	2018-11-10	24	131	Low	CME
130	2019-05-09	2019-05-16	24	134	Low	CME
131	2019-08-03	2019-08-10	24	139	Low	CME
132	2019-08-29	2019-09-05	24	146	Low	HSS
133	2019-10-23	2019-10-30	24	143	Low	CME
134	2020-04-18	2020-04-25	25	143	Low	CME
135	2021-05-10	2021-05-17	25	146	Low	CME
136	2021-10-10	2021-10-17	25	167	Low	CME
137	2021-11-01	2021-11-09	25	216	Moderate	CME
138	2022-02-01	2022-02-08	25	153	Low	CME
139	2022-03-11	2022-03-18	25	139	Low	CME
140	2022-04-08	2022-04-15	25	138	Low	CME
141	2022-09-02	2022-09-09	25	148	Low	CME
142	2022-11-05	2022-11-12	25	139	Low	HSS
143	2022-12-05	2022-12-12	25	144	Low	HSS

4.2.2 Forecasting assessment metrics

Previous works on the SW field have proposed different metrics to evaluate the performance of the forecasting models. For instance, Liemohn et al. [59] classified the metrics on fit performance metrics, which calculates the error made by the model on each time-step compared to the observed time series; and event detection metrics, which focus on whether the model would have predicted the existence of a storm, comparing the forecast against an event threshold. Camporeale [57] also gave an overview of different metrics for SW related models. Then, geomagnetic indices forecasting works [36], [38], [80] have mainly used the RMSE, defined in Equation 2.1, and the Coefficient of determination (R^2), defined in Equation 2.2, being the RMSE the key metric of the assessment. For both metrics, y is the observed value, $\bar{y}$ is the average of the observed value and $\hat{y}$ is the prediction for N samples.

When evaluating geomagnetic indices forecasting models, the performance has often been evaluated using the fit performance metrics, which are point-to-point. These metrics, while useful, fail to provide a complete picture of model performance across variable storm duration and intensities. As identified by Siciliano et al. [36] and evidenced later

in Section 4.2.4, these metrics can mask the model's effectiveness during critical storm peaks due to their sensitivity to the inclusion of extended calm periods, which inherently yields lower error values. The average RMSE across the entire storm, therefore, skews towards the model's ability to forecast these quiet periods rather than the peak activity, potentially overshadowing the model's accuracy during the more intense storm phases.

When comparing different models, several issues arise. A fair comparison requires evaluating models on shared test storms. Additionally, the considered storm duration has a large influence on the evaluated metrics. Typically, error metrics are lower during periods of low activity. Therefore, a longer evaluation period, which includes more low-activity periods, can artificially lower the overall prediction error. For example, in the comparison made by Siciliano et al. [36] with the models of Cai et al. [155] and Bhaskar and Vichare [80], they were restricted to only one test storm. However, the considered storm duration were different between the models. Cai et al. [155] considered less days for their evaluated storm while Bhaskar and Vichare [80] considered a longer evaluation period. This discrepancy makes impossible to perform a rigorous comparison, as models covering longer periods tend to yield a lower RMSE value, benefiting from the inclusion of more predictable, quieter periods. Nevertheless, both the RMSE and R^2 remain essential when evaluating the performance of the forecasting model and should be included. However, they need to be compared when they evaluated the same temporal frame, ensuring that variations in performance due to considering different evaluation periods are not present.

4.2.3 Binned Forecasting Error metric

To address the limitations mentioned in the previous section, we introduce the Binned forecasting Error (BFE) metric. Its primary goal is to offer a robust measure of the model's performance that is resilient to variations in the duration of geomagnetic storms considered for evaluation, while also providing insights on the model's capabilities when forecasting at different intensity levels. This metric aims to facilitate the comparison between different models, even when the exact dates of storms differ. It addresses the inherent bias present in RMSE and R^2 metrics, which can be skewed by prolonged periods of inactivity or quiet conditions within the storm's timeline. By adjusting for these quiet periods, BFE provides a more balanced and accurate assessment of a model's capability to predict storm peaks, ensuring that the true predictive power of the model during critical storm phases is adequately represented.

The calculation of the BFE is as follows:

1. **Calculate the Absolute Difference (AD):** compute the absolute differences between the observed values and their corresponding predicted values.

$$\mathrm{AD} = |y_i - \hat{y}_i| \tag{4.1}$$

   where y_i is the observed value, $\hat{y}_i$ is the predicted value.

2. **Group differences into bins:** organize the observed values y_i into bins of size 10 nT. To maintain consistency, the bins are always structured in tens with a left closed interval, such as $\{[-10, 0), [0, 10), [10, 20), ...\}$. For each bin b, calculate the Mean

Absolute Difference (MAD):

$$\mathrm{MAD}_b = \frac{1}{N_b} \sum_{i \in b} \mathrm{AD} \tag{4.2}$$

where N_b is the number of observations in bin b. This offers an insight into the prediction accuracy for the different intensities of the storm. This analysis allows to separate the computation on different intensity levels, separating the accuracy of the model during inactive times to more intense periods.

3. **Calculate the Mean of MAD across all bins:**

$$\mathrm{BFE} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{MAD}_b \tag{4.3}$$

where B is the total number of bins, and MAD_b is the mean absolute difference for bin b. Finally, we obtain the BFE by averaging the bin-wise MAD values. This gives us a single value that represents the overall prediction accuracy across all intensity levels.

Figure 4.18 depicts the BFE computation on the test storm of March 2005 using a persistence model. The BFE offers a visually intuitive representation of the model accuracy, significantly enhancing our understanding and providing a more human-friendly assessment. In the previous figure, the visual representation of the BFE highlights where the model is most and least accurate. This visualization offers significant insights, especially in aspects that might be overlooked by the RMSE and R^2 metrics, where the error is uniformly weighted across all the predictions. In comparison, the BFE reveals the specific intensity ranges where errors are most significant, such as during the critical, high-intensity phases of a storm. This level of detail is crucial for operational SW forecasting, where predicting intense storm peaks accurately is more critical than overall performance. Additionally, the BFE provides a more nuanced evaluation of a model's performance, compared to the average RMSE. For example, a model with a lower overall RMSE but a worse BFE score may perform adequately in less active conditions but inadequately during high-intensity events. These events, though infrequent and short, have a minimal impact on the overall RMSE but are the most relevant ones. This differentiation is essential in geomagnetic indices forecasting, emphasizing the importance of accurately predicting intense events over consistent performance across varying conditions.

Additionally, one of the most important problems, as highlighted by Sciliano et al. [36], is the difficulty to compare different models when they are not perfectly aligned in terms of the time periods they evaluate. As discussed before, the RMSE is sensitive to the duration of the storm, as exemplified in Figure 4.19, where an extension of the time frame around the storm leads to a substantial reduction in the RMSE, of around 30%. Instead, the proposed BFE exhibits remarkable stability with only a marginal difference of less than 0.3%. This robustness makes the proposed metric an interesting option for comparative analysis, allowing for consistent evaluations across models even when the exact time frames differ.

By incorporating BFE alongside traditional metrics, we can achieve a more comprehensive understanding of a model's capabilities. It reveals the model's strengths and

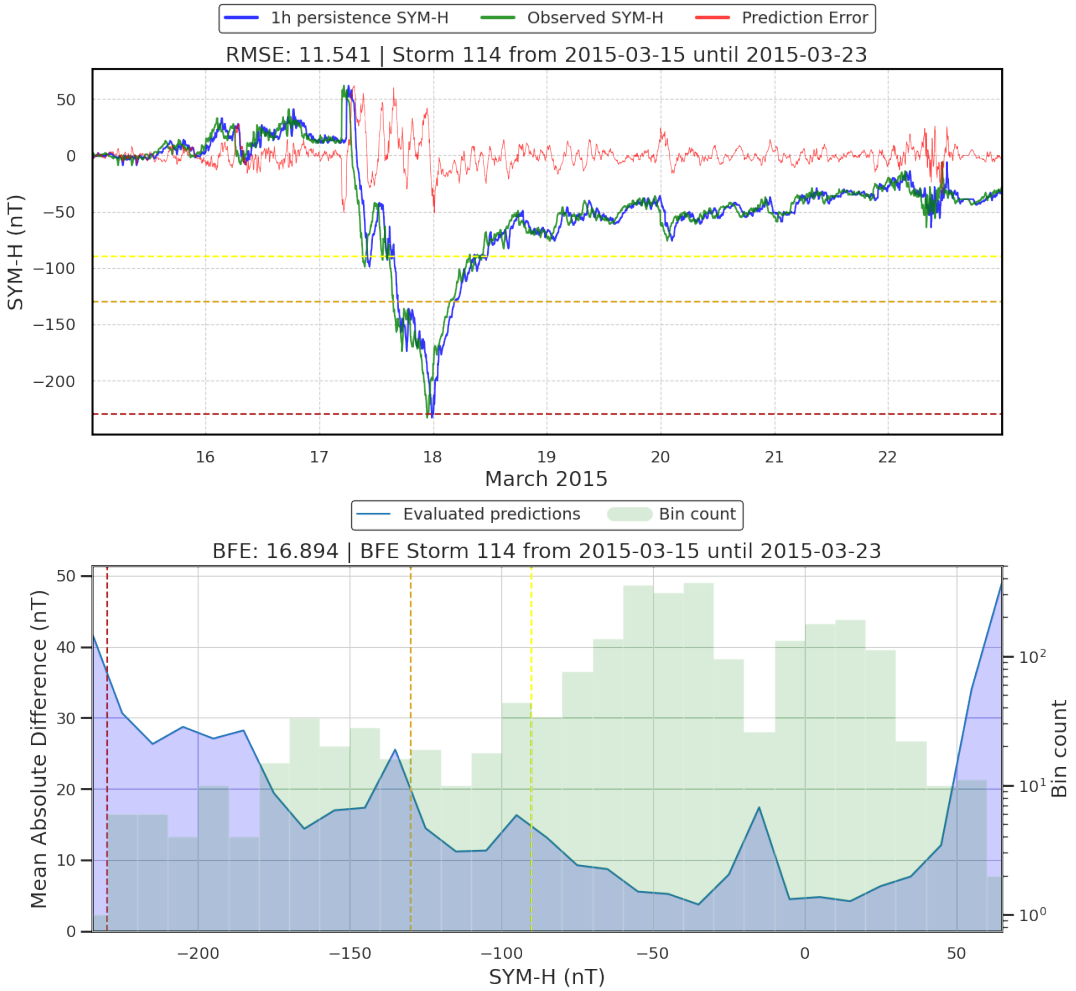


Figure 4.18: Example of the evaluation of the BFE on the Storm of March, 2005 for the 1 hour persistence of the SYM-H index. On the top figure the observed values are displayed in green, the persistence in blue and the error in red. The bottom figure depicts the evaluation of the BFE, the bin-wise MAD values are displayed in blue, the green shaded histogram shows the bin count in a logarithmic scale. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

weaknesses in areas where RMSE alone might not suffice, ensuring the model's reliability not just on average, but particularly in moments of critical importance.

4.2.4 Case study and discussion

In this section, we provide a comparative study between a baseline NN and a persistence model that employs the last known value of the geomagnetic index as its prediction, using the proposed geomagnetic storm sets and evaluated on the commonly used metrics and the proposed BFE.

The architecture for the baseline model is similar to the one used by Siciliano et al. [36]: it consists on an LSTM layer of 128 units followed by three of dense layers, the first two with 128 units using the ELU activation function and the last one, the output, using only one neuron without any activation function. To train our baseline model, we employed

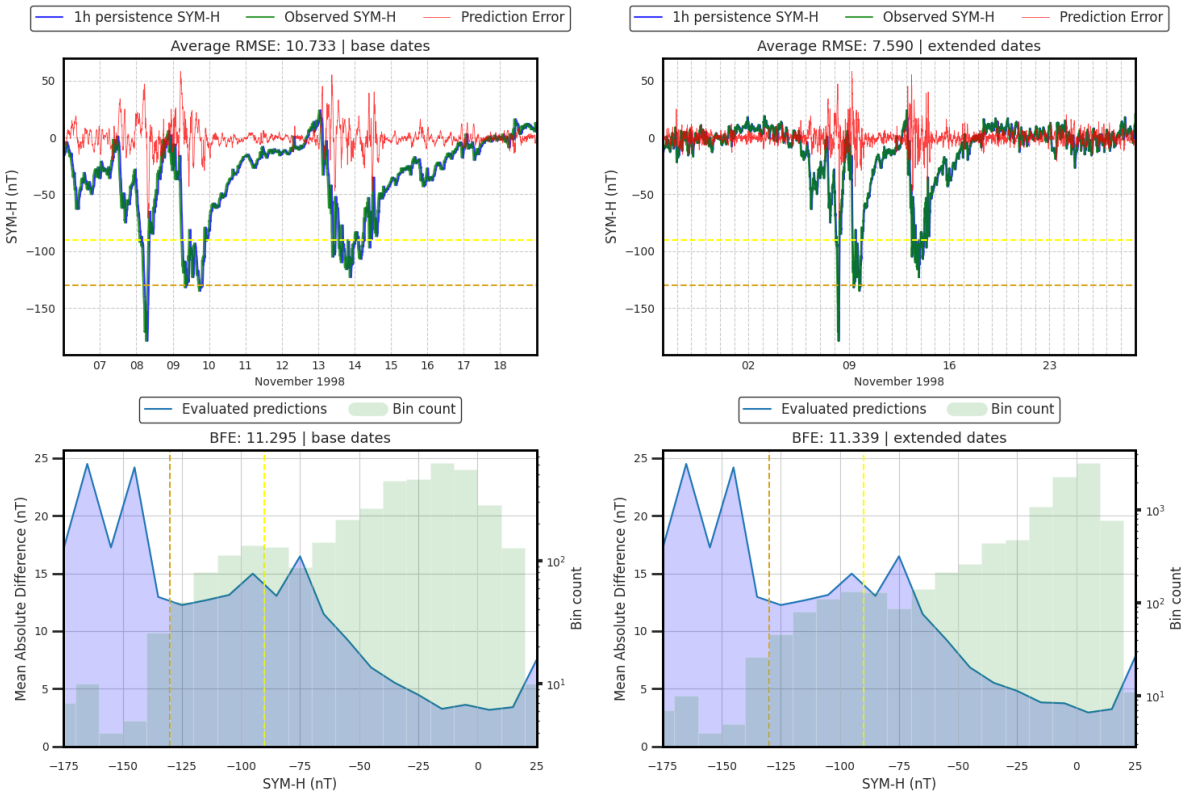


Figure 4.19: Geomagnetic storm of November 1998. Comparison of the RMSE and BFE with the start and end times of the storm assigned in Section 4.1 and extending them by 10 days.

the MAE as the loss function and the Adam optimizer (using a learning rate of 3^{-4}). The model was trained using the storm data sets described in section 4.2.1. The validation set was used to tune hyperparameters and prevent overfitting. Finally, the model is evaluated on the test set and compared to the persistence model. The model is trained for 100 epochs or until the loss on the validation set no longer improves for at least 10 epochs, restoring the weights to those that yielded the best performance on the validation set.

The persistence model, uses the most recent value of the geomagnetic index as its prediction for the next time step. Given its naive nature, the persistence model doesn't require any training. Despite its simplicity, the persistence model often serves as a tough baseline for many time series forecasting problems.

4.2.4.1 Data description and pre-processing

In line with the latest works [37], [38] we have used the IMF and solar wind plasma parameters measured by ACE. Particularly, for each storm we use the following input features:

- IMF: We use the Magnitude of the magnetic field and its X, Y and Z components from the ACE's MAG instrument. For the test storms we use the dataset `AC_H0_MFI` which provides said features with a resolution of 16-seconds. For the storms where the key parameters are available we use the measurements in the dataset `AC_K0_MFI` which provides the measurements in 5-minute intervals.

- Plasma: We use the proton density, speed and temperature from ACE's SWEPAM suite. For the test storms we use the dataset `AC_H0_SWE` which provides said features with a resolution of 64-seconds. For the test key storms we use the data from the `AC_K0_SWE` dataset which, similar to the IMF, providing the measurements in 5-minute intervals. The missing values in the plasma datasets have been filled using interpolation if the value to interpolate to is known; otherwise the last valid value is propagated forward.
- Derived parameters: We also use the y component of the electric field: $E_y = V_x B_z$, where V_x is the x component of the bulk proton speed and B_z is the z component of the IMF and the dynamic pressure, defined as $P = \rho V_x^2$. Both features are related to the geomagnetic indices as evidenced by Iong et al. [38].
- Geomagnetic index: the SYM-H or ASY-H index, depending on the model, expressed in nT. The indices are the variables that we are forecasting, but we will also use the index's values up to the time of the prediction as another input feature. Both indices are retrieved form the `OMNI_HRO_5` with a 5-minute resolution.

The input features are grouped into 5 minutes averages to match the indices resolution. For the baseline model we are using the last 3 hours of the input features. Each index has been trained using their respective training sets presented in section 4.2.1.

4.2.4.2 SYM-H

When evaluating the performance of the network across multiple test storms a summarized evaluation can be done in two ways: averaging the results of the metrics for all the storms separately or concatenating all the evaluated periods and perform the computation over the whole dataset. While both approaches obtain similar results for the RMSE and R^2 metrics, the BFE changes considerably. This is due to the relevance given by the metric to the most intense storms, which is emphasized when computing the metric over the whole dataset and depreciated when averaging the metric across storms.

Tables 4.12 and 4.13 present the comparison of the baseline model with the persistence forecast for 1 hour ahead forecasts for the test and test key storms, respectively. The Mean rows calculate the metrics by averaging the individual storm metrics while the Global rows perform the evaluation over the whole test storms dataset. The baseline model achieves a global RMSE of 7.446 across all the test storms, a R^2 of 0.959, and a BFE of 20.818, significantly better than the persistence model that has a RMSE of 9.89, a R^2 of 0.928 and a BFE of 31.795. The situation is similar for the test key storms, where the baseline model consistenly performs better than the persistence.

Table 4.12: Metrics for the 1-hour forecast of the SYM-H index over the test storms, comparing the baseline and persistence models.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
81	8.314	13.657	5.481	8.066	0.926	0.839
82	14.167	16.261	8.832	10.814	0.951	0.926
83	6.852	9.062	6.633	9.027	0.945	0.898
84	9.405	11.295	8.683	10.733	0.942	0.911
85	6.715	7.337	4.856	6.235	0.963	0.939
86	4.488	6.146	4.811	7.257	0.965	0.921
87	8.486	12.198	5.885	7.994	0.937	0.884
88	6.122	8.136	5.72	6.564	0.953	0.938
89	11.215	18.988	7.229	11.855	0.983	0.955
90	8.856	11.634	6.316	9.988	0.959	0.897
91	26.678	36.269	15.429	21.449	0.962	0.927
92	15.194	20.932	8.821	10.372	0.855	0.800
93	6.905	8.038	6.127	7.968	0.930	0.882
94	6.302	7.878	5.013	6.473	0.964	0.939
95	7.51	10.483	6.351	9.647	0.969	0.929
96	20.108	20.425	11.399	11.177	0.754	0.763
97	10.027	15.483	6.997	9.524	0.932	0.874
98	8.651	11.119	6.829	8.455	0.914	0.868
99	12.224	16.525	8.86	10.742	0.892	0.841
100	5.1	8.234	4.982	6.682	0.945	0.902

Continued on next page

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
101	14.218	25.157	12.512	18.624	0.974	0.942
102	8.042	11.806	5.906	8.524	0.898	0.788
103	10.328	11.631	6.558	8.990	0.931	0.870
104	6.974	8.562	5.342	5.712	0.932	0.922
105	4.585	7.197	5.051	6.188	0.917	0.876
106	5.42	10.558	4.215	6.532	0.929	0.829
107	17.698	14.456	6.531	8.163	0.931	0.892
108	15.865	16.596	7.121	8.799	0.945	0.916
109	8.812	9.287	6.789	7.867	0.965	0.953
110	6.009	8.519	5.475	7.016	0.955	0.927
111	7.778	10.185	5.24	8.046	0.960	0.906
112	5.533	6.372	4.778	5.571	0.966	0.954
113	10.783	19.805	5.991	9.038	0.872	0.710
114	13.582	16.894	8.753	11.541	0.965	0.939
115	14.306	17.068	7.005	8.049	0.917	0.890
116	6.015	7.743	5.992	8.037	0.933	0.879
Mean	9.98	13.109	6.903	9.103	0.959	0.89
Global:	20.818	31.795	7.446	9.89	0.959	0.928

Table 4.13: Metrics for the 1-hour forecast of the SYM-H index over the test key storms, comparing the baseline and persistence models.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
117	9.615	11.864	6.211	7.744	0.972	0.956
118	4.371	5.235	4.285	5.367	0.944	0.913
119	12.666	13.464	8.03	9.383	0.877	0.831
120	3.993	6.514	4.152	5.735	0.964	0.932
121	14.493	21.112	6.723	9.147	0.831	0.688
122	6.067	7.881	4.063	5.392	0.964	0.937
Mean	8.534	11.012	5.577	7.128	0.925	0.876
Global:	12.408	13.606	5.803	7.378	0.946	0.912

Figure 4.20 depicts the evaluation of the BFE made by the baseline model on the next hour for the test storms. It provides some interesting information:

- Error peaks at extreme values: the model has the biggest errors on the lowest and highest SYM-H values. This indicates that the model struggles to predict during intense geomagnetic disturbances. It is mainly caused because these extreme events

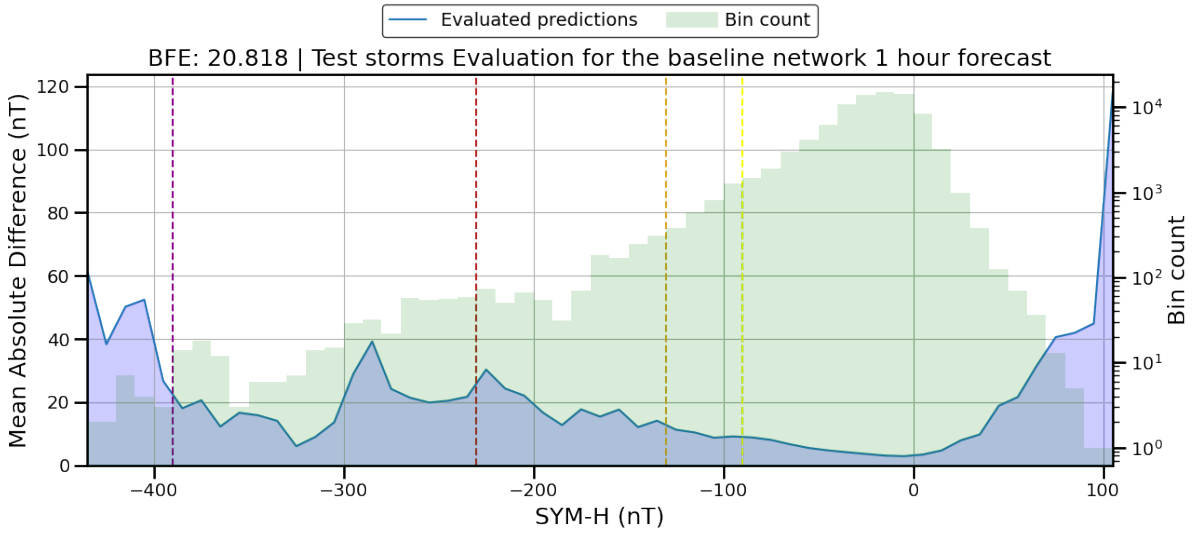


Figure 4.20: BFE evaluated on the predictions over the test storms made by the baseline model for the SYM-H index forecasting the next hour. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

are relatively rare, making the model less exposed to such samples during training. It also shows that the model fails to predict the sudden spikes in the SYM-H before the storm, known as sudden storm commencement.

- Consistent Low Errors at Mid-Range Values: the model performs consistently well for mid-range SYM-H values, as evidenced by the relatively low and stable error values in the range between -200 and 0 nT.

- Sample Distribution Insight: the histogram (in green) provides an indication of the distribution of the observed SYM-H values in the test storms in a logarithmic scale. Most of the samples are concentrated around 0 nT, which explain the model's better performance in that region, since those are the values that the has been trained on the most. This great imbalance also contributes to the fact that only using the RMSE as the performance metric skews the results.

Figure 4.21 compares the BFE on the baseline model and the persistence model. The red curve represents the MAE of the persistence model across different SYM-H values. The blue curve denotes the MAD of the baseline network model. The shaded blue region between the two curves provides a visual representation of the difference in errors between the two models. They can be objectively compared subtracting the BFE of one model from that one of the other. In this case the baseline model consistently outperforms the persistence, except for one bin around the -300 nT mark and some bins that mark the sudden storm commencement that the model, generally, fails to forecast.

Figure 4.22 presents the BFE calculated on the test key storms, for which we only have ACE's preliminary measurements. The performance is similar to the complete test storms, albeit slightly higher error values, which is to be expected considering the provisional nature of the data used. We have chosen to separate the calculations and depictions of the BFE between the test storms using the definitive parameters and those using the

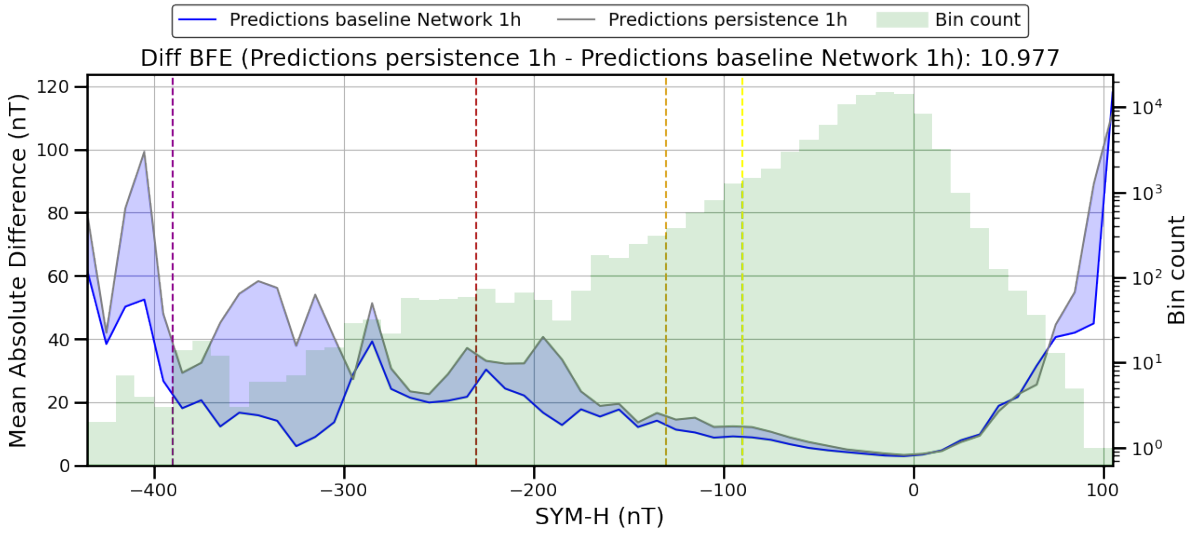


Figure 4.21: Comparison of the BFE on the predictions on the test storms made by the baseline model for the SYM-H index forecasting the next hour compared to the persistence model. The vertical colored lines mark the separation of the different intensities categories as classified in Section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

preliminary ones because the difficulty of forecasting the index using the later is greater, making the direct comparison not entirely fair.

Another application of the BFE is to compare different forecasting horizons for the same model. This is crucial for assessing a model's predictive reliability over varying future intervals. This is illustrated in Figure 4.23, where the baseline model's forecasting capabilities at one-hour and two-hours horizons are compared. While the differences in forecasting accuracy during non-active periods are relatively minor, the disparity becomes pronounced on the more extreme values of the index. This observation is critical because it demonstrates that the model accuracy is notably lower when forecasting further in the future, which is more challenging.

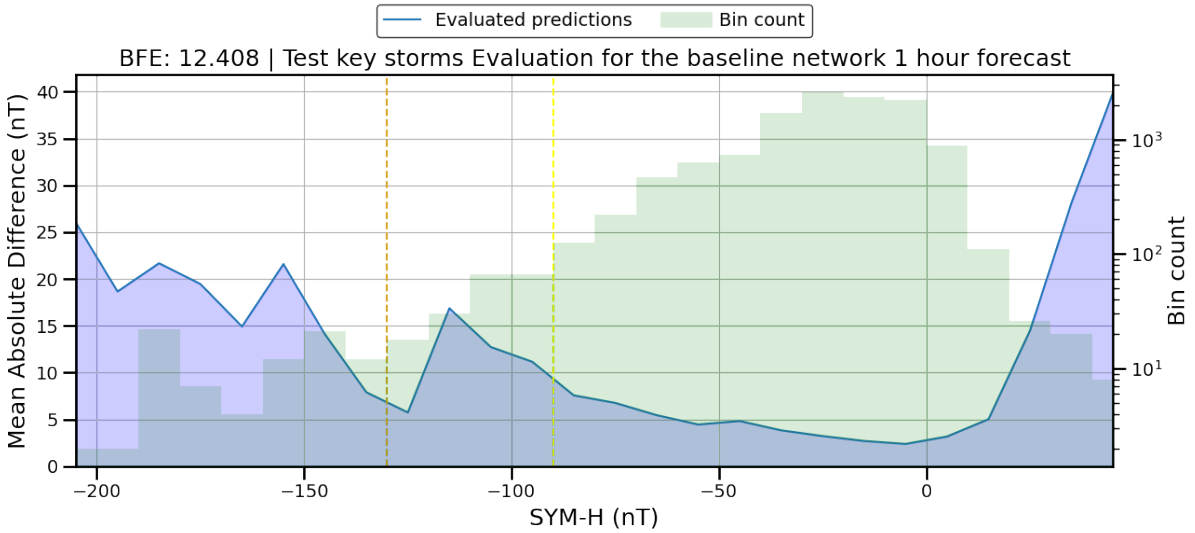


Figure 4.22: Evaluation of the BFE on the predictions on the test key parameters storms made by the baseline model for the SYM-H index forecasting the 1 hour ahead. The vertical colored lines mark the separation of the different intensities categories as classified in Section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

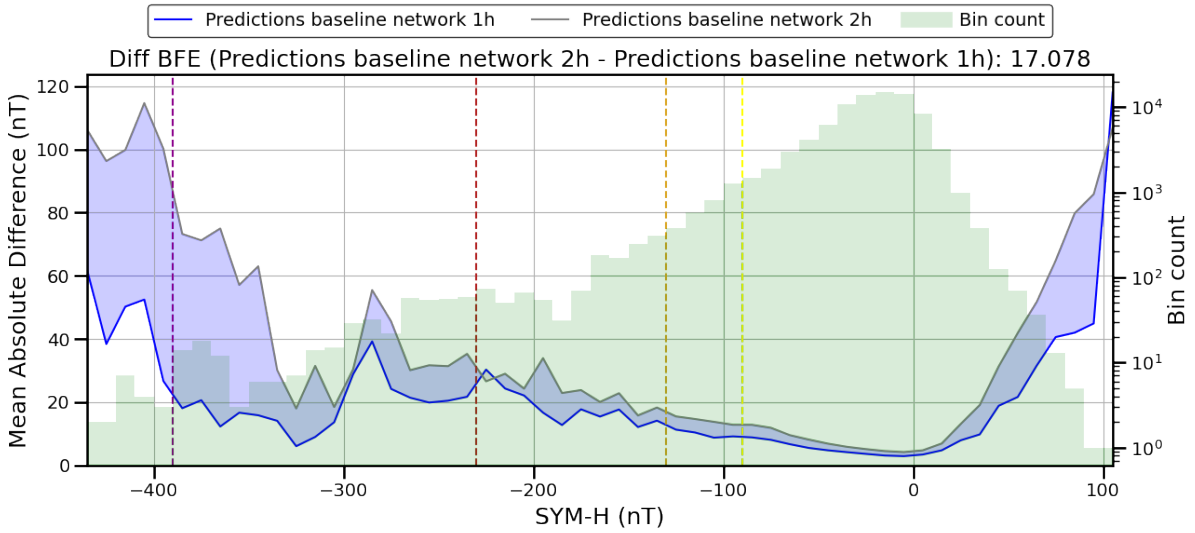


Figure 4.23: Comparison using the BFE for the baseline model for the SYM-H index for the time horizons of 1 and 2 hours on the test storms.

Then, comparing the metrics for the two hours forecast, Tables 4.14 and 4.15 depict the computation of the metrics for the baseline and the persistence models, both in a per-storm manner, the mean of all the storms and the global computation of the metric. Nevertheless, in this case is very important to also plot the BFE. Figure 4.24 shows the computation of the BFE over the test storms. In this case, the situation is pretty similar to the 1 hour situation, the baseline model significantly outperforms the persistence in most of the bins, except on some of the positive values of the SYM-H. However, when comparing the performance on the test key storms, depicted in Figure 4.25, despite having a better overall BFE, the baseline model fails to accurately forecast the two most intense bins. This discrepancy is only caused by one particular storm in which there are significant data gaps, but it is important to note. Moreover, this particular case underscores the importance of plotting the BFE to pinpoint where the model is performing or under-performing.

Table 4.14: Metrics for the 2-hours forecast of the SYM-H index over the test storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
81	14.013	25.47	9.133	13.298	0.794	0.563
82	16.455	20.865	11.555	15.155	0.916	0.856
83	9.815	12.681	10.211	12.958	0.869	0.789
84	17.216	20.651	13.564	16.428	0.858	0.792
85	8.102	10.833	5.895	10.092	0.946	0.84
86	6.269	10.173	7.145	10.693	0.924	0.829
87	15.715	21.545	9.557	12.355	0.834	0.723
88	13.371	11.819	9.736	9.968	0.864	0.858
89	19.062	32.143	11.311	19.606	0.959	0.876
90	14.339	19.67	9.665	14.682	0.904	0.777
91	43.254	59.669	22.318	34.325	0.921	0.814
92	20.322	29.454	12.572	13.031	0.706	0.684
93	8.345	11.262	8.707	11.976	0.86	0.734
94	8.555	11.293	6.619	8.854	0.936	0.886
95	12.395	18.606	9.931	15.207	0.925	0.823
96	14.593	28.948	13.56	15.501	0.651	0.544
97	14.754	24.218	9.955	14.276	0.863	0.718
98	11.539	17.598	9.447	11.804	0.835	0.742
99	16.385	29.558	12.931	16.357	0.77	0.631
100	7.914	14.427	7.073	9.957	0.89	0.781
101	28.314	40.487	18.612	28.788	0.943	0.863
102	12.329	18.256	8.777	11.841	0.775	0.591
103	13.646	15.855	9.322	12.883	0.861	0.734
104	8.567	13.174	6.969	7.778	0.884	0.855

Continued on next page

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
105	8.423	11.513	9.093	9.231	0.732	0.724
106	10.159	17.143	6.88	9.804	0.81	0.615
107	22.263	21.928	8.332	11.899	0.888	0.771
108	21.97	27.546	11.071	14.414	0.867	0.774
109	13.787	13.669	11.022	11.053	0.908	0.907
110	7.2	15.061	7.248	10.722	0.922	0.828
111	9.21	16.456	7.004	13.015	0.929	0.754
112	7.187	7.363	6.03	6.966	0.946	0.928
113	19.303	33.467	9.121	13.118	0.704	0.388
114	20.253	24.649	12.87	17.641	0.924	0.858
115	14.538	26.934	8.91	12.158	0.865	0.749
116	9.297	11.709	9.37	11.998	0.835	0.73
Mean:	14.413	20.725	10.042	13.606	0.862	0.759
Global:	37.895	53.272	10.796	14.967	0.914	0.834

Table 4.15: Metrics for the 2-hours forecast of the SYM-H index over the test key storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
117	16.048	22.326	9.564	13.153	0.933	0.874
118	6.842	9.084	6.416	8.009	0.875	0.806
119	18.533	19.595	11.534	12.998	0.745	0.677
120	5.483	11.372	5.479	8.537	0.938	0.85
121	16.866	31.543	9.051	13.291	0.694	0.34
122	11.754	12.947	7.777	8.934	0.869	0.827
Mean:	12.588	17.811	8.304	10.82	0.842	0.729
Global:	19.367	24.028	8.557	11.128	0.882	0.801

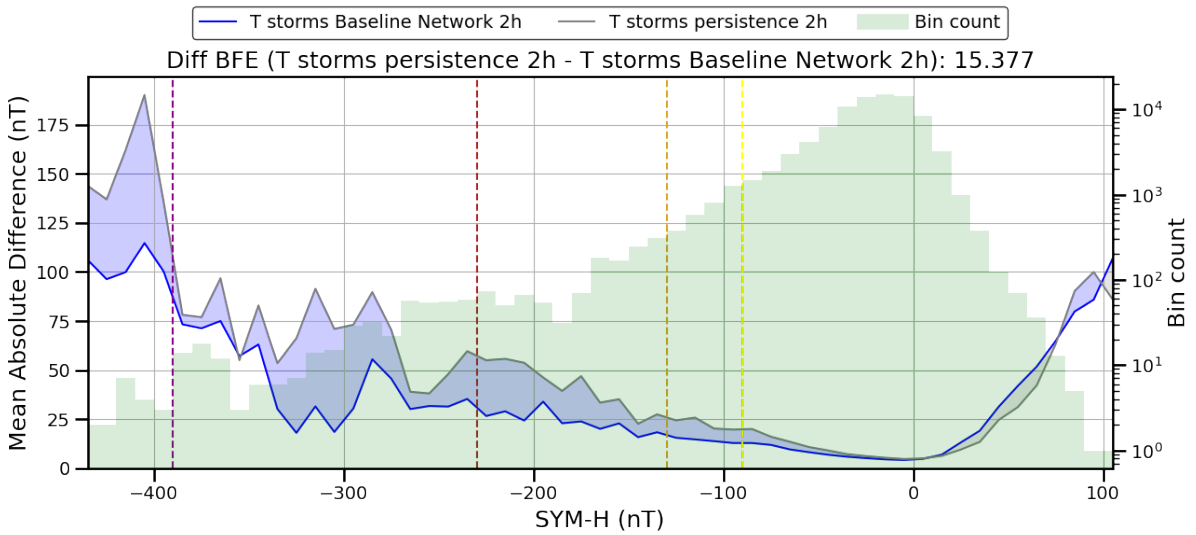


Figure 4.24: Comparison of the BFE on the predictions on the test storms made by the baseline model for the SYM-H index for the 2 hours ahead forecast compared to the persistence model. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

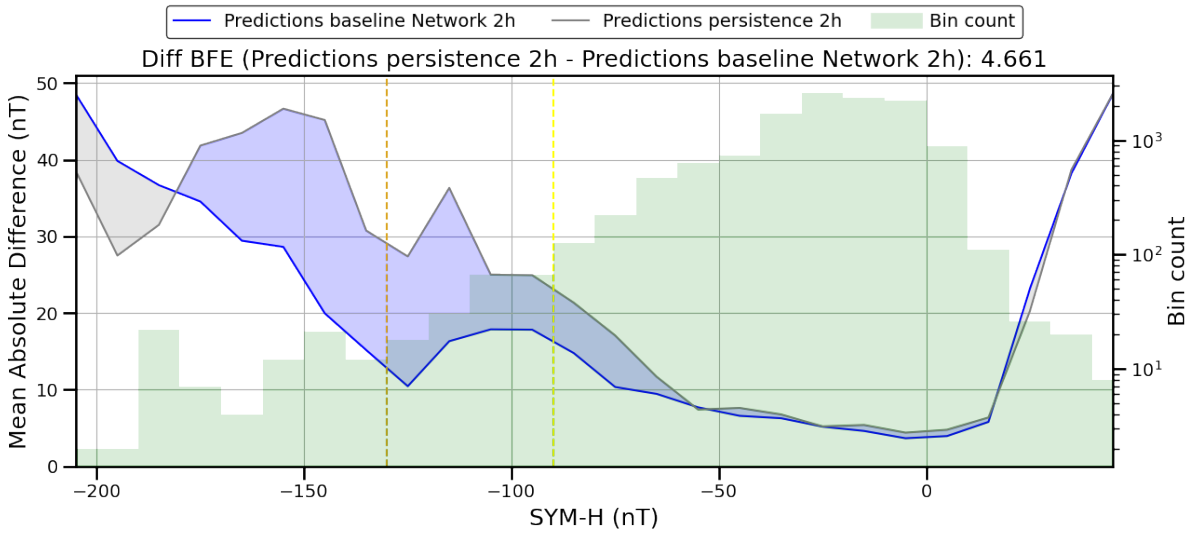


Figure 4.25: Comparison of the BFE on the predictions on the test key storms made by the baseline model for the SYM-H index for the 2 hours ahead forecast compared to the persistence model. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

4.2.4.3 ASY-H

The ASY-H index, given its asymmetrical nature is significantly harder to forecast. Additionally, for the more extreme storms, the amount of examples is extremely scarce, as there is only two superintense storms. Moreover, they are considerably different, as evidenced in the superposed epoch plot in Figure 4.17.

Tables 4.16 and 4.17 present the comparison of the baseline model compared to the persistence one for the one hour ahead forecast for the test and test key storms, respectively. The baseline model generally surpasses the persistence model, especially in terms of average RMSE and R^2 metrics, with an average RMSE approximately 20% better. However, this improvement is mostly observed during inactive periods of the index. Despite its overall superior performance, the baseline model's accuracy decreases significantly when predicting the most intense storms. To illustrate that, Figure 4.27 shows the BFE comparing the baseline model for the ASY-H index to the persistence on two scenarios. The left plot in Figure 4.27 evaluates all test storms, while the right plot specifically excludes the superintense storm, which the baseline model fails to forecast accurately, as it predicts a moderate storm instead of a superintense one. This storm is further showcased in Figure 4.26. The evaluation of the ASY-H thus underscores the crucial impact of the most intense storms on the BFE's overall evaluation, revealing the baseline model's limitations in accurately forecasting the most extreme storm events.

Table 4.16: Metrics for the 1-hour forecast of the ASY-H index over the test storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
93	29.406	28.753	9.595	11.656	0.775	0.668
94	72.492	82.811	19.044	27.012	0.714	0.424
95	23.928	27.943	10.376	13.845	0.793	0.632
96	28.411	17.050	13.966	14.356	0.715	0.699
97	17.016	16.814	10.124	11.968	0.774	0.684
98	95.988	74.928	25.87	26.959	0.692	0.665
99	14.936	17.935	10.142	11.721	0.714	0.619
100	24.725	22.781	11.208	13.554	0.767	0.660
101	31.998	25.370	14.12	15.725	0.722	0.656
102	58.284	70.897	19.493	21.789	0.625	0.532
103	39.127	25.758	16.716	17.738	0.681	0.641
104	24.303	37.779	13.319	20.445	0.489	-0.203
105	28.025	30.084	11.803	15.450	0.516	0.170
106	39.561	46.166	12.625	17.688	0.620	0.254
107	33.572	26.360	16.16	19.742	0.637	0.458
108	24.55	19.122	15.589	17.699	0.697	0.609
109	97.09	116.003	22.02	28.322	0.509	0.188

Continued on next page

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
110	33.661	36.515	10.95	13.514	0.664	0.488
111	27.021	39.742	14.051	16.402	0.759	0.672
112	25.612	30.230	14.026	18.174	0.673	0.450
113	38.915	64.382	13.46	19.106	0.702	0.400
114	17.233	30.290	7.939	10.953	0.668	0.368
115	32.069	38.393	9.655	12.204	0.645	0.433
116	23.998	26.793	9.172	11.184	0.692	0.542
117	18.565	43.974	13.75	17.199	0.599	0.372
118	21.622	26.890	8.329	10.125	0.807	0.715
119	22.243	26.409	10.144	13.087	0.699	0.499
120	31.652	39.296	9.698	12.801	0.784	0.623
121	58.858	69.530	18.978	20.181	0.752	0.720
122	37.728	37.387	16.191	19.479	0.661	0.509
123	47.621	43.571	16.007	21.354	0.568	0.231
124	37.975	32.829	15.087	20.802	0.464	-0.019
125	26.905	24.221	10.449	13.010	0.660	0.474
Mean:	35.912	39.303	13.638	16.826	0.674	0.48
Global:	112.548	92.757	14.4	17.684	0.704	0.553

Table 4.17: Metrics for the 1-hour forecast of the ASY-H index over the test key storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
126	22.971	22.590	8.465	10.738	0.707	0.528
127	38.24	22.141	16.502	14.678	0.722	0.780
129	27.652	39.437	9.287	14.010	0.641	0.183
130	26.757	27.540	11.134	13.449	0.710	0.578
131	23.581	28.192	8.393	10.939	0.681	0.458
132	33.074	36.563	13.452	18.068	0.554	0.195
133	36.802	29.738	11.45	12.999	0.611	0.498
134	28.823	27.320	9.511	9.913	0.680	0.653
135	28.053	39.214	6.969	10.208	0.689	0.332
136	35.647	22.871	15.152	14.395	0.565	0.608
137	36.977	34.547	13.507	16.106	0.701	0.574
138	26.699	32.847	10.024	13.111	0.734	0.545
139	28.625	24.472	12.097	12.461	0.638	0.616
140	26.358	27.991	9.984	12.573	0.705	0.532

Continued on next page

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
141	31.636	22.111	15.762	17.102	0.443	0.344
142	27.662	14.627	14.316	11.123	0.656	0.793
143	25.595	25.126	11.431	11.209	0.711	0.722
Mean:	29.569	27.739	11.551	13.151	0.653	0.516
Global:	43.846	27.775	11.865	13.368	0.68	0.593

While this does not have a strong impact on the overall RMSE when all the storms are taken into account, it has a great impact on the BFE. Despite being only a few samples above the 400 nT mark, those peaks are the most important values to accurately forecast since are the moments when the storm has the potential to cause significant disruptions. This results in the baseline model having a worse BFE than the persistence model, caused by its inability to forecast the more intense scenarios, while certainly being better in the more common, inactive state.

This particular example illustrates the hypothetical case presented in the previous section. The baseline model has a better average RMSE and R^2 score, but this improvement is highly condensed in the quiet activity time, which comprises more than 90% of the evaluated data, while not being able to properly forecast the remaining, considerably more important, intense active time.

The analysis of our case study reveals a critical insight: the simple NN models struggle to accurately forecast extreme ASY-H storm events. This finding highlights the importance of using BFE as a key metric for evaluation. Unlike traditional metrics that might overlook a model's performance in rare, high-intensity scenarios, BFE provides a more comprehensive assessment. It specifically highlights the model's capability (or lack of) in predicting these crucial extreme events. Therefore, when evaluating models for geomagnetic storm forecasting, particularly for the asymmetric and unpredictable nature of ASY-H storms, it is essential to consider BFE. By doing so, we can determine whether the evaluated model possesses the expected capabilities to reliably forecast across the full spectrum of geomagnetic conditions, especially during the most intense and challenging scenarios.

Finally, Figure 4.28 depicts the comparison of the BFE for the baseline model and the persistence for the test key storms. In this case the baseline model performs better across all the ASY-H range, having a global BFE score 10 units better.

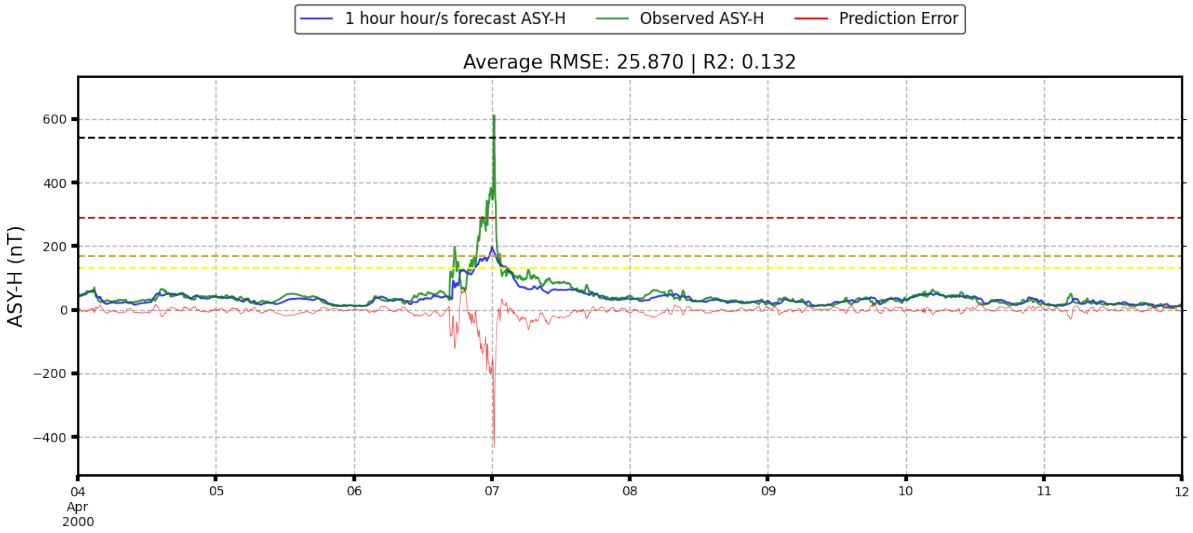


Figure 4.26: Forecast of the superintense storm 97 for the ASY-H index using the baseline model.

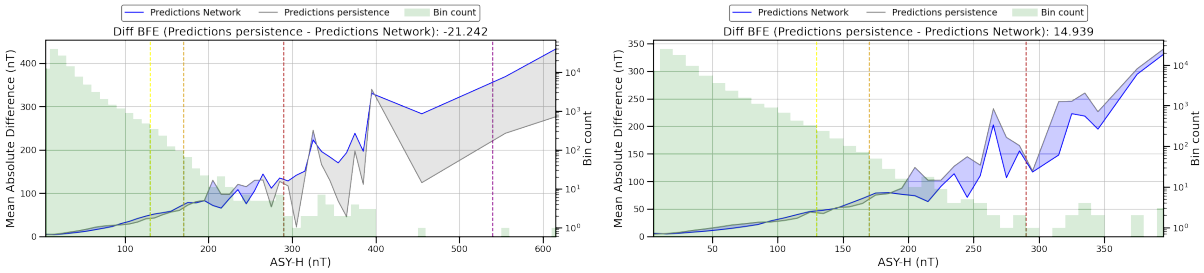


Figure 4.27: Comparison of the BFE on the predictions made by the baseline model and the persistence model for the ASY-H index 1 hour ahead on the all the test storms on the left, and the test storms, except the superintense storm of April, 2000, on the right. The vertical colored lines mark the separation of the different intensities categories as classified in Section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

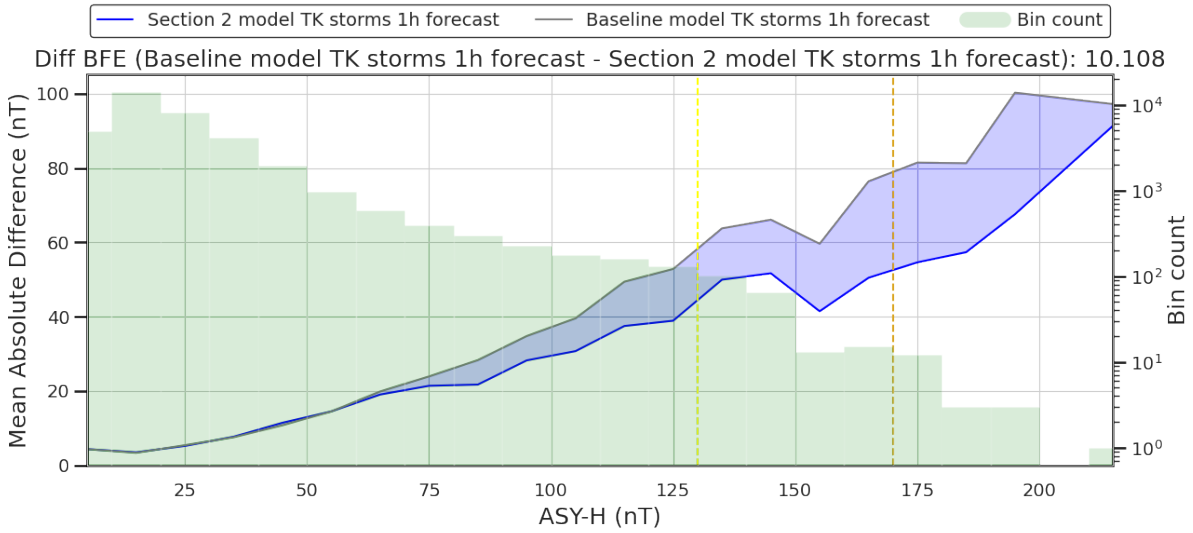


Figure 4.28: Comparison of the BFE on the predictions made by the baseline model and the persistence model for the ASY-H index 1 hour ahead on the test key storms. The vertical colored lines mark the separation of the different intensities categories as classified in Section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

In the 2 hours evaluation, the situation remains largely similar, Tables 4.18 and 4.19 show the computation of the metrics comparing the baseline model and the persistence one on the test and test key storms respectively. Additionally, Figures 4.29 and 4.30 depict the computation of the BFE metric on the test and test key storms. For both the average and global evaluation, the baseline model performs better on the BFE and R^2 metric. However, falls short on the BFE. Once again, the largest geomagnetic storm plays a vital role in the BFE, making the baseline model lose in the global computation of the metric.

Table 4.18: Metrics for the 2-hours forecast of the ASY-H index over the test storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
93	33.413	44.728	11.147	15.961	0.696	0.378
94	83.443	83.321	22.329	28.834	0.606	0.343
95	22.386	36.101	10.542	18.202	0.786	0.363
96	35.574	28.273	17.009	19.801	0.577	0.427
97	18.371	25.247	12.516	16.055	0.654	0.431
98	109.559	96.469	29.804	35.379	0.591	0.424
99	17.344	26.336	11.698	14.823	0.62	0.39
100	24.164	24.985	11.844	17.123	0.74	0.457
101	32.722	30.782	16.128	18.893	0.638	0.503
102	75.711	72.457	21.857	23.245	0.529	0.467
103	39.375	38.702	17.803	22.485	0.638	0.423
104	36.85	44.283	17.334	25.107	0.135	-0.815
105	34.472	37.484	13.231	18.496	0.391	-0.19
106	46.436	54.206	15.025	21.278	0.462	-0.08
107	39.829	38.604	19.549	24.673	0.469	0.154
108	21.387	27.47	16.873	22.055	0.645	0.393
109	112.031	130.47	26.768	33.257	0.274	-0.12
110	35.144	35.251	11.874	15.167	0.605	0.355
111	31.529	57.421	14.929	21.305	0.728	0.447
112	22.775	43.214	15.76	22.171	0.587	0.182
113	53.683	56.454	16.323	22.621	0.562	0.158
114	16.968	43.419	9.433	15.182	0.532	-0.213
115	38.483	49.405	11.44	16.097	0.502	0.013
116	20.712	39.828	9.72	15.619	0.654	0.107
117	47.831	50.259	16.385	21.407	0.43	0.027
118	29.052	33.624	10.384	12.802	0.7	0.544
119	30.547	42.583	13.658	18.65	0.454	-0.018
120	38.21	51.525	11.982	17.193	0.67	0.32

Continued on next page

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
121	97.252	96.506	25.556	25.423	0.551	0.556
122	44.21	45.651	18.816	24.133	0.542	0.247
123	49.784	46.921	17.32	25.35	0.494	-0.084
124	47.437	40.75	17.775	23.71	0.256	-0.324
125	28.69	31.021	11.827	16.61	0.565	0.142
Mean:	42.89	48.598	15.898	20.882	0.554	0.194
Global:	133.604	117.476	16.811	21.728	0.596	0.325

Table 4.19: Metrics for the 2 hours forecast of the ASY-H index over the test key storms, comparing the baseline model and the persistence.

#	BFE		RMSE		R2	
	Baseline	Persistence	Baseline	Persistence	Baseline	Persistence
126	21.363	25.566	8.668	12.828	0.692	0.326
127	38.991	33.952	17.116	18.229	0.701	0.661
128	30.946	32.298	12.301	17.278	0.466	-0.053
129	37.681	46.025	11.335	16.763	0.465	-0.17
130	30.621	31.643	13.075	17.427	0.601	0.291
131	24.973	28.863	9.292	13.117	0.609	0.221
132	41.948	43.024	16.138	22.402	0.358	-0.237
133	42.301	31.320	13.764	15.644	0.438	0.274
134	30.196	30.088	9.998	14.225	0.647	0.285
135	32.545	43.184	8.793	12.031	0.505	0.072
136	42.807	31.599	19.207	19.406	0.301	0.287
137	47.283	48.694	16.824	20.119	0.536	0.336
138	28.251	43.055	11.342	17.838	0.659	0.157
139	31.382	30.965	14.722	17.306	0.464	0.259
140	29.367	34.853	12.064	16.328	0.569	0.21
141	42.285	33.777	20.361	22.608	0.07	-0.146
142	33.032	23.271	17.57	17.695	0.482	0.475
143	33.234	38.606	13.592	17.051	0.591	0.357
Mean:	34.400	35.044	13.676	17.128	0.509	0.2
Global:	50.357	43.812	14.133	17.386	0.546	0.312

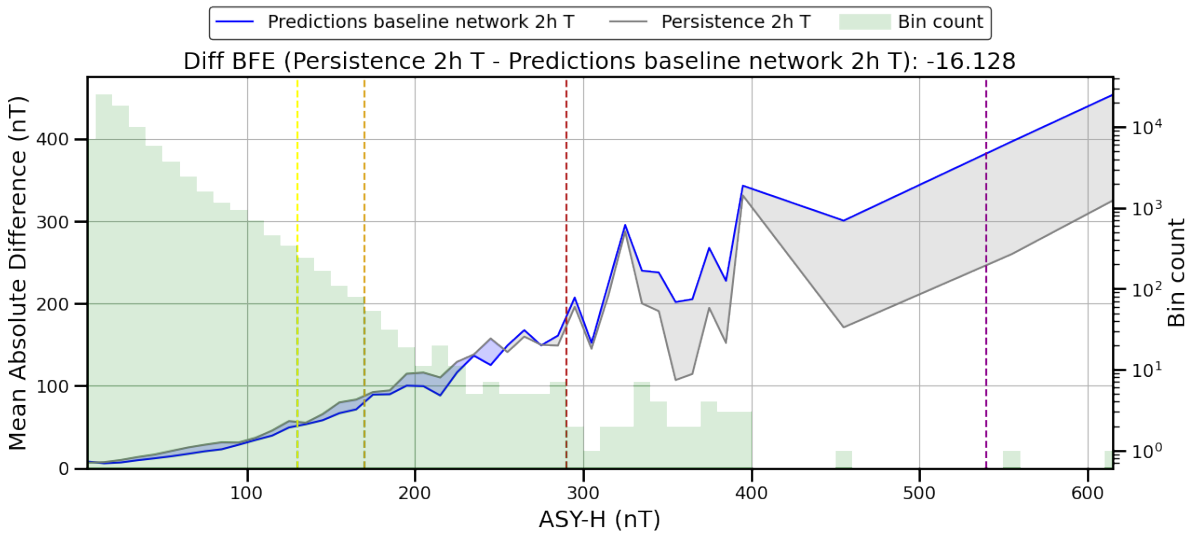


Figure 4.29: Comparison of the BFE on the predictions on the test storms made by the baseline model for the ASY-H index for the 2 hours ahead forecast compared to the persistence model. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

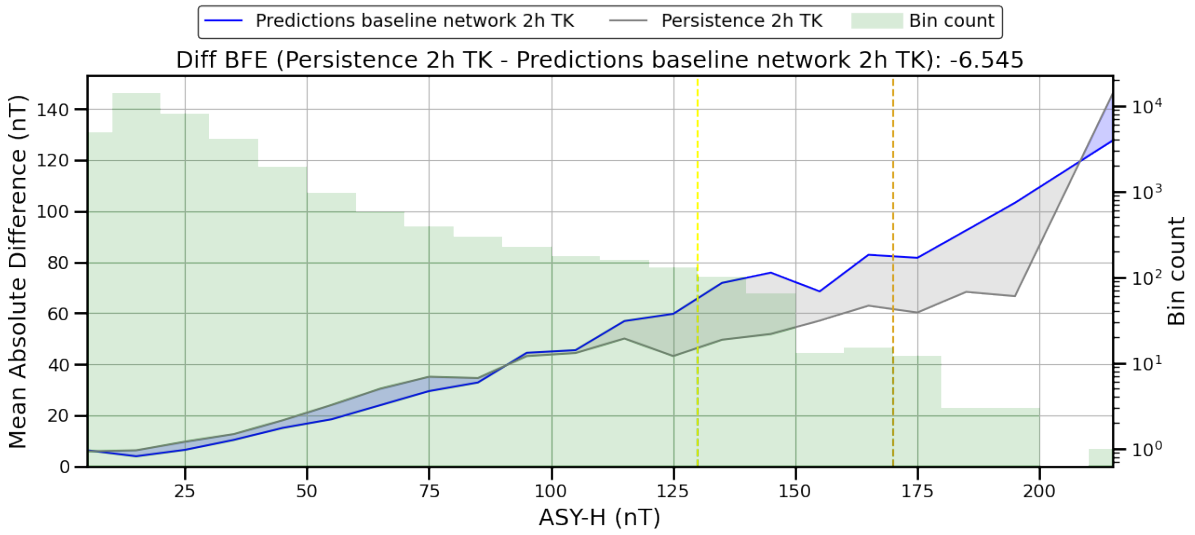


Figure 4.30: Comparison of the BFE on the predictions on the test key storms made by the baseline model for the ASY-H index for the 2 hours ahead forecast compared to the persistence model. The vertical colored lines mark the separation of the different intensities categories as classified in section 4.1. The green shaded histogram depicts the number of samples on each bin, on a logarithmic scale.

4.3 Conclusions

Geomagnetic disturbances have been quantified by the scientific community using different indices, but the use of these indices by the operational community is not straightforward. Historically, geomagnetic activity has been measured using indices such as Kp or Dst; however, their operational utility has been limited due to low time resolution. In contrast, the SYM-H and ASY-H indices, with their high temporal resolution, provide a more accurate representation of geomagnetic disturbances, particularly for low and mid latitude regions. SYM-H, in particular, has proven to be a reliable proxy for the local disturbance with an uncertainty equal to the ASY-H index.

Aside from the selection of the index, a quantification of the severity of the disturbance to provide an estimation of the vulnerability is also of great importance. As such, a significant contribution of this study is the development of a statistically-backed classification system for geomagnetic storms based on geomagnetic indices. This was applied to indices with different resolutions: one-minute SYM-H and ASY-H indices and the hourly Dst. This classification addresses gaps in existing previous thresholds. By resampling the indices in 27-day intervals to address their autocorrelation, we provide a robust classification framework based on industry-wide percentiles. This system categorizes storms into four intensity levels –low, moderate, intense, and superintense– offering a clear and objective method for assessing geomagnetic risks.

The identification and classification of geomagnetic storms eases the development of downstream applications. For instance, the development of forecasting models that need to separate the available storms into training and testing sets while maintaining enough representative samples of all the classes in the sets. Having an intensity classification of geomagnetic storms simplifies the selection of the storms to train and validate forecasting models. With this classification, we can expand the SYM-H sets used in Chapter 3 and create a new set specifically tailored for the ASY-H, taking its particularities into account using the identified storms, instead of reusing the separation made for the SYM-H. This expansion should help the model learn from a wider variety of storms. Moreover, the proposed procedure can be applied to other geomagnetic indices assuming there is sufficient data available to resample the time-series into independent variables.

To improve model evaluation, we also address the limitations of traditional metrics like RMSE and R^2. To ensure that model evaluations accurately reflect performance across a range of geomagnetic conditions, we have designed a new metric named BFE that evaluates the performance of the model based on the real value of the index. As the previous metrics give the same weight to all the samples both inactive and active periods of the storm. Therefore, since there are more inactive samples a good performance forecasting them overshadows less accurate predictions during active times. BFE fills this gap by providing a sharp view of model's performance, especially in the critical prediction of high-intensity events. It enhances our understanding of the model predictive accuracy by highlighting its capabilities and drawbacks across all the possible observed values of the geomagnetic indices. It also facilitates the comparison of different models, without overvaluing the inactive time, being resilient to different start and end dates for the storms.

In conjunction with the new metric, the revised storm sets for SYM-H index takes into account storms that were not previously considered, enhancing all the sets. For the ASY-H index, the proposed sets serve as a starting point for future predictive models.

Moreover, assessing models against preliminary parameters storms provides insights into how these models are likely to perform in a real-time operational context, for both the SYM-H and ASY-H indices.

The BFE proves to be an invaluable tool, both for comparing different models on similar storms and for providing a comprehensive overview of a model's performance when it is evaluated across all test storms. However, it is crucial to acknowledge the metric's inherent skew towards the most intense storms. These storms are usually the most relevant to forecast accurately due to their potential for dealing significant damage. However, they largely affect the metric computation, having the most effect in negative performance evaluations. Nonetheless, for studies specifically focused on moderate or intense storms, excluding the superintense category when computing the BFE offers a detailed evaluation. This approach complements traditional metrics, providing a more refined assessment of a model's capabilities in forecasting storms of specific intensities. Such a targeted evaluation strategy is essential for developing more precise and reliable forecasting models.

Chapter 5

Improving the network

Don't compare your life to others. There's no comparison between the Sun & the Moon, they shine when it's their time.

This chapter focuses on enhancing the forecasting capabilities of the DNN model presented in Chapter 3 for the SYM-H and ASY-H indices. The DNN model is retrained using the extended subsets for SYM-H and ASY-H defined in Chapter 4, incorporating the newly identified storms. By leveraging these expanded datasets, the retrained models are evaluated using both traditional regression metrics and the BFE metric introduced earlier. The results demonstrate a significant improvement in forecasting accuracy, particularly for the most intense geomagnetic disturbances, highlighting the efficacy of this approach. These updated models outperform those presented in Chapter 3, underscoring the value of the extended datasets.

In addition to point forecasts, we also introduce quantile-based confidence intervals, offering a probabilistic range within which the observed SYM-H values are expected to fall. This advancement significantly improves forecast reliability and usability, providing decision-makers with a more comprehensive understanding of forecast uncertainty.

The system is evaluated using both historical level 2 science-ready data and preliminary observations, which simulate the operational environment where the model is intended to function. The ability of the model to perform well in real-time scenarios demonstrates its robustness and practical utility for SW forecasting. By integrating quantile forecasts into SYM-H prediction models, we provide more accurate and trustworthy information to manage the hazards posed by geomagnetic storms, representing a substantial improvement in real-time forecasting capabilities.

Additionally, a preliminary version of the SYM-H forecasting model, trained with the datasets proposed in Section 4 and including the prediction intervals, was deployed for real-time operation and is accessible at https://www.senmes.es/pub/ISG/lastSYMforUAHinterval.png.

5.1 Retraining the model with the extended datasets

This section focuses on updating the DNN model using the extended subsets for the SYM-H and ASY-H indices defined in Section 4.2. The performance of these retrained models is evaluated using both the traditional metrics and the newly introduced BFE metric. The results demonstrate a notable improvement in the models' performance, particularly in predicting the most intense values of the SYM-H. This underscores the effectiveness of the approach in enhancing geomagnetic indices forecasting.The retrained models outerperform those presented in section 3.4 for the SYM-H index and in section 3.5 for the ASY-H index, highlighting the value of the extended datasets.

Tho achieve this, the DNN model described in Chapter 3 is retrained using the extended subsets for the SYM-H and ASY-H indices defined in Section 4.2.1. The rationale behind this is to evaluate the impact of the extended data set on the SYM-H index forecasting, particularly on the BFE metric. Regarding the ASY-H, we can not directly compare with the Section 3.5 model, the advanced model using Attention. This is because of the rework of the sets. As such, we compare it with Section 4.2.4.3 model, the baseline one using LSTM and Dense layers.

5.1.1 Data preparation and processing

Following the methodology outlined in Section 3.1, we maintain the same data preparation and processing strategies. The key difference lies in the selection of storm subsets for training, validation, and testing, ensuring a comprehensive representation of geomagnetic disturbances. We also employ the secondary test set using only the preliminary key parameters to give a closer look to the performance that the model will have in the operational scenario.

5.1.2 SYM-H index evaluation

The SYM-H models can be directly compared because the database revision ensured that neither model was trained or validated on the new subset's test storms. The retrained model demonstrates an improvement in the forecast of the SYM-H index. Although the architectural framework remains unchanged, the extended training data set has refined the model's performance. The BFE improved from 21.170 to 19.273, with significant gains in predicting extreme storm intensities. However, it is important to take into account that the hyper-parameter optimization was done for the original model trained with the old sets, not for the new ones.

The comprehensive comparison of the network performance on old and new data splits is shown in Table 5.1. It reveals the impact of the expanded dataset. The improvement on the RMSE, R2, and BFE metrics across the evaluated storms underscore the effectiveness of the retrained network. The new split data generally shows improved performance, as evidenced by lower RMSE values and higher R2 scores in most cases, indicating enhanced predictive accuracy and model reliability. The reduction in BFE for the new split further confirms the model's refined capability in forecasting, especially in handling extreme conditions. These results collectively highlight the significance of dataset selection

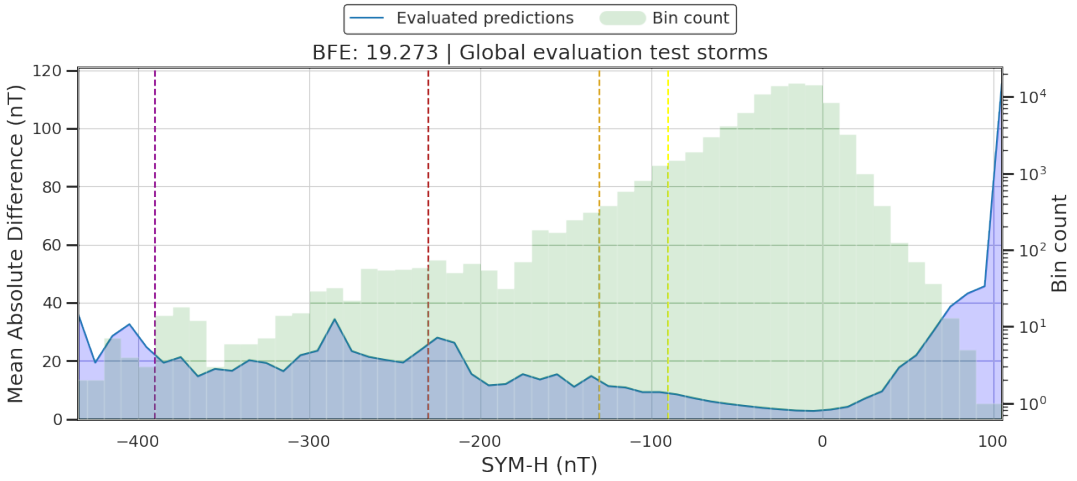


Figure 5.1: Graph of the BFE evaluated on the predictions for the SYM-H index in the next hour for the test storms made by the model presented in Chapter 3 and trained with the SYM-H sets presented in Section 4.2.

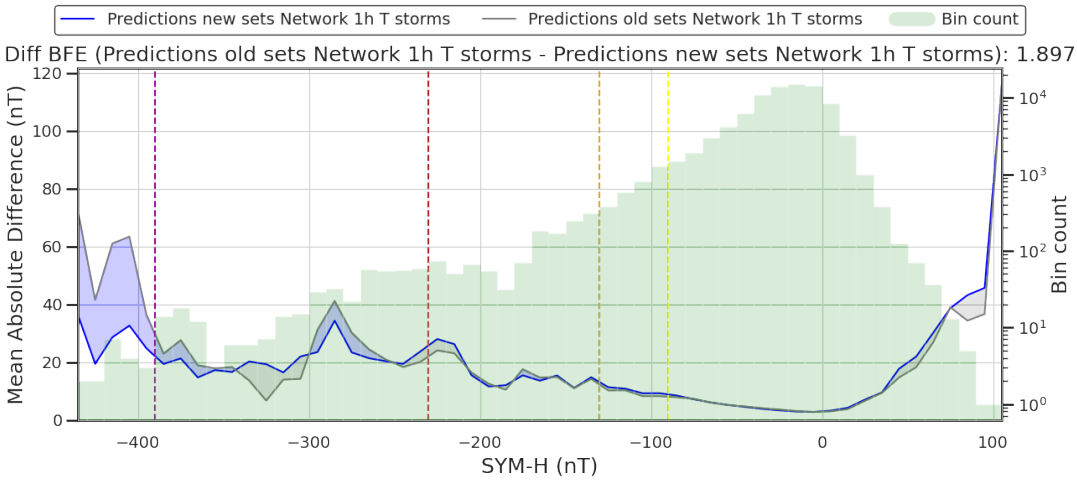


Figure 5.2: Comparison of the BFE evaluated on the predictions of the SYM-H index in the next hour on the test storms made by the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2.

and the potential benefits of retraining with carefully curated data in enhancing model performance for geomagnetic storm forecasting.

Additionally, Figure 5.1 depicts the global computation of the BFE metric over the test storms. As we can see, the retrained model performs remarkably well, not increasing the error on the most intense bins of the SYM-H. The only area where the model has not improved is on the very positive values of the SYM-H index. However, this is not that surprising, because even with the expanded dataset, the number of samples for those bins has not increased, as they are very scarce. Moreover, Figure 5.2 depicts the comparison of the BFE obtained by both models, showing a remarkable improvement in the superintense area.

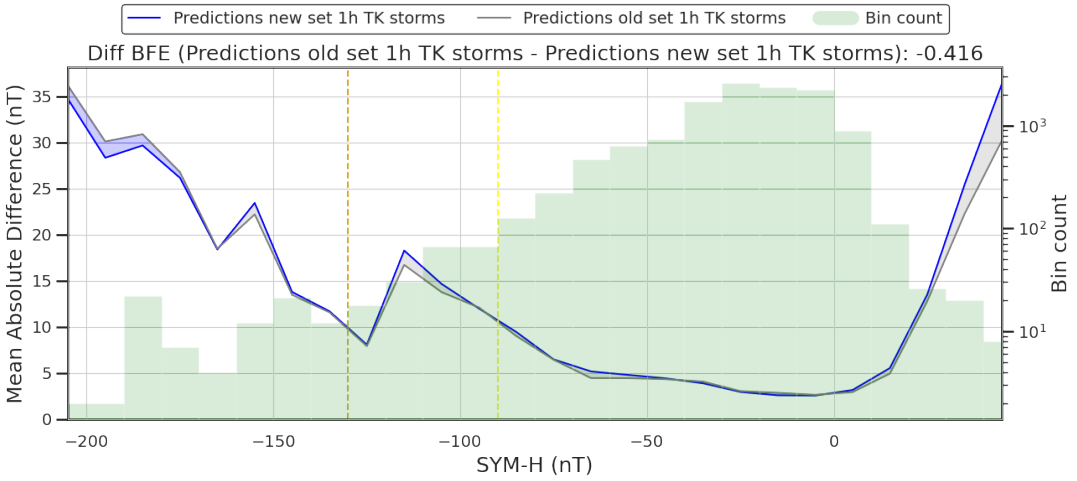


Figure 5.3: Comparison of the BFE evaluated on the predictions of the SYM-H index in the next hour on the test key storms made by the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2.

Table 5.1: Comparison of the performance of the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2 for the SYM-H forecast 1 hour ahead on the test storms.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
81	7.888	8.625	5.385	5.494	0.928	0.925
82	15.111	14.245	8.784	8.603	0.951	0.953
83	6.233	6.721	6.290	6.565	0.950	0.946
84	9.589	9.658	8.176	8.244	0.948	0.948
85	5.887	5.634	4.826	4.217	0.964	0.972
86	4.018	3.634	4.628	4.442	0.968	0.970
87	7.327	8.065	5.404	5.069	0.947	0.953
88	4.118	5.346	4.916	4.903	0.965	0.966
89	12.243	13.064	7.478	7.795	0.982	0.980
90	8.796	8.729	6.223	6.124	0.960	0.961
91	28.141	22.106	14.350	12.485	0.968	0.975
92	13.031	15.523	7.353	7.565	0.900	0.894
93	6.035	6.489	5.713	5.459	0.940	0.945
94	6.391	7.178	5.078	5.180	0.963	0.961
95	7.728	7.745	6.337	6.271	0.969	0.970
96	12.324	13.060	10.054	9.972	0.808	0.811
97	9.988	10.934	7.119	6.885	0.930	0.934
98	7.697	7.723	6.571	6.338	0.920	0.926
99	10.298	11.603	8.173	8.294	0.908	0.905

Continued on next page

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
100	5.411	6.022	4.787	4.803	0.949	0.949
101	13.482	15.380	10.820	11.494	0.981	0.978
102	9.520	9.048	7.021	6.391	0.856	0.881
103	9.539	9.720	6.612	6.156	0.930	0.939
104	6.005	7.193	5.006	4.330	0.940	0.955
105	4.758	4.006	5.514	4.677	0.902	0.929
106	4.966	5.009	4.276	4.103	0.927	0.933
107	15.183	17.124	6.380	6.672	0.934	0.928
108	11.493	12.503	5.265	6.032	0.970	0.960
109	8.719	8.586	6.414	6.822	0.969	0.965
110	7.450	7.158	5.314	5.188	0.958	0.960
111	6.529	6.432	4.829	4.725	0.966	0.968
112	5.221	5.483	4.550	4.557	0.969	0.969
113	10.323	11.097	5.656	5.689	0.886	0.885
114	14.344	13.713	8.831	8.310	0.964	0.968
115	14.928	14.791	6.883	6.809	0.920	0.921
116	5.990	6.141	6.219	5.863	0.927	0.936
Mean	9.353	9.597	6.590	6.459	0.939	0.942
Global:	21.170	19.273	7.035	6.895	0.963	0.965

Table 5.2 illustrates the network's performance on test key storms with only preliminary parameters available for the 1 hour ahead forecast. While the network trained with the new split exhibits a marginally higher BFE, it performs significantly better when forecasting the intense values. This trend is particularly noteworthy, as the model's slight underperformance is primarily observed when the SYM-H index reaches very positive values. These nuances in model behavior can be observed in Figure 5.3, which visually represents these findings, providing a clearer perspective of the model's strengths and limitations across different storm intensities. This particular case underscores the importance of the human evaluation of the metric.

Table 5.2: Comparison of the performance of the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2 for the SYM-H forecast 1 hour ahead on the test key storms, using the preliminary parameters.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
117	11.351	11.505	6.716	6.647	0.967	0.968
118	4.675	4.639	4.441	4.176	0.940	0.947
119	12.823	12.589	8.091	8.000	0.875	0.877
120	3.694	4.233	4.041	4.130	0.966	0.965

Continued on next page

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
121	11.192	12.607	5.843	6.178	0.872	0.857
122	8.291	7.827	5.137	4.867	0.943	0.949
Mean	8.671	8.900	5.711	5.666	0.927	0.927
Global:	13.632	14.049	5.875	5.847	0.944	0.945

On the 2 hours ahead comparison, the results are mostly similar, confirming the initial hypothesis that the increased training subset has a positive impact on the network evaluation. Tables 5.3 and 5.4 compare the metrics for the two splits on the test and test key storms respectively. For the test storms there is an important improvement across all the metrics, most notably on the BFE, as the new model forecasts better the most intense values of the SYM-H, as shown in Figure 5.4. Regarding the evaluation of the test key storms, the new split presents better RMSE and R^2 metrics but slightly worse BFE. This discrepancy is mainly due to the lower performance on the storm 117, the most intense among the test key storms. This causes the global computation of the BFE to be slightly worse, as shown in Figure 5.5. Nevertheless, the difference is extremely minor.

Table 5.3: Comparison of the performance of the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2 for the SYM-H forecast 2 hours ahead on the test storms.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
81	12.434	13.311	8.517	8.748	0.821	0.811
82	15.790	16.840	11.642	11.462	0.915	0.917
83	8.637	8.739	9.939	9.040	0.876	0.897
84	18.428	17.796	12.604	12.058	0.877	0.888
85	6.786	6.413	5.353	5.072	0.955	0.960
86	5.741	5.829	7.224	6.577	0.922	0.935
87	12.971	12.860	8.554	8.007	0.867	0.884
88	8.987	7.785	7.940	6.861	0.910	0.933
89	16.857	18.162	9.773	10.241	0.969	0.966
90	13.678	13.610	9.300	9.142	0.911	0.914
91	38.404	39.799	19.120	20.056	0.942	0.937
92	18.848	20.979	10.689	11.036	0.788	0.774
93	8.430	8.498	9.145	8.076	0.845	0.879
94	8.879	9.334	6.773	6.854	0.933	0.932
95	12.444	13.631	9.790	10.338	0.927	0.918
96	15.924	15.740	14.263	13.800	0.614	0.639
97	13.578	15.811	9.827	10.012	0.866	0.861
98	12.077	11.754	9.282	8.738	0.841	0.859

Continued on next page

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
99	13.829	13.902	11.394	11.877	0.821	0.806
100	7.386	8.739	6.215	6.383	0.915	0.910
101	24.577	21.445	18.001	17.700	0.946	0.948
102	12.557	12.586	9.073	8.791	0.760	0.775
103	12.993	12.966	9.196	8.854	0.864	0.874
104	10.990	10.616	8.956	5.819	0.808	0.919
105	8.103	6.301	9.076	7.332	0.733	0.826
106	9.833	10.555	7.006	6.681	0.803	0.821
107	20.619	21.927	7.797	8.300	0.902	0.888
108	19.024	19.205	9.077	9.343	0.910	0.905
109	10.834	10.888	7.524	7.880	0.957	0.953
110	9.096	10.387	6.953	7.234	0.928	0.922
111	7.743	9.174	6.412	7.076	0.940	0.927
112	7.124	7.109	5.831	5.736	0.950	0.951
113	18.390	13.856	8.373	7.810	0.751	0.783
114	19.617	20.721	12.486	12.201	0.929	0.932
115	15.906	15.473	8.434	8.376	0.879	0.881
116	9.515	9.013	9.442	8.659	0.833	0.859
Mean:	13.528	13.660	9.472	9.227	0.872	0.883
Global:	34.271	31.720	10.161	9.995	0.924	0.926

Table 5.4: Comparison of the performance of the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2 for the SYM-H forecast 2 hours ahead on the test key storms, using the preliminary parameters.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
117	17.547	17.915	10.003	10.163	0.927	0.925
118	7.594	6.685	6.717	6.186	0.863	0.884
119	19.231	18.989	11.159	11.370	0.762	0.752
120	5.117	6.239	5.402	5.593	0.940	0.935
121	13.617	13.551	8.521	8.227	0.729	0.747
122	14.329	13.965	8.579	8.385	0.840	0.847
Mean:	12.906	12.891	8.397	8.321	0.844	0.849
Global:	20.901	21.556	8.611	8.557	0.881	0.882

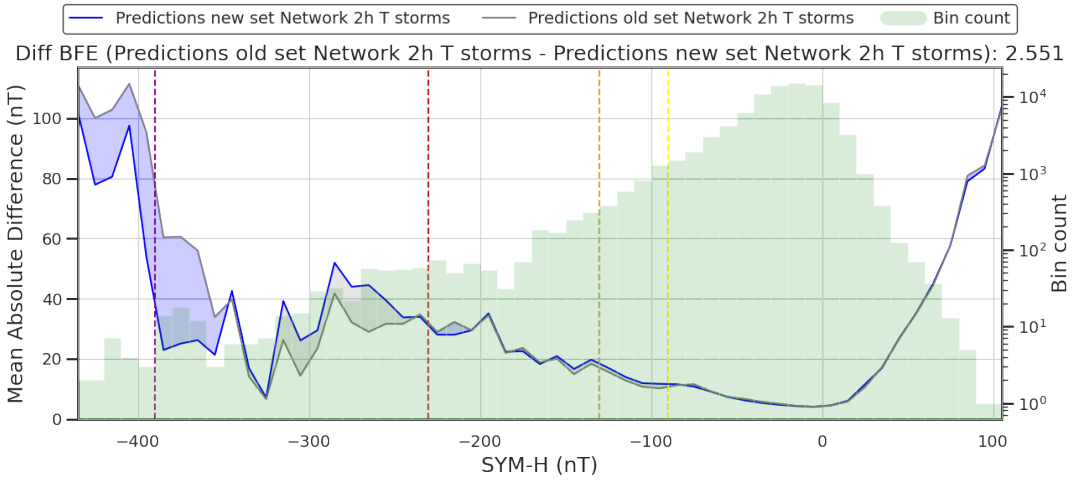


Figure 5.4: Comparison of the BFE evaluated on the predictions of the SYM-H index for the 2 hours ahead forecast on the test storms made by the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2.

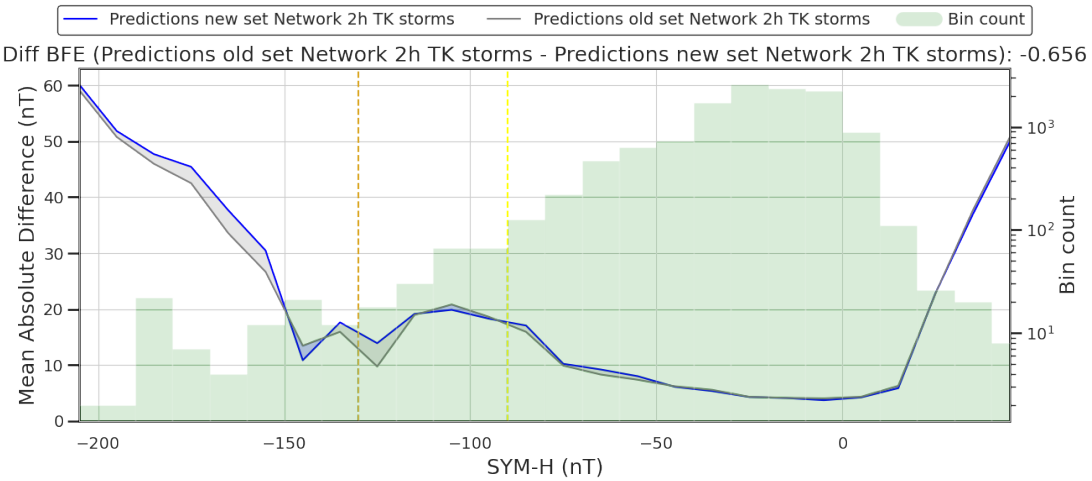


Figure 5.5: Comparison of the BFE evaluated on the predictions of the SYM-H index for the 2 hours ahead forecast on the test key storms made by the model presented in Section 3.4 trained with the original subsets and the same model trained with the subsets proposed in Section 4.2.

5.1.3 ASY-H index evaluation

When evaluating the ASY-H index, a direct comparison with the model presented in Section 3.5 is not feasible. This is because, in the existing literature, a previous training, validation and test sets separation for the ASY-H did not exist. A such, we created a new one as described in Section 4.2. In this proposed separation, some of the storms that are used for testing the model in the proposed separation, were used to train the model in Section 3.5. Considering that, a direct comparison of the models is not entirely fair.

Nevertheless, the model can be compared to the baseline model presented in Section 4.2.4.3. Tables 5.5 and 5.6 depict the computation of the BFE, RMSE and R^2 metrics made by the baseline model and the Chapter 3 model trained with the ASY-H sets for the test storms and the test key storms respectively. As expected, the model

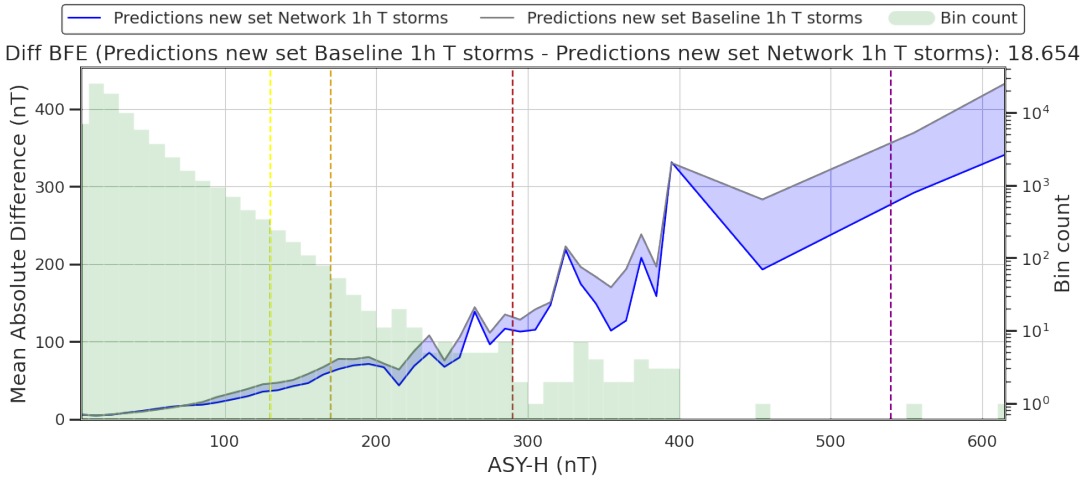


Figure 5.6: Comparison of the BFE evaluated on the predictions of the ASY-H index for the 1 hour ahead forecast on the test storms made by the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2.

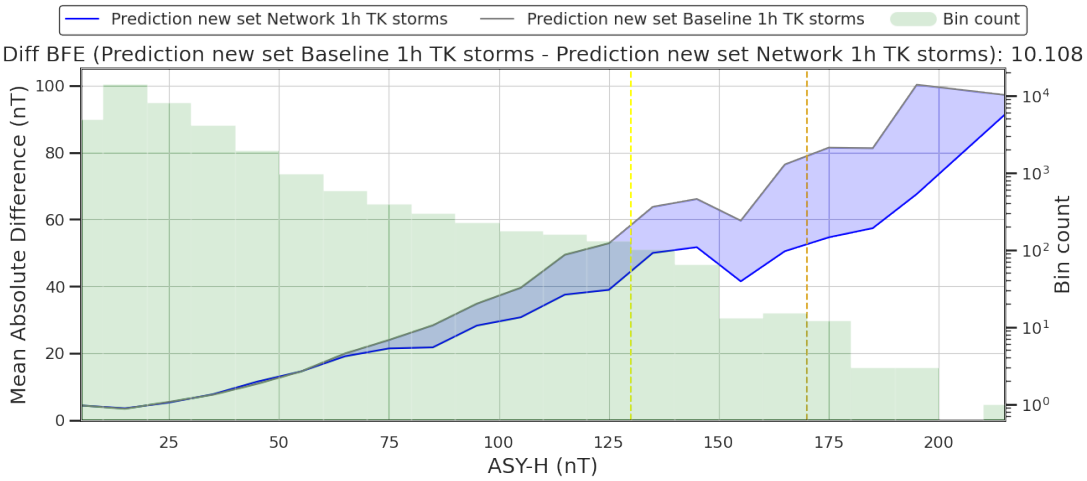


Figure 5.7: Comparison of the BFE evaluated on the predictions of the ASY-H index for the 1 hour ahead forecast on the test key storms made by the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2.

presented in Chapter 3 significantly outperforms the baseline model in all the evaluated metrics.

Figures 5.6 and 5.7 depict the comparison of the BFE for both models on the test storms and the test key storms respectively. Despite the Chapter 3 model still having a considerable error in the most intense values of the ASY-H, it has been reduced substantially.

Table 5.5: Comparison of the performance of the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2 for the 1 hour ahead forecast on the test storms.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
93	29.406	25.444	9.595	9.243	0.775	0.791
94	72.492	64.254	19.044	18.553	0.714	0.728
95	23.928	19.053	10.376	9.223	0.793	0.837
96	28.411	22.004	13.966	12.172	0.715	0.784
97	17.016	12.937	10.124	9.478	0.774	0.802
98	95.988	73.432	25.87	21.694	0.692	0.783
99	14.936	11.771	10.142	9.541	0.714	0.747
100	24.725	17.775	11.208	8.986	0.767	0.850
101	31.998	24.805	14.12	13.077	0.722	0.762
102	58.284	59.591	19.493	19.735	0.625	0.616
103	39.127	29.255	16.716	14.525	0.681	0.759
104	24.303	28.801	13.319	13.735	0.489	0.457
105	28.025	28.149	11.803	11.787	0.516	0.517
106	39.561	39.305	12.625	12.635	0.620	0.619
107	33.572	27.689	16.16	15.369	0.637	0.672
108	24.55	17.705	15.589	13.446	0.697	0.774
109	97.09	97.886	22.02	23.797	0.509	0.427
110	33.661	32.026	10.95	10.838	0.664	0.671
111	27.021	24.173	14.051	14.034	0.759	0.760
112	25.612	21.258	14.026	13.695	0.673	0.688
113	38.915	32.271	13.46	13.326	0.702	0.708
114	17.233	15.583	7.939	7.508	0.668	0.703
115	32.069	29.270	9.655	9.052	0.645	0.688
116	23.998	17.940	9.172	7.773	0.692	0.779
117	18.565	23.541	13.75	11.936	0.599	0.698
118	21.622	16.970	8.329	7.457	0.807	0.845
119	22.243	21.861	10.144	10.145	0.699	0.699
120	31.652	21.231	9.698	7.670	0.784	0.865
121	58.858	51.800	18.978	15.986	0.752	0.824
122	37.728	31.768	16.191	14.698	0.661	0.721
123	47.621	42.172	16.007	16.078	0.568	0.564
124	37.975	37.979	15.087	15.621	0.464	0.425
125	26.905	25.462	10.449	10.340	0.660	0.667

Continued on next page

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
Mean:	35.912	31.671	13.638	12.823	0.674	0.704
Global:	112.548	93.894	14.4	13.653	0.704	0.734

Table 5.6: Comparison of the performance of the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2 for the 1 hour ahead forecast on the test key storms, using preliminary parameters.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
126	22.971	21.708	8.465	8.509	0.707	0.704
127	38.24	26.910	16.502	13.241	0.722	0.821
128	27.093	26.931	10.48	10.557	0.612	0.607
129	27.652	32.028	9.287	9.797	0.641	0.600
130	26.757	22.295	11.134	9.966	0.710	0.768
131	23.581	25.270	8.393	8.826	0.681	0.647
132	33.074	35.024	13.452	13.726	0.554	0.535
133	36.802	33.500	11.45	10.972	0.611	0.643
134	28.823	20.744	9.511	7.820	0.680	0.784
135	28.053	29.110	6.969	7.362	0.689	0.653
136	35.647	26.329	15.152	12.858	0.565	0.687
137	36.977	27.763	13.507	10.639	0.701	0.814
138	26.699	23.505	10.024	10.353	0.734	0.716
139	28.625	24.100	12.097	11.344	0.638	0.682
140	26.358	24.638	9.984	9.694	0.705	0.721
141	31.636	25.787	15.762	14.217	0.443	0.547
142	27.662	20.057	14.316	11.325	0.656	0.785
143	25.595	20.441	11.431	9.731	0.711	0.790
Mean:	29.569	25.897	11.551	10.608	0.653	0.695
Global:	43.846	33.738	11.865	10.774	0.68	0.736

For the 2 hours ahead forecast, the situation is mostly similar. Tables 5.7 and 5.8 present the comparison of the metrics for the baseline model and the Chapter 3 model, both trained on the ASY-H sets presented in Section 4.2 for the test and test key storms respectively. Then, Figures 5.8 and 5.9 depict the comparison of the BFE metric. As evidenced in both images, the model presented in Chapter 3 outperforms the baseline model in all the evaluated range for the index. With the most notable improvementas observed in the most intense values of the index. Nevertheless, the model still struggles to accurately forecast the superintense storm, leaving room for further improvement for those cases.

Table 5.7: Comparison of the performance of the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2 for the 2 hours ahead forecast on the test storms.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
93	33.413	32.560	11.147	11.17	0.696	0.695
94	83.443	73.64	22.329	20.416	0.606	0.671
95	22.386	22.41	10.542	10.685	0.786	0.781
96	35.574	26.843	17.009	14.97	0.577	0.673
97	18.371	17.561	12.516	12.121	0.654	0.676
98	109.559	95.183	29.804	27.381	0.591	0.655
99	17.344	17.972	11.698	12.212	0.62	0.586
100	24.164	21.086	11.844	11.393	0.74	0.76
101	32.722	27.793	16.128	15.489	0.638	0.666
102	75.711	71.252	21.857	20.655	0.529	0.579
103	39.375	32.771	17.803	16.244	0.638	0.699
104	36.85	37.548	17.334	17.387	0.135	0.13
105	34.472	35.003	13.231	13.094	0.391	0.404
106	46.436	48.444	15.025	15.392	0.462	0.435
107	39.829	35.542	19.549	18.894	0.469	0.504
108	21.387	21.826	16.873	16.741	0.645	0.65
109	112.031	117.428	26.768	29.631	0.274	0.111
110	35.144	31.993	11.874	11.419	0.605	0.634
111	31.529	34.493	14.929	17.503	0.728	0.626
112	22.775	22.03	15.76	14.549	0.587	0.648
113	53.683	53.211	16.323	17.545	0.562	0.494
114	16.968	22.513	9.433	9.576	0.532	0.517
115	38.483	35.835	11.44	11.149	0.502	0.527
116	20.712	19.614	9.72	9.464	0.654	0.672
117	47.831	45.299	16.385	15.358	0.43	0.499
118	29.052	22.718	10.384	9.325	0.7	0.758
119	30.547	23.321	13.658	12.65	0.454	0.532
120	38.21	36.208	11.982	12.712	0.67	0.628
121	97.252	96.51	25.556	24.068	0.551	0.602
122	44.21	41.019	18.816	18.241	0.542	0.57
123	49.784	45.838	17.32	18.519	0.494	0.422
124	47.437	45.589	17.775	18.124	0.256	0.227
125	28.69	24.546	11.827	11.292	0.565	0.603

Continued on next page

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
Mean:	42.89	40.473	15.898	15.617	0.554	0.565
Global:	133.604	122.963	16.811	16.677	0.596	0.603

Table 5.8: Comparison of the performance of the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2 for the 2 hours ahead forecast on the test key storms, using preliminary parameters.

#	BFE		RMSE		R2	
	Old Split	New Split	Old Split	New Split	Old Split	New Split
126	21.363	22.011	8.668	8.986	0.692	0.669
127	38.991	35.376	17.116	15.963	0.701	0.74
128	30.946	30.34	12.301	12.195	0.466	0.475
129	37.681	38.533	11.335	11.291	0.465	0.469
130	30.621	25.497	13.075	11.801	0.601	0.675
131	24.973	27.035	9.292	9.618	0.609	0.581
132	41.948	42.327	16.138	16.689	0.358	0.313
133	42.301	39.254	13.764	13.126	0.438	0.489
134	30.196	20.729	9.998	9.001	0.647	0.714
135	32.545	24.698	8.793	8.684	0.505	0.517
136	42.807	35.61	19.207	17.367	0.301	0.429
137	47.283	37.634	16.824	15.439	0.536	0.609
138	28.251	25.143	11.342	11.816	0.659	0.63
139	31.382	26.846	14.722	13.899	0.464	0.522
140	29.367	25.343	12.064	11.107	0.569	0.634
141	42.285	36.153	20.361	18.773	0.07	0.21
142	33.032	27.383	17.57	15.4	0.482	0.602
143	33.234	26.418	13.592	12.072	0.591	0.678
Mean:	34.400	30.352	13.676	12.957	0.509	0.553
Global:	50.357	42.96	14.133	13.315	0.546	0.597

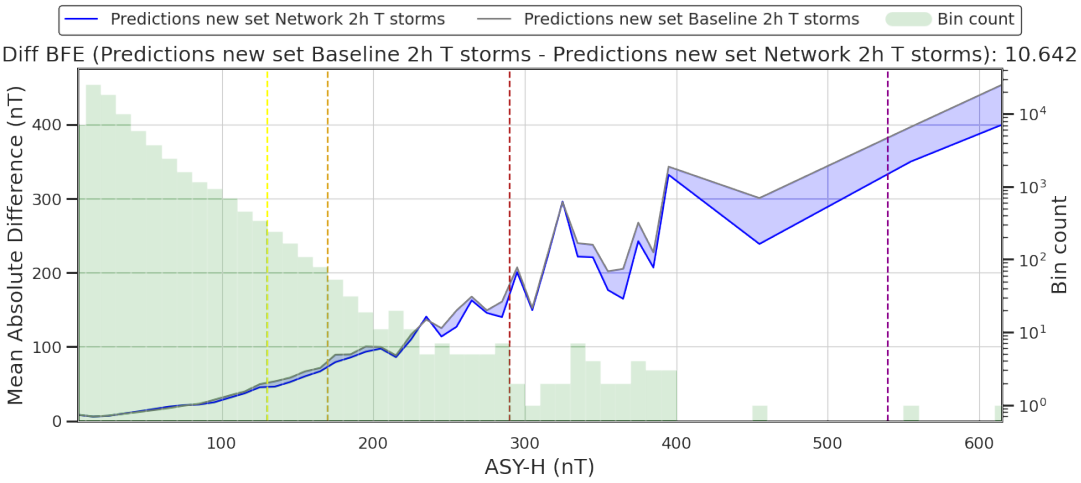


Figure 5.8: Comparison of the BFE evaluated on the predictions of the ASY-H index for the 2 hours ahead forecast on the test storms made by the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2.

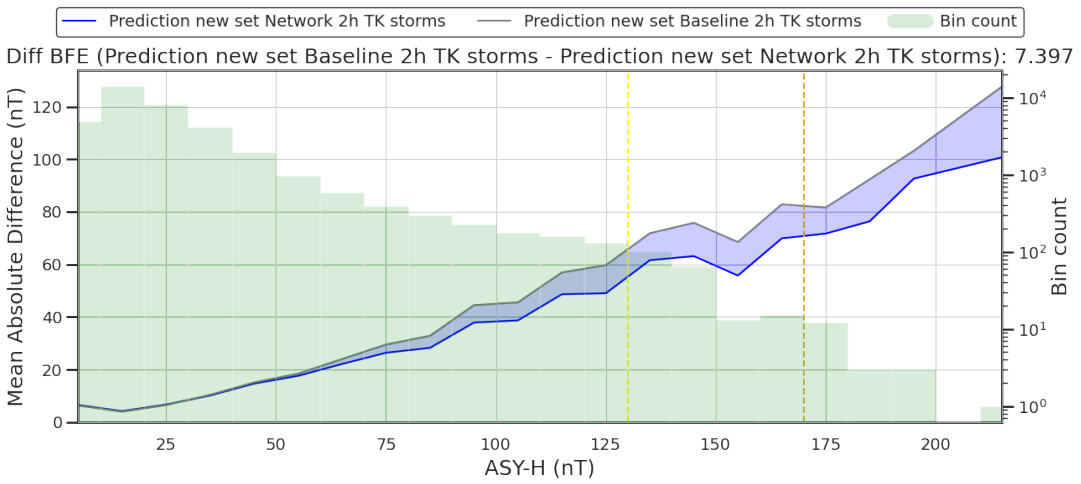


Figure 5.9: Comparison of the BFE evaluated on the predictions of the ASY-H index for the 2 hours ahead forecast on the test key storms made by the baseline model presented in Section 4.2.4.3 and the Chapter 3 model retrained with the new subset for the ASY-H proposed in Section 4.2.

5.2 Forecasting confidence intervals

Decision-makers, such as those managing power grids or satellite operations, could rely on the forecasts made by the DNN model to implement timely and appropriate measures to mitigate the impacts of geomagnetic storms. However, previous works only provide the forecasted value without providing the associated prediction error. Consequently, end users do not have the necessary tools to effectively utilize these outputs in their decision-making processes.

In this regard, providing a confidence interval for the SYM-H forecast is crucial for easing decision-making, because it quantifies the uncertainty associated with the predictions. A confidence interval offers a range within which the true value of the SYM-H index is expected to lie, giving a clearer picture of the forecast's reliability. This additional layer of information allows end users to assess the risk more accurately, while enabling more informed decisions, balancing caution with operational efficiency. By understanding the potential variability in the forecast, they can better prepare for worst-case scenarios and allocate resources more effectively, ultimately enhancing the resilience of critical technological systems against geomagnetic disturbances.

Hence, to make the real-time forecast relevant for end users, another contribution of this work is to define and evaluate the implementation of a confidence interval for geomagnetic indices, operationalized in the 1 and 2 hours real-time SYM-H forecasting with a DNN model. The model used in this work is inherited from the one presented in Chapter 3, re-trained with a larger training dataset as reported in Section 5.1 and enhanced with a confidence interval output as described later.

One of the first approaches found in the literature to provide confidence intervals in forecasting systems is the quantile regression, introduced by Koenker et al. [156]. It was employed in econometrics to estimate conditional quantile functions, enabling a more comprehensive analysis beyond mean predictions. Since then, it has been particularly influential in economics and finance, where understanding the distribution tails, such as the risk of rare events, is essential. It enhances prediction models with robustness against outliers and provides a richer understanding of data distributions.

Quantile forecasting and confidence interval analysis have become essential tools in the time series forecasting domain, especially in fields that require reliable uncertainty estimates. In the context of geomagnetic indices forecasting, such as the SYM-H, the addition of quantile forecasts provides significant benefits, as they enable a more comprehensive understanding of possible values in the forecast, not just the average or most likely outcome. They are particularly beneficial when the distribution is skewed, such as the forecasting during intense geomagnetic storms, as they enhance the forecast reliability and interpretability in a high-stakes, real-time environment [157], [158].

In the field of ML, quantile forecasts have been widely adopted for its ability to create Prediction Intervals (PIs) along with mean predictions. For example Meinshausen [159] introduced Quantile Regression Forests as an extension to Random Forests to estimate conditional quantiles without assuming a parametric distribution of residuals. They have also been integrated with NN [160], [161] to create probabilistic forecasts.

Quantile forecast have been applied in various fields, such as the energy sector, especially with renewable energies like wind and solar. This technique is essential due to the

inherent variability and the necessity of knowing the full distribution of potential outcomes [162]. It has also been applied to urban water demand forecasting [163], financial forecasts [164], weather forecast [165] among others [53].

Additionally, there are specific metrics to evaluate the effectiveness of the prediction intervals, such as the Prediction Interval Coverage Probability (PICP) or the Prediction Interval Average Width (PIAW) [166], [167].

The next subsection details the database used to train the model, followed by the methodology used to forecast confidence intervals and its specific metrics. Then we present how we have adapted the model presented in Chapter 3 to include the confidence intervals, along with the evaluation of the model. Is important to highlight that we are only presenting results for the SYM-H index as is more relevant than the ASY-H in the literature.

5.2.1 Database

We use the storms proposed in Section 4.2.1.1. While using a new dataset can complicate the comparison with previous works [37], [38], as the evaluation is based on comparing performance metrics among the same set of test storms, the expanded set remains compatible with past models. The original training, validation and test storms are preserved in the new dataset, with the addition of 25 new storms that were previously not considered with adjusted start and end dates. This ensures that the new subset does not train on any data used for validation or testing in previous works, and vice versa, maintaining the integrity of comparative analysis. However, for the comparison to be fair, older models would need to be rerun on the adjusted dates.

Moreover, we use a secondary test set, named "test key storms" for which we use the preliminary Solar Wind and IMF measurements. These measurements are not definitive, as they have not undergone the heavy post-processing that is used on the level 2 science-ready data used for the other storms, but they are closer to the data that the model will end up using during operational deployment. This secondary test set allows us to assess the model's performance under more realistic conditions, ensuring its robustness and reliability in an operational environment. The main drawback is that the preliminary measurements are only available from October 2017 onward, so we can only test a limited amount of storms in such conditions.

The input data to forecast the SYM-H are the same as the original DNN presented in Section 3.1, using 5-minute averages of each feature. Meanwhile, the storms set are the ones defined in Section 4.2.1.1.

5.2.2 Quantile forecast

Quantile regression forecasts provide a distributional prediction by estimating specific quantiles (e.g., the 5th, 25th or 90th percentiles), which can be interpreted as providing a range within which the actual value is expected to fall with a certain probability. This method can be particularly useful in scenarios where understanding the range of possible outcomes is more important than just a single point estimation.

To construct a prediction interval using quantile forecasts, we create a band using the desired quantiles. For instance, to create a prediction interval of 90% certainty we use the 5% and 95% quantiles. This means that 90% of the predicted values are expected to fall within this range. This approach allows us to effectively estimate the uncertainty of the predictions. However, it is important to note that greater coverage, such as 95% or 97.5%, results in wider interval widths. While wider intervals may increase the likelihood that the true value falls within the interval, they also limit the usefulness of the prediction by reducing the precision.

Furthermore, the adaptability of quantile forecasts to user-specific needs enhances their utility. Different users may have varying thresholds for what they consider critical levels of geomagnetic activity, depending on their sensitivity to SW impacts. For instance, satellite operators might be concerned with higher quantiles that indicate extreme conditions posing a risk to satellite integrity, while power grid operators may focus on a wider range of quantiles to monitor potential fluctuations in geomagnetic activity that could affect grid stability. By customizing alert levels based on user-defined quantiles, the forecasting system can provide tailored warnings, ensuring that stakeholders receive relevant and actionable information.

Quantile forecasts are deterministic once the model is trained, meaning that given the same input, the model will always produce the same quantiles. This characteristic makes quantile forecasts reproducible and stable. The Quantile Loss has been widely used in the literature [53], [168]–[170], defined as in Equation 5.1, where y is the real value, $\hat{y}$ is the predicted value, and τ is the specified quantile, representing the probability threshold we are interested in forecasting. In quantile loss, for values above the prediction, the loss is scaled by τ, emphasizing overestimations for higher quantiles. Conversely, for underestimations, the loss is scaled by (1-τ), placing greater importance on underestimations for lower quantiles. This asymmetry allows the model to adjust predictions based on the desired confidence level, providing a range of outcomes with varying probabilities, which can be more informative than a single-point forecast.

$$\text{QuantileLoss}_{\tau}(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}) & \text{if } y \geq \hat{y}, \\ (1 - \tau)(\hat{y} - y) & \text{if } y < \hat{y}. \end{cases} \tag{5.1}$$

Existing SW forecasting agencies like NOAA's SWPC and ESA's Space Weather Service Network provide not only numerical forecasts but also probabilities associated with various geomagnetic conditions. However, precise numerical forecasts may not always capture the full range of possible outcomes, potentially impacting user confidence if the actual conditions differ slightly from the forecasted values. Introducing quantile forecasts that offer a range, within which we are 90% confident the actual value will lie, can enhance user trust by transparently communicating uncertainty. Tailoring alerts to specific user needs, such as differentiating between the likelihoods of reaching certain activity levels at the 90% versus 75% quantiles, allows users to better assess risks and make informed decisions based on their unique requirements.

This approach is preferred over post-training methods that estimate the uncertainty solely based on the predicted SYM-H value. Our integrated approach allows the prediction intervals to vary with the surrounding conditions. For example, two forecasts with the same predicted SYM-H value might have different intervals if the input conditions differ.

This ensures that the intervals are more representative of the actual uncertainties in the forecast, enhancing reliability and usefulness.

By incorporating interval estimation directly into the training process, our model simultaneously learns to predict the SYM-H value and its associated uncertainty. This method aligns with recent advancements in uncertainty quantification in DL, which emphasize the importance of training models to understand and represent uncertainty as part of their predictive capabilities [171], [172].

5.2.2.1 Interval coverage metrics

In forecasting, assessing the accuracy and reliability of prediction intervals is crucial. These intervals represent the range within which future observations are expected to fall, providing a measure of uncertainty in the forecasts. A fundamental aspect of evaluating these intervals is understanding how well they capture the possible range of values. This is typically done by examining the interval coverage, which is the percentage of actual observations that fall within the prediciton interval. Another critical metric is the interval width, which indicates the size of the interval. Both metrics are essential as they help in determining the effectiveness of the forecast in covering the range of possible outcomes while maintaining a balance between breadth and precision [173]–[175].

The Prediction Interval Coverage Probability (PICP) measures the proportion of times the observed values fall within the PI. Ideally, this percentage should match the intended confidence level associated with the trained intervals. Variations from this expected coverage are caused by the noise in the data, the impact of uncertainty and along with the imperfect nature of the NN. It is defined as in Equation 5.2, where N is the total number of observations and 1 is an indicator function that is 1 if y_t is within the interval $[L_t, U_t]$ and 0 otherwise.

$$\text{PICP} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}(y_t \in [L_t, U_t]) \tag{5.2}$$

Nevertheless, solely evaluating the PICP is insufficient. While a high PICP can easily be obtained by setting the upper and lower bounds as the extreme values of the target; in that case even 100% can be achieved. However, extremely wide intervals offer almost no practical insight about the target. Therefore, it is crucial to assess the width of the intervals along with the PICP to ensure that the predictions provide meaningful and actionable information. The most common metric is the PIAW; it is a measure of the average width of the prediction intervals. It is calculated as in Equation 5.3, where U_t and L_t are the upper and lower bounds of the prediction interval for each observation t.

$$\text{PIAW} = \frac{1}{N}\sum_{t=1}^{N}(U_t - L_t) \tag{5.3}$$

Other variants also normalize the average width, such as the Prediction Interval Normalized Average Width (PINAW), typically obtained by dividing the width by range of the observed values. However, for the particular case of forecasting the geomagnetic indices during a storm, simply calculating the average width of all the prediction intervals

does not provide enough information. As the interval width during the storm peak is considerably wider than the interval during inactive time (see Figure 5.14 for reference).

To address this, we have chosen to follow behind the BFE to measure the model's forecasting performance across the different possible values of the geomagnetic indices, defined in Equation 5.4, where y_i is the observed value, $\hat{y}_i$ is the predicted value, AD is the Absolute Difference, b are bins of 10 nT, structured in tens with a left closed interval, N_b is the number of observations in bin b, MAD is the Mean of the Absolute Differences and B is the total number of bins.

$$\begin{aligned} \text{AD} &= |y_i - \hat{y}_i| \\ \text{MAD}_b &= \frac{1}{N_b} \sum_{i \in b} \text{AD} \\ \text{BFE} &= \frac{1}{B} \sum_{b=1}^{B} \text{MAD}_b \end{aligned} \tag{5.4}$$

Then, we have applied the binning strategy to the calculation of the interval widths, calculating the Prediction Interval Binned Width (PIBW) as defined in Equation 5.5, where B is the number of bins, n_b is the number of observations in bin b, and U_{b_i} and L_{b_i} are the upper and lower bounds of the interval for the i^{th} observation in bin b.

$$\begin{aligned} \text{Average Width}_b &= \frac{1}{n_b} \sum_{i=1}^{n_b} (U_{b_i} - L_{b_i}) \\ \text{PIBW} &= \frac{1}{B} \sum_{b=1}^{B} \text{Average Width}_b \end{aligned} \tag{5.5}$$

5.2.2.2 Deep neural network architecture

To both forecast the SYM-H index while also providing Prediction Intervals tailored to the chosen quantiles, that capture the underlying uncertainty in the forecasts, we have extended the DNN architecture presented in Chapter 3. Specifically, we have added two copies of the last two Dense Layers, each one tailored to forecast a specific quantile, 5% and 95%, of the SYM-H forecast distribution. The full architecture is depicted in Fig 5.10. The added layers are inside the yellow dotted box. Using those predictions we can create a prediction interval with a 90% confidence. We have chosen a 90% confidence level for our prediction intervals as it provides a good trade-off between interval width and coverage. This level of confidence ensures that a significant majority of the actual values fall within the predicted range, offering reliable forecasts while maintaining reasonably broad intervals.

Considering this addition, the output for the network will now be 3 SYM-H values. These outputs are all simultaneously optimized during the training process. More specifically, the two quantiles are optimized using the Quantile Loss function described previously, with the specific τ for their respective quantile, meanwhile, the original dense layer is still optimized using the MSE, aimed at predicting the "actual" SYM-H value. This

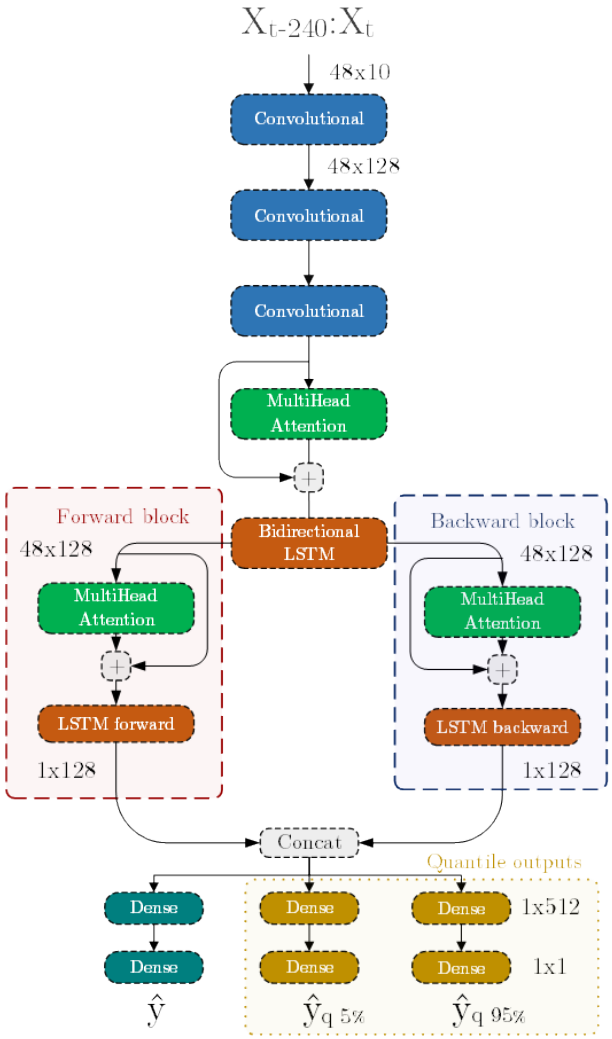


Figure 5.10: DNN architecture for the SYM-H forecast with the quantile forecasts. The input shape is expressed in t ± minutes. The shapes between layers are expressed in time-steps × features. $\hat{y}_{q\ 5\%}$ and $\hat{y}_{q\ 95\%}$ represent the 5 and 95 quantiles forecasts, respectively.

combined approach in loss functions allows the network to not only forecast a reliable central value but also to provide a meaningful assessment of the uncertainty around this forecast by specifying the range within which the actual values are likely to fall, with the desired confidence.

5.2.2.3 Training and validation

Following the framework presented in 4.2, we use the proposed standardized set of geomagnetic storms for training and validation of the expanded network. This set allows for a backwards comparison with previous models that were trained using Siciliano et al. [36] set, such as Collado-Villaverde et al. [37]; Iong et al. [38], since none of the storms used for training in both sets are used in validation or testing.

Considering the extra outputs, we have used a weighted sum of the given loss functions during the training process, in order to balance the influence of each output on the overall model performance. Specifically, each quantile loss (Equation 5.1) is given a weight of 1/2

of the weight assigned to the actual prediction loss. Consequently, the combined weight of the two quantile losses equals the weight of the actual prediction loss, ensuring that each aspect of the model's output receives appropriate attention during training, The Model loss is defined in Equation 5.6 where:

- $L_{\text{quantile},0.05}$ and $L_{\text{quantile},0.95}$ are the quantile loss terms for the 5% and 95% quantiles, respectively.
- $L_{\text{prediction}}$ is the MSE loss for the actual prediction.
- α and β are weights applied to balance the influence of each loss component during training and validation, where α is set to half of β to ensure that the combined weight of the two quantile losses equals the weight of the prediction loss.

$$L_{\text{model}} = \alpha \cdot L_{\text{quantile},0.05} + \alpha \cdot L_{\text{quantile},0.95} + \beta \cdot L_{\text{prediction}} \quad \text{with } \alpha = 0.5 \text{ and } \beta = 1 \tag{5.6}$$

The validation phase uses this same weighted loss approach, it is used to monitor the model's performance on the validation storms, when the validation performance no longer improves (the model loss no longer decreases), the training is stopped. This approach to training and validation is more complex than traditional single-output NN training, but the network learns to not only predict the central tendency of the SYM-H but to also provide the prediction confidence intervals, which is essential for comprehensive SW forecasting. It is important to note that while PICP, PIAW and PIBW metrics are critical for evaluating the performance of the model, they are not directly incorporated into the training process. These metrics serve as evaluation tools to assess the goodness of the results, once the model has been trained, rather than being components of the ML architecture itself.

5.2.3 Model evaluation

The model evaluation is critical in assessing the performance of different forecasting models for the SYM-H index, particularly focusing on how well these models perform during test storms. In this section we compute the performance metrics of the updated model including the quantile forecasts trained on the subset proposed in Section 4.2.1.1. Following the framework, we use the BFE, RMSE, and the R^2 as primary metrics, with additional evaluation using PICP and PIAW as metrics to evaluate the confidence intervals produced by the new model. We do not compare against previous models in order to establish a new baseline that follows the new test set and metrics, which provide more relevant information for a reliable assessment.

It is important to note that, in order to obtain an evaluation of the performance that is as realistic as possible, we have chosen to handle the missing values in the dataset similarly to how they would be handled in a real-time environment. That is, for each forecast, previous values are interpolated only if the gap has already ended, otherwise we propagate the last valid value. Moreover, the test storms contain corrected data that is

commonly not available for real-time forecasts. Therefore, it is expected that, when using the model to forecast in real-time, the accuracy is reduced. For such reasons we also provide the evaluation of the model in near real-time conditions, using a set of storms that only contains preliminary (non-corrected) data to assess the accuracy of the model in an operational scenario.

Tables 5.9 and 5.10 present a comprehensive evaluation of the model over the test storms, using the level 2 science-ready historical data, for both 1-hour and 2-hour forecasts, respectively. The results are fairly good. For the 1-hour ahead forecast, the performance is similar to the model without the quantile intervals presented in Section 5.1. However, the 2-hours ahead forecasts shows a substantial improvement. By incorporating quantiles into the model and optimizing the quantile loss during training, the model is forced to learn from a broader range of patterns in the data, enabling it to learn from both central tendencies and the variability around them. This not only enhances the model's ability to provide reliable prediction intervals but can also improve its overall performance in point forecasts. It is also important to note that the confidence intervals have been successfully trained, as they are close to the target 90% coverage in both the 1 and 2 hours ahead forecasts with both the average and binned width being higher in the 2 hours ahead forecast. An example is shown in Figure 5.11, forecasting one of the most intense storms, where the most intense values of the storm are within the confidence interval.

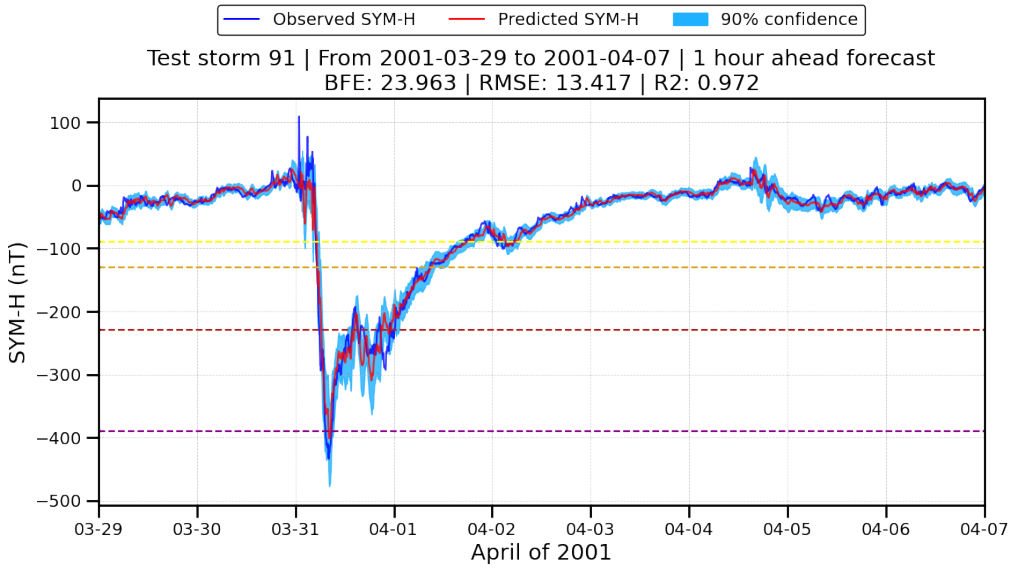


Figure 5.11: 1 hour ahead forecast for the superintense test storm from April 2001.

Figures 5.12 and 5.13 depict the BFE computation on all the test storms for the 1 hour and 2 hour ahead forecasts, respectively. Additionally, the bottom heatmap indicates the percentage of samples that are inside the confidence interval, in green, and outside, in red. The general shape of the 1 hour and 2 hour error is similar, with the 2 hour errors generally higher and with the error increasing on the extreme values, most notably in the positive SYM-H spikes right before the storm. However, they are extremely uncommon. Nevertheless the interval coverage, even during the most intense periods is still quite successful. The visual evaluation of the BFE offers a more rich representation of the model performance compared to the numerical information in the tables. It highlights

Table 5.9: Metrics computation for the test storms on the 1 hour ahead forecast. Dates are in DD/MM/YY-DD/MM/YY format.

#	Storm date	BFE	RMSE	R2	PICP	PIAW	PIBW
81	24/06/98-01/07/98	8.596	5.861	0.915	0.905	17.242	25.666
82	25/08/98-01/09/98	13.119	8.282	0.957	0.837	20.600	30.077
83	17/10/98-24/10/98	6.341	6.657	0.944	0.892	19.617	25.320
84	06/11/98-19/11/98	8.965	8.638	0.942	0.916	20.873	31.655
85	11/01/99-19/01/99	5.213	4.670	0.966	0.964	15.834	23.690
86	27/02/99-06/03/99	3.882	4.858	0.965	0.958	18.518	22.432
87	15/04/99-22/04/99	7.885	5.242	0.950	0.949	18.016	31.731
88	21/01/00-28/01/00	6.875	5.791	0.952	0.954	18.171	27.312
89	04/04/00-12/04/00	12.633	7.965	0.980	0.894	19.996	48.500
90	15/05/00-29/05/00	7.569	6.283	0.959	0.904	18.051	31.821
91	29/03/01-07/04/01	23.963	13.417	0.972	0.848	24.846	62.047
92	15/08/01-22/08/01	16.427	8.023	0.880	0.867	17.111	36.913
93	31/01/02-07/02/02	6.181	5.694	0.940	0.895	17.648	19.252
94	22/03/02-29/03/02	6.372	5.405	0.958	0.906	15.324	20.525
95	02/09/02-13/09/02	6.730	6.078	0.972	0.901	17.929	27.687
96	27/05/03-07/06/03	16.825	10.804	0.779	0.872	23.885	55.100
97	15/06/03-23/06/03	10.876	7.165	0.929	0.883	19.659	30.499
98	10/07/03-21/07/03	7.017	6.789	0.915	0.893	19.594	26.313
99	20/01/04-27/01/04	9.825	8.689	0.896	0.839	21.575	35.771
100	09/02/04-16/02/04	5.906	5.030	0.944	0.934	17.068	22.514
101	05/11/04-17/11/04	16.244	11.314	0.979	0.920	26.486	57.368
102	05/01/05-13/01/05	9.949	6.890	0.862	0.928	19.157	30.592
103	10/06/05-18/06/05	9.470	6.491	0.932	0.889	17.623	26.159
104	21/06/05-28/06/05	7.339	4.698	0.947	0.900	14.305	22.839
105	18/08/06-25/08/06	4.514	5.201	0.912	0.936	15.997	22.669
106	07/03/08-14/03/08	4.505	4.266	0.927	0.964	16.608	26.094
107	03/08/11-11/08/11	15.512	6.532	0.931	0.892	13.420	26.831
108	22/10/11-30/10/11	16.237	7.622	0.937	0.869	10.638	20.465
109	13/07/12-21/07/12	8.829	7.307	0.960	0.875	17.355	24.166
110	29/09/12-18/10/12	7.903	6.051	0.945	0.860	15.280	23.291
111	30/05/13-06/06/13	6.392	4.911	0.965	0.920	15.409	26.307
112	27/06/13-04/07/13	5.190	4.971	0.963	0.911	14.359	20.772
113	10/09/14-17/09/14	8.809	5.529	0.891	0.955	16.745	33.031
114	15/03/15-23/03/15	12.759	9.164	0.962	0.889	21.926	35.867
115	06/06/15-13/06/15	15.208	6.973	0.918	0.890	15.935	22.005
116	06/09/15-16/09/15	5.978	6.194	0.928	0.903	18.296	22.737
	Mean:	9.612	6.818	0.935	0.903	18.086	29.889
	Global:	20.168	7.263	0.961	0.901	18.282	62.113

the increment in difficulty when the lead times increase, especially for the most extreme values of the index, in values lower than -400 nT.

Additionally, Figures 5.14 and 5.15 show the analysis of the confidence intervals, depicting the average interval width for each of the bins for the observed values of the SYM-H. As expected, the interval width rapidly increases with the increased activity level. While the general shape of the width is similar for the 1 and 2 hours forecast, the 2 hours ahead one is considerably wider. Another important point is that most of the values that are outside the prediction interval are concentrated on the right side of the image, corresponding to the most positive SYM-H values reached during the initial phase of the storm.

Table 5.10: Metrics computation for the test storms on the 2 hours ahead forecast. Dates are in DD/MM/YY-DD/MM/YY format.

#	Storm date	BFE	RMSE	R2	PICP	PIAW	PIBW
81	24/06/98-01/07/98	12.967	9.008	0.800	0.914	25.867	39.602
82	25/08/98-01/09/98	15.359	11.182	0.921	0.864	29.355	41.755
83	17/10/98-24/10/98	9.025	9.441	0.888	0.892	28.874	37.199
84	06/11/98-19/11/98	16.407	13.336	0.863	0.907	31.189	43.593
85	11/01/99-19/01/99	7.372	5.564	0.952	0.983	23.378	37.708
86	27/02/99-06/03/99	5.608	6.762	0.932	0.963	27.247	32.281
87	15/04/99-22/04/99	14.124	9.034	0.852	0.913	27.353	44.744
88	21/01/00-28/01/00	11.790	9.090	0.882	0.941	27.692	40.709
89	04/04/00-12/04/00	15.269	9.267	0.972	0.914	28.731	68.317
90	15/05/00-29/05/00	14.796	10.177	0.893	0.920	26.395	48.338
91	29/03/01-07/04/01	34.397	17.372	0.952	0.861	35.912	96.767
92	15/08/01-22/08/01	18.766	11.430	0.757	0.872	25.352	55.039
93	31/01/02-07/02/02	8.219	8.545	0.865	0.882	26.853	28.494
94	22/03/02-29/03/02	8.695	7.156	0.926	0.929	22.879	28.914
95	02/09/02-13/09/02	11.839	9.764	0.927	0.905	26.521	40.058
96	27/05/03-07/06/03	30.368	18.315	0.364	0.871	34.520	86.829
97	15/06/03-23/06/03	15.251	10.325	0.852	0.850	28.376	43.749
98	10/07/03-21/07/03	11.431	9.277	0.841	0.886	27.778	35.947
99	20/01/04-27/01/04	15.667	13.654	0.743	0.829	31.833	56.494
100	09/02/04-16/02/04	8.115	6.770	0.899	0.943	23.844	33.761
101	05/11/04-17/11/04	21.203	17.027	0.952	0.918	38.107	82.204
102	05/01/05-13/01/05	13.181	9.551	0.734	0.952	28.601	48.224
103	10/06/05-18/06/05	13.548	9.573	0.853	0.873	25.877	36.958
104	21/06/05-28/06/05	10.135	6.555	0.897	0.873	20.393	38.118
105	18/08/06-25/08/06	7.416	7.765	0.805	0.941	23.831	34.136
106	07/03/08-14/03/08	10.130	7.089	0.799	0.953	23.100	37.181
107	03/08/11-11/08/11	20.991	8.242	0.890	0.841	18.897	30.548
108	22/10/11-30/10/11	22.750	11.230	0.863	0.839	14.984	29.091
109	13/07/12-21/07/12	12.282	9.702	0.929	0.879	25.059	35.578
110	29/09/12-18/10/12	10.108	7.818	0.909	0.853	21.751	33.239
111	30/05/13-06/06/13	7.989	6.487	0.939	0.929	21.275	38.437
112	27/06/13-04/07/13	7.155	6.348	0.940	0.915	20.467	29.260
113	10/09/14-17/09/14	18.928	10.077	0.639	0.962	26.248	59.864
114	15/03/15-23/03/15	19.322	12.704	0.926	0.901	31.627	49.258
115	06/06/15-13/06/15	14.507	8.278	0.884	0.867	21.935	35.482
116	06/09/15-16/09/15	9.397	9.493	0.831	0.881	27.061	33.181
	Mean:	14.014	9.817	0.857	0.900	26.366	44.196
	Global:	27.408	10.559	0.917	0.898	26.628	91.518

5.2.4 Operational evaluation

To properly complete the evaluation of the model, we also consider the secondary test set described in Section 4.2.1, known as "Test key storms". These storms represent a closer approximation of real-time operational conditions, as they rely on preliminary observations instead of the definitive level 2 science ready data that is used on the historical test storms. This aspect is crucial because, in operational scenarios, forecasters often have to make predictions based on provisional data, despite it being categorized as data only suitable for browsing on the CDAWeb, they are closer to the eventual real-time data that the model will be used on. Therefore, the ability of a model to perform well under these

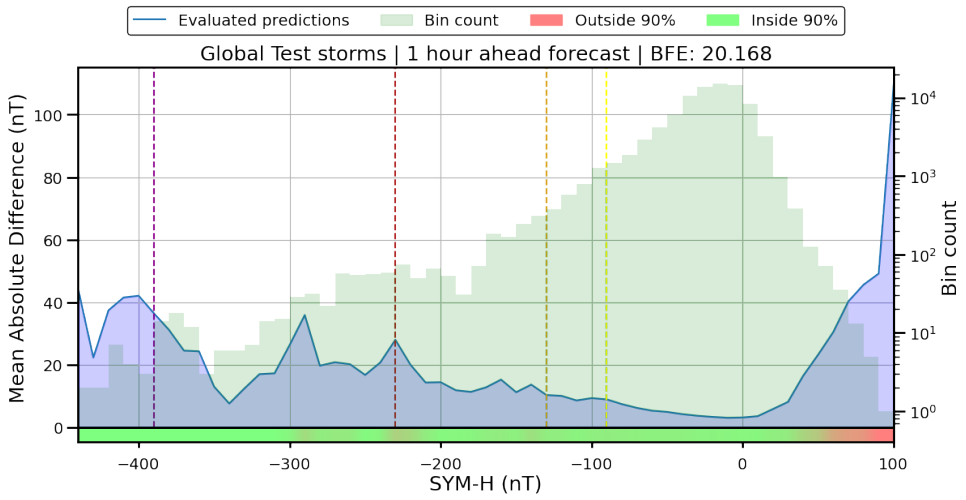


Figure 5.12: BFE plot for the 1 hour ahead forecast on all the test storms. The bottom heatmap shows the percentage of values inside the confidence interval.

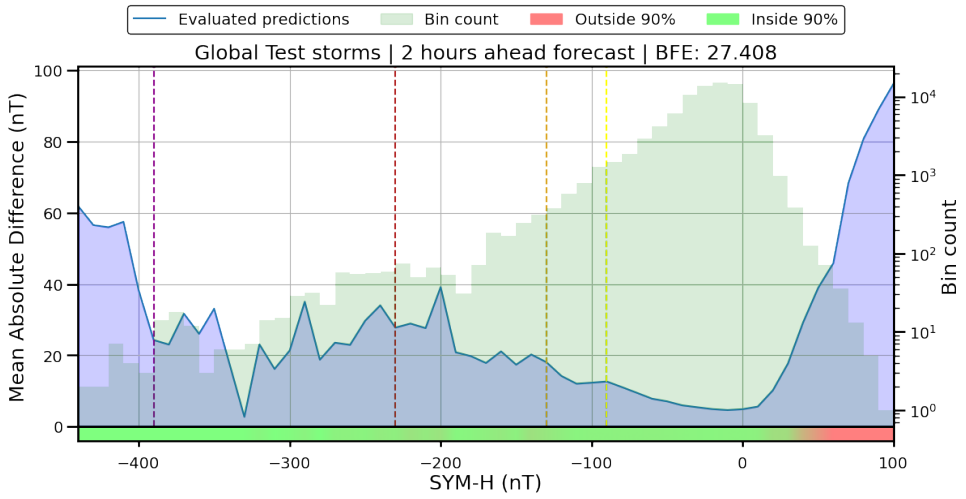


Figure 5.13: BFE plot for the 2 hours ahead forecast on all the test storms. The bottom heatmap shows the percentage of values inside the confidence interval.

conditions is of great importance, arguably more so than its performance on test storms with definitive data.

Test key storms are selected to reflect near-real-time operational scenarios. This selection process ensures that the model's robustness and reliability are tested under the same constraints that would be present in actual forecasting situations. The provisional nature of the data for these key storms introduces additional uncertainty, incorrect or missing values values and variability, making the forecasting task more challenging. This evaluation of the model using test key storms provides a more comprehensive understanding of its real-world applicability. It demonstrates how well the model can handle the uncertainties and data imperfections that are inherent in real-time forecasting.

Tables 5.11 and 5.12 present the performance computation of the models on the test key storms for the 1 and 2 hours ahead forecast, respectively. In both cases, the model presents good point-forecast metrics (BFE, RMSE and R^2). However, the interval effectiveness is lower, as the interval coverage has decreased from the target 90% in the test storms to 86%

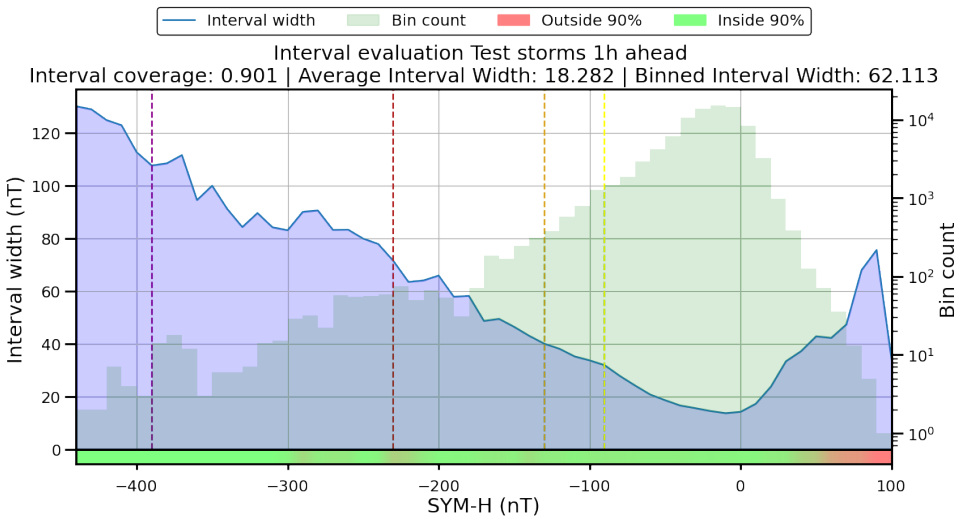


Figure 5.14: Confidence interval analysis for the 1 hour ahead forecast on all the test storms.

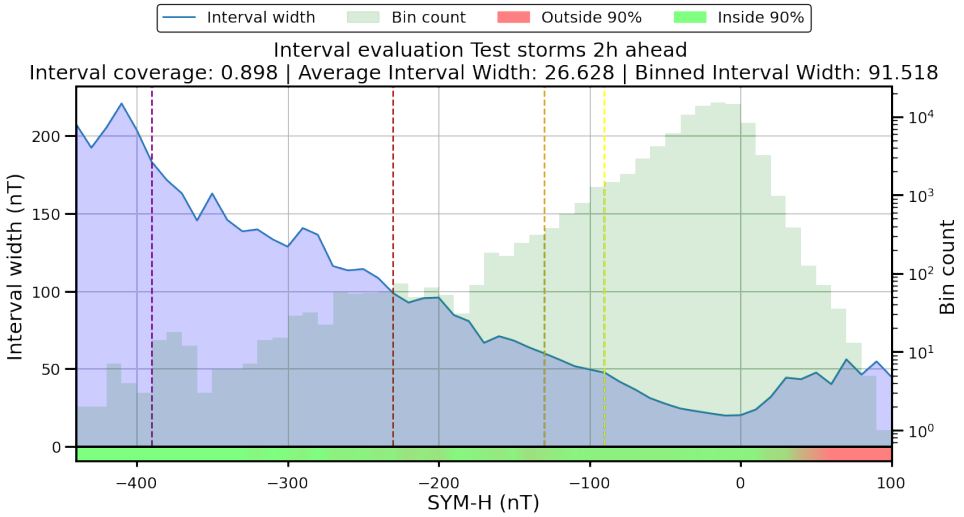


Figure 5.15: Confidence interval analysis for the 2 hours ahead forecast on all the test storms.

and 83% for the 1 and 2 hours forecasts. This is caused by the usage of the preliminary data in the solar wind and IMF measurements, making the forecast more difficult and the uncertainties higher.

Table 5.11: Metrics computation for the test key storms on the 1 hour ahead forecast. Dates are in DD/MM/YY-DD/MM/YY format.

#	**Storm date**	BFE	RMSE	R2	PICP	PIAW	PIBW
117	24/08/18-31/08/18	9.719	5.966	0.974	0.884	15.135	27.199
118	26/08/21-02/09/21	4.660	4.467	0.940	0.882	12.675	17.720
119	02/11/21-09/11/21	11.881	8.074	0.875	0.757	14.307	21.092
120	12/01/22-19/01/22	4.474	4.528	0.958	0.920	13.942	19.158
121	11/03/22-19/03/22	11.795	6.522	0.841	0.813	12.619	26.648
122	05/11/22-12/11/22	7.468	4.706	0.952	0.903	12.322	17.814
	Mean:	8.333	5.711	0.923	0.860	13.500	21.605
	Global:	11.961	5.875	0.944	0.859	13.480	27.611

Table 5.12: Metrics computation for the test key storms on the 2 hours ahead forecast. Dates are in DD/MM/YY-DD/MM/YY format.

#	Storm date	BFE	RMSE	R2	PICP	PIAW	PIBW
117	24/08/18-31/08/18	16.503	9.680	0.932	0.858	22.233	39.701
118	26/08/21-02/09/21	6.960	6.451	0.874	0.881	18.787	26.465
119	02/11/21-09/11/21	17.621	11.080	0.765	0.733	20.535	28.681
120	12/01/22-19/01/22	6.715	6.235	0.920	0.911	19.727	28.025
121	11/03/22-19/03/22	15.375	9.531	0.661	0.766	19.174	37.230
122	05/11/22-12/11/22	12.186	7.806	0.868	0.853	18.722	25.941
	Mean:	12.560	8.464	0.837	0.834	19.863	31.007
	Global:	19.691	8.670	0.879	0.832	19.847	38.451

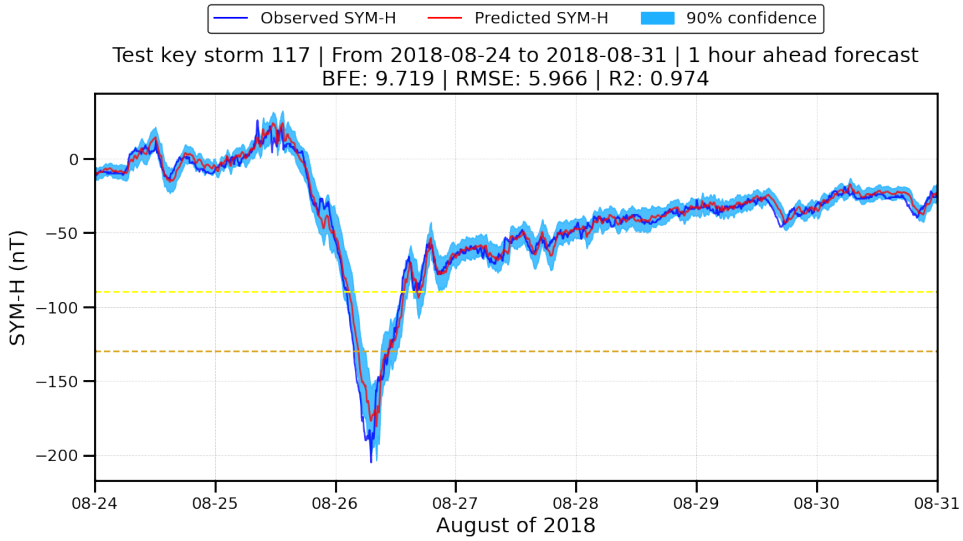


Figure 5.16: 1 hour ahead forecast for the moderate test key storm of August 2018.

Figure 5.16 shows an example of the 1 hour ahead forecast for the moderate test key storm of August 2018. Figures 5.17 and 5.18 show the BFE computation for the test key storms on both lead times. Similarly to the global test analysis, the BFE increases for the more intense values. However, due to the increased difficulty in the forecast, a higher percentage of values are outside of the target confidence interval, especially on the extremes of the possible values.

The prediction confidence interval presented in Figure 5.19 and 5.20 depict the interval analysis for the test key storms, both for the 1 and 2 hours ahead forecast. Similar to the test storms, the intervals are wider during periods of intense activity, reflecting the increased uncertainty in these conditions. However, forecasting using the preliminary parameters is considerably more difficult. This is reflected in the worse interval coverage compared to the target 90%, and generally more values outside of the prediction interval around the storm peak values and the most positive values of the SYM-H.

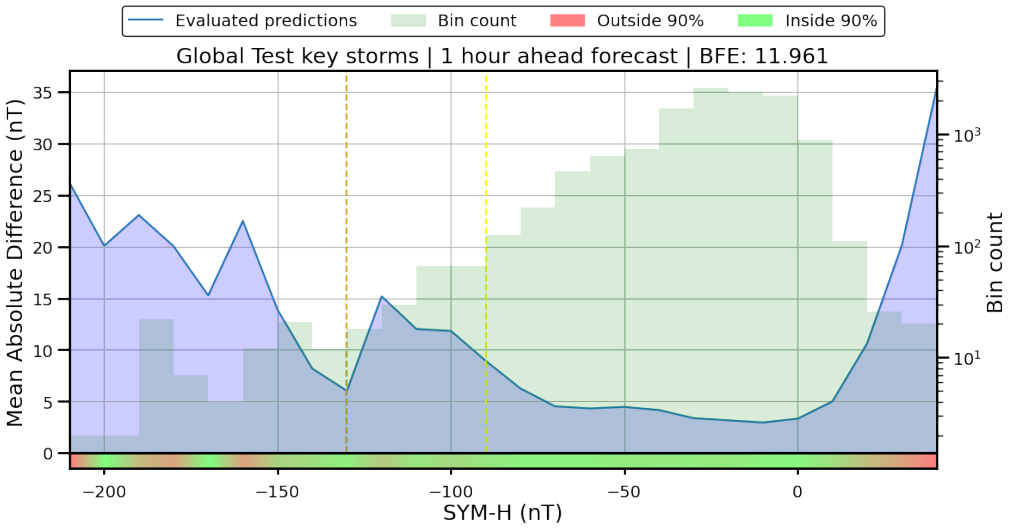


Figure 5.17: BFE plot for the 1 hour ahead forecast on all the test key storms. The bottom heatmap shows the percentage of values inside the prediction confidence interval.

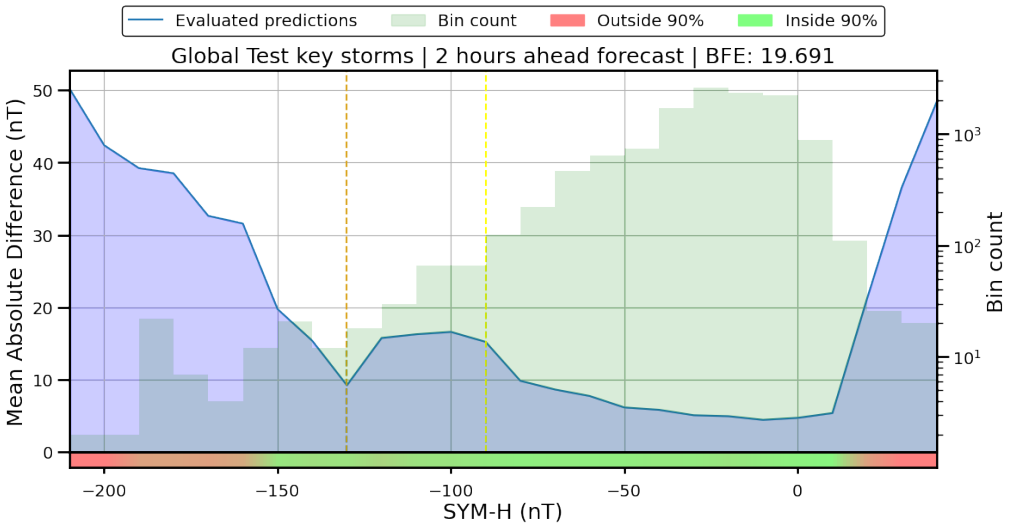


Figure 5.18: BFE plot for the 2 hours ahead forecast on all the test key storms. The bottom heatmap shows the percentage of values inside the prediction confidence interval.

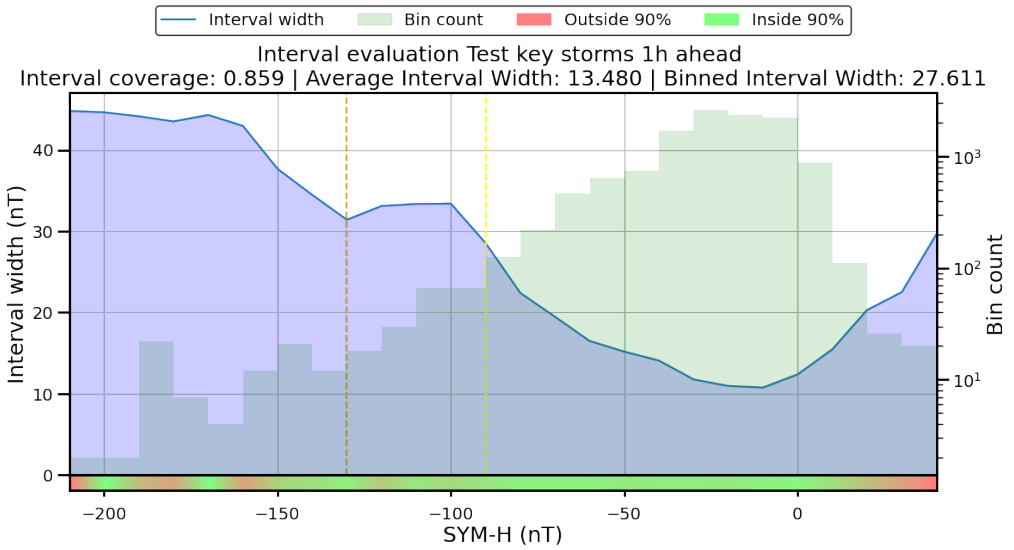


Figure 5.19: Prediction confidence interval analysis for the 1 hour ahead forecast on all the test key storms.

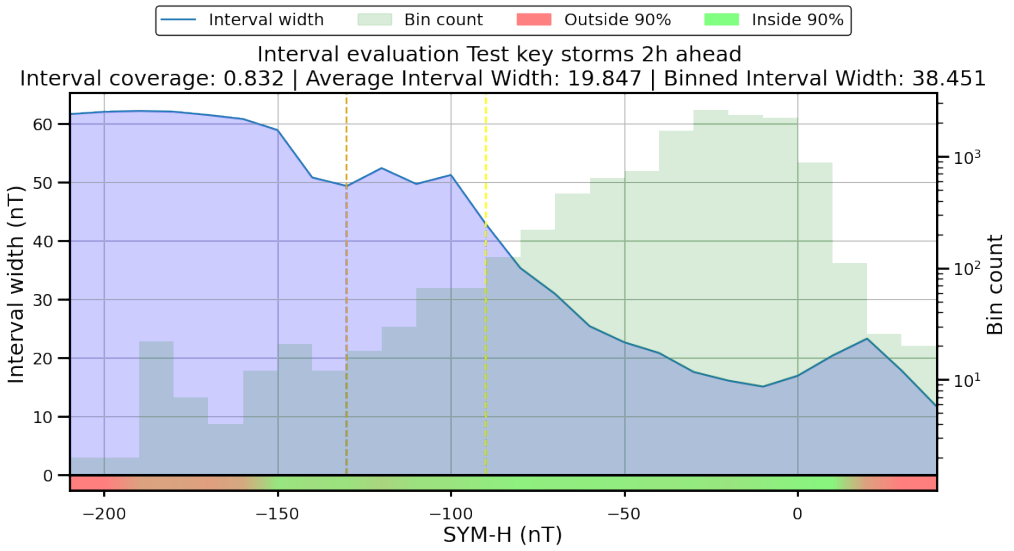


Figure 5.20: Prediction confidence interval analysis for the 2 hours ahead forecast on all the test key storms.

5.3 Conclusions

The retraining of the DNN models using the extended datasets for SYM-H and ASY-H indices has led to improvements in forecasting performance, particularly for the most extreme values of the SYM-H index. This highlights the critical role of careful data separation and expansion in the training process, as geomagnetic storm forecasting models rely heavily on accurately predicting rare, high-impact events. By exposing the model to a broader variety of solar wind phenomena, including disturbances caused by HSSs and CMEs, we ensured that the network could learn the distinct geomagnetic disturbance patterns associated with each type of event. This diversity in training data enhances the model's generalization capabilities, improving its ability to forecast extreme activity.

A key advancement demonstrated in this study is the successful integration of quantile forecasting into the SYM-H prediction models. The use of quantile-based confidence intervals provides valuable insights into the uncertainty surrounding the forecasted values, tailoring prediction intervals to reflect the varying conditions of each forecast. This allows for more adaptive confidence intervals that account for the phase of the geomagnetic storm, enhancing the model's reliability and operational usefulness. By learning these intervals directly during the training process, the model is able to capture not only the central forecast but also the uncertainty associated with each prediction, offering decision-makers a richer set of information for managing geomagnetic storm risks.

The integration of quantile forecasting did not come at the cost of predictive accuracy. On the contrary, this addition complements the model's core forecasting ability by offering both point forecasts and predictions intervals. The model remains robust and shows improved performance in predicting extreme values of SYM-H, an essential feature for operational forecasting during severe geomagnetic storms. The operational evaluation using preliminary observations, further underscores the model's potential for real-time applications, demonstrating that the model can perform reasonably well even with provisional data.

However, while the model shows substantial improvement, there are limitations, particularly when predicting the most extreme geomagnetic events. The uncertainties associated with extreme value predictions remain large, primarily due to the scarcity of training samples in these ranges. Although a 90% interval coverage is commonly used in risk analysis and provides broad intervals to cover the uncertainty, the wide range of predictions in the most extreme cases may reduce the practical utility of the model in operational settings. Tailoring the prediction intervals to the specific needs of end-users could mitigate this issue, enabling more focused and contextually appropriate forecasts. Importantly, the SW community has yet to establish consensus on acceptable levels of uncertainty for the SYM-H index, though guidelines like the 30% accuracy requirement for Dst under ESA's Space Situational Awareness program for SW serve as a useful reference [176].

In summary, this chapter has demonstrated that expanding training datasets and incorporating quantile forecasting into geomagnetic prediction models significantly improves their performance and utility. While there is still room for improvement, particularly in addressing the large uncertainties during extreme events, the advancements presented here represent a step forward in the operational forecasting of space weather phenomena.

Chapter 6

Local indices

"Space weather forecasting: because predicting rain was too down-to-Earth."

The SYM-H index is a global measure of geomagnetic disturbances, derived by averaging the horizontal component of the magnetic field (H) from six magnetic observatories distributed evenly across different longitudes. These observatories are selected from a larger set of 11, allowing for flexibility based on data availability. However, this method of averaging data from different observatories to compute a global index, while useful for a broad understanding of geomagnetic activity, introduces significant limitations.

The process of deriving SYM-H from these observatories can result in critical local geomagnetic information being overlooked. The aggregation of data from observatories worldwide inherently masks critical localized geomagnetic disturbances. In some cases, local disturbances can be significantly more intense than what the global SYM-H index reflects, particularly in mid-latitude regions or areas near auroral zones. Conversely, during global geomagnetic storms, localized disturbances in certain regions may be less intense, despite the global storm activity suggested by SYM-H.

This mismatch between local and global geomagnetic indices underscores the necessity of a more localized approach to SW forecasting. While global indices like SYM-H provide a valuable overall view of geomagnetic activity, they fail to capture the intricacies of localized disturbances, which can be critical for understanding SW effects on regional technological systems.

This chapter focuses on forecasting a local geomagnetic index, the Local Disturbance Index (LDi). We will use a selection of stations detailed in Section 6.1.1 and train a modified version of the NN presented in Chapter 3, adapted to work with the local indices while incorporating the confidence intervals presented in Section 5.2.

6.1 Local disturbance index

The LDi is a geomagnetic index specifically designed to quantify localized geomagnetic disturbances. Developed by Cid et al. [35], the LDi focuses on capturing geomagnetic variations at individual observatories by subtracting the solar quiet (Sq) variation and a baseline from the horizontal component (H) of the geomagnetic field. This method removes the regular solar influence and isolates the disturbances that are locally generated, allowing for a clearer representation of the geomagnetic conditions at mid-latitude stations.

We have chosen the LDi to measure local geomagnetic disturbances due to its ability to provide localized disturbances compared to global indices like SYM-H. While SYM-H offers a valuable global perspective by averaging data from multiple observatories, it inevitably masks critical local information due to its global averaging process. In contrast, LDi captures the nuances of geomagnetic activity at specific geographic locations, making it particularly effective for monitoring regional SW effects, which are vital for protecting local technological infrastructure from SW-induced disruptions, such as GICs.

Another key advantage of using the LDi is its flexibility in contributing to global indices like SYM-H. By combining data from several LDi calculations across different longitudes, a global picture of geomagnetic activity can be reconstructed. Specifically, the LDi can be used to derive the SYM-H index by averaging local disturbance measurements from the observatories used to calculate it. This approach allows for a better representation of both global and local disturbances, enhancing the accuracy of SW forecasting systems.

By leveraging the localized insights provided by the LDi, we aim to develop a forecasting system that not only captures the overall geomagnetic storm intensity but also identifies and predicts regional disturbances, which are often hidden in global indices.

The LDi has been validated for use not only in the Northern Hemisphere but also in the Southern Hemisphere, as demonstrated in Nahayo et al. [96]. Their study successfully applied the LDi and its derivative, the Local Current Index (LCi), to nowcast local geomagnetic disturbances during major geomagnetic storms, including the Halloween storm of 2003 and the Saint Patrick's Day storm of 2015. The results from magnetic observatories and GIC data in South Africa confirmed that LDi can effectively capture localized SW impacts, even in regions where global indices like SYM-H may underestimate the severity of geomagnetic events. This validation underscores the robustness and adaptability of LDi in providing precise local forecasts across hemispheres.

6.1.1 Selected observatories

For our analysis, we have initially selected five geomagnetic observatories, which are detailed in Table 6.1. These observatories are chosen to provide sufficient coverage across different longitudes, ensuring that we capture the full range of geomagnetic disturbances globally.

All of the observatories are part of the INTERMAGNET network, which is a global network of observatories that continuously monitor the Earth's magnetic field. The INTERMAGNET observatories are recognized for their high data quality and have been

Table 6.1: Selected Geomagnetic Observatories from the INTERMAGNET network.

Identifier	Full Name	Country	Latitude	Longitude
ABG	Alibag	India	18.638	72.872
HON	Honolulu	United States of America	21.32	-158.0
MMB	Memambetsu	Japan	43.91	144.19
TUC	Tucson	United States of America	32.17	-110.729
SPT	San Pablo-Toledo	Spain	39.55	-4.35

utilized extensively in geomagnetic research, including in the calculation of global indices such as SYM-H.

The SYM-H index is derived using data from a set of observatories located at mid-latitudes, distributed across different longitudes to provide a global view of geomagnetic activity. All the selected observatories, except for San Pablo-Toledo (SPT), are part of this official network used to calculate SYM-H. For the European region, we have opted to use the SPT observatory in Spain instead of the Chambon la Forêt (CLF) observatory in France. This choice was made due to SPT's proximity to our research location in Spain, as it covers a similar geographic area while still offering high-quality data.

We have selected these five observatories to ensure sufficient longitudinal coverage while also prioritizing data availability and quality. Where multiple observatories exist within a region, such as in the United States area, we have selected the Tucson (TUC) observatory, since is the one that had the best data availability and quality. This allows us to produce a comprehensive analysis that accounts for both local and global geomagnetic variations. Figure 6.1 depicts the selected observatories.

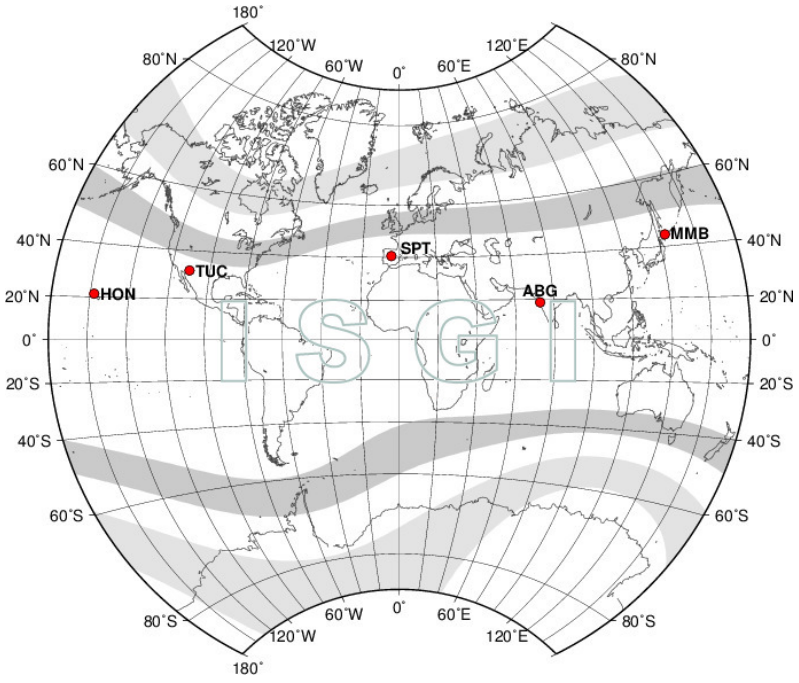


Figure 6.1: Selected observatories for the local indices forecasting model.

6.1.2 Importance of local indices: Case studies of major geomagnetic storms

To highlight the importance of local geomagnetic indices in capturing regional SW disturbances, we present two case studies using data from the SYM-H index and two LDis during two major geomagnetic storms. Figures 6.2 and 6.3 show the variation in disturbances across specific observatories compared to the global SYM-H index.

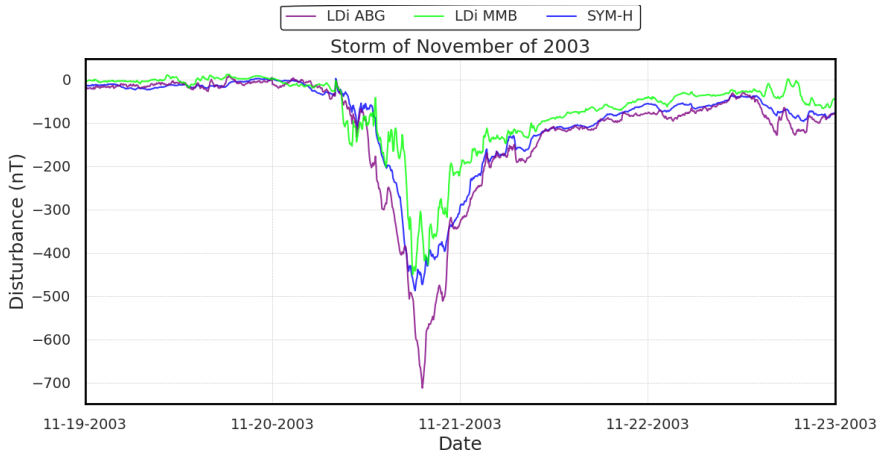


Figure 6.2: Comparison of the SYM-H index with the LDi at ABG (India) and MMB (Japan) during the geomagnetic storm of November 2003.

In Figure 6.2, we observe the differences between the global SYM-H index and the local disturbances experienced at the Alibag (ABG) and Memambetsu (MMB) observatories during the geomagnetic storm of November 2003. The disturbance at ABG is significantly larger than that captured by the SYM-H index, while the disturbance at MMB is smaller overall. This demonstrates the variability in how different regions experience geomagnetic storms, which may be obscured when relying solely on global indices like SYM-H. The local index at ABG reveals a stronger geomagnetic response compared to the global average, which can have significant implications for regional systems, such as power grids, that are sensitive to geomagnetic activity.

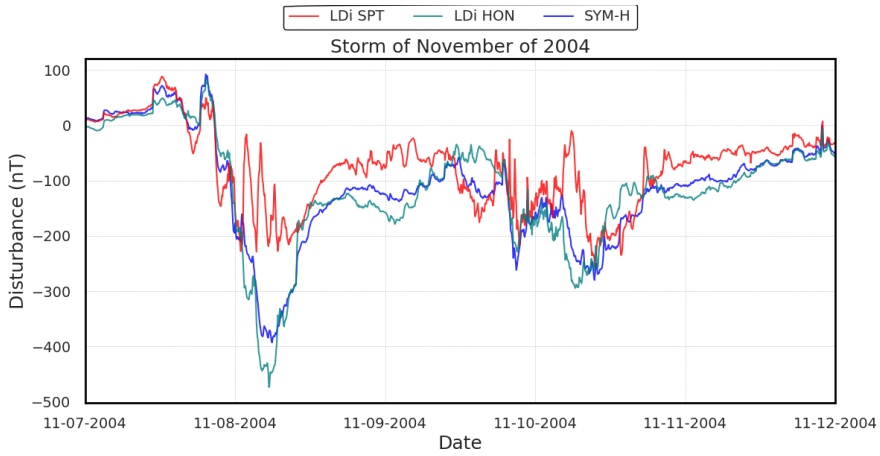


Figure 6.3: Comparison of the SYM-H index with the LDi at SPT (Spain) and HON (USA) during the geomagnetic storm of November 2004.

In Figure 6.3, we compare the local disturbances at SPT (Spain) and Honolulu (HON) (USA) during the geomagnetic storm of November 2004. The figure illustrates that, at

the most intense peak, the disturbance at SPT is notably lower than at HON. However, in other phases of the event, the disturbance at SPT surpasses that of HON, and in some instances, the disturbance at SPT aligns more closely with the global SYM-H index. This variability further reinforces the point that local disturbances can behave very differently from the global average, with consequences that can be severe in some regions while remaining moderate in others. These regional differences highlight the need for local indices to understand and mitigate the effects of geomagnetic storms more precisely at the local level.

These examples clearly illustrate that local geomagnetic indices like LDi are essential for capturing the full scope of geomagnetic activity during storms. Relying solely on global indices like SYM-H can lead to a misrepresentation of the true impact of SW events on specific regions, underscoring the importance of local indices for regional forecasting and mitigation efforts.

6.1.3 Influence of magnetic local time on geomagnetic disturbances

Geomagnetic disturbances exhibit significant variability depending on the rotation of the Earth, as these disturbances are highly dependent on the Magnetic Local Time (MLT) of the affected region [177]. MLT, a measure of time based on the geomagnetic longitude of an observatory, influences the magnitude and nature of the disturbance experienced at different points on the Earth's surface [178]. Due to the Earth's rotation, regions that are in the dawn sector of the magnetosphere often experience different disturbances compared to those in the dusk sector [179]. For example, geomagnetic storms tend to be more intense around dawn [180], [181]. These variations can lead to significant differences in the observed geomagnetic disturbances at local observatories even during the same storm event [182], [183].

Magnetic Local Time is calculated by adjusting the Universal Time (UT) of a location based on its geomagnetic longitude. The Earth's magnetic field is slightly offset from its rotational axis, meaning that MLT does not correspond directly to geographic local time. Specifically, the MLT of a location is calculated using Equation 6.1, where λ is the geomagnetic longitude of the observatory in degrees, and UT is the Universal Time. The factor 15° corresponds to the Earth's rotation rate, where 360 degrees equals 24 hours, making each 15 degrees equivalent to one hour of time. This results in a time system where "midnight" MLT corresponds to the point on Earth directly opposite the Sun in terms of geomagnetic alignment, and "noon" MLT is the point directly facing the Sun. The rotation of the Earth causes local disturbances to vary with MLT, which is why geomagnetic storms often affect regions differently depending on their magnetic longitude.

$$\mathrm{MLT} = \mathrm{UT} + \frac{\lambda}{15^\circ} \tag{6.1}$$

In addition to the spatial variability of geomagnetic disturbances observed during storms, the temporal aspect, specifically the MLT, plays a critical role in determining the intensity and characteristics of the disturbance. The MLT accounts for the Earth's rotation and its impact on how different regions experience geomagnetic activity. Figures 6.4 and 6.5 show the previous storms, now including the MLT for the selected observatories during the storms.

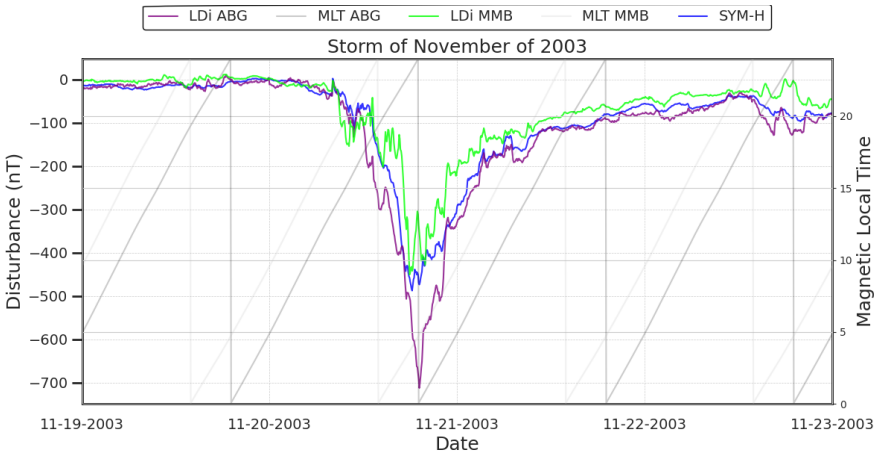


Figure 6.4: Comparison of the SYM-H index with the LDi at ABG (India) and MMB (Japan) during the geomagnetic storm of November 2003, with corresponding MLT.

In Figure 6.4 the disturbance at ABG (India) is notably more intense than at MMB (Japan), and this variation is influenced in part by their differing MLTs during the storm. ABG is positioned in the dawn sector during key phases of the storm, where disturbances tend to be more severe due to enhanced ionospheric currents. In contrast, MMB, located in the dusk sector during most of the storm, experiences a weaker disturbance. The differences in MLT between these two locations clearly illustrate the impact of Earth's rotation on geomagnetic activity.

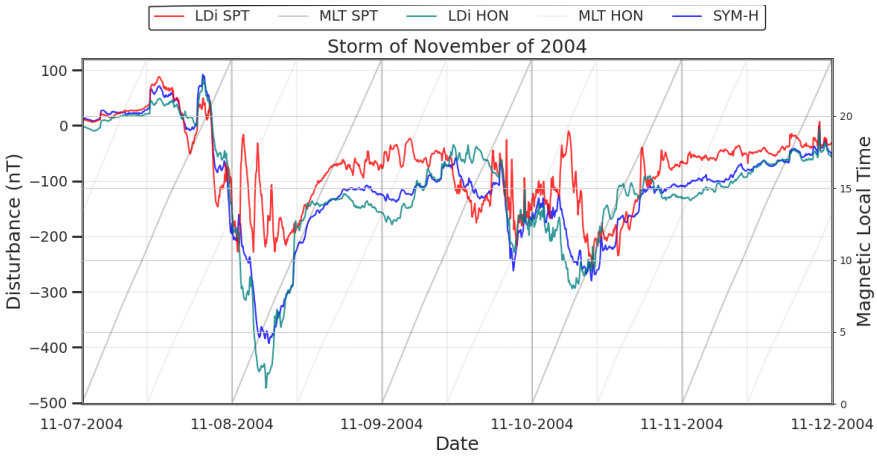


Figure 6.5: Comparison of the SYM-H index with the LDi at SPT (Spain) and HON (USA) during the geomagnetic storm of November 2004, with corresponding MLT.

Similarly, in the Figure 6.5 we observe variations in disturbance levels at SPT (Spain) and HON (USA), which are also influenced by their respective MLTs. During the most intense peak, SPT is in a sector closer to midnight MLT, where the disturbance is generally lower compared to the noon sector, where HON is located during the peak. However, at other times, SPT moves into regions of MLT where geomagnetic activity increases, sometimes exceeding that of HON. These fluctuations in disturbance levels underscore the importance of considering MLT when analyzing local geomagnetic effects during global storms.

To explore the influence of MLT on the LDi more thoroughly, we have performed a detailed analysis by separating geomagnetic activity into active and inactive periods.

The active periods are defined when the SYM-H index drops below -100 nT, indicating a significant geomagnetic storm, while inactive periods are defined when SYM-H remains above -50 nT. By differentiating between these periods, we aim to highlight how the influence of MLT varies depending on the level of geomagnetic activity.

For each observatory, we calculate the difference between the LDi of the station and the global SYM-H index. These differences are then grouped according to the MLT of the respective station in 15-minute intervals. The results are plotted against MLT to reveal how the local disturbances deviate from the global average during both active and inactive periods.

As expected, during the inactive periods, shown in Figure 6.6, the differences between LDi and SYM-H are generally small and stable across all stations. The variations in MLT do not significantly affect the magnitude of local disturbances, as the global geomagnetic activity remains low.

However, during the stormy periods, depicted in Figure 6.7, there is a clear pattern of variability depending on the MLT. The most significant differences are observed around 5 MLT (dawn), where the local disturbances greatly exceed the SYM-H index at the SPT, MMB and TUC stations. Conversely, the differences reach their lowest point around 18 MLT (dusk). This pattern indicates that dawn sectors are more likely to experience stronger disturbances, while dusk sectors generally observe weaker local effects.

It is also important to note the role of latitude in this pattern. Observatories at lower latitudes, such as HON and ABG, exhibit less pronounced differences at dawn and even experience negative differences at dusk, indicating weaker local disturbances compared to the global SYM-H index. This further suggests that the latitude of the station, in addition to MLT, plays a critical role in determining the magnitude of local geomagnetic disturbances during storm events.

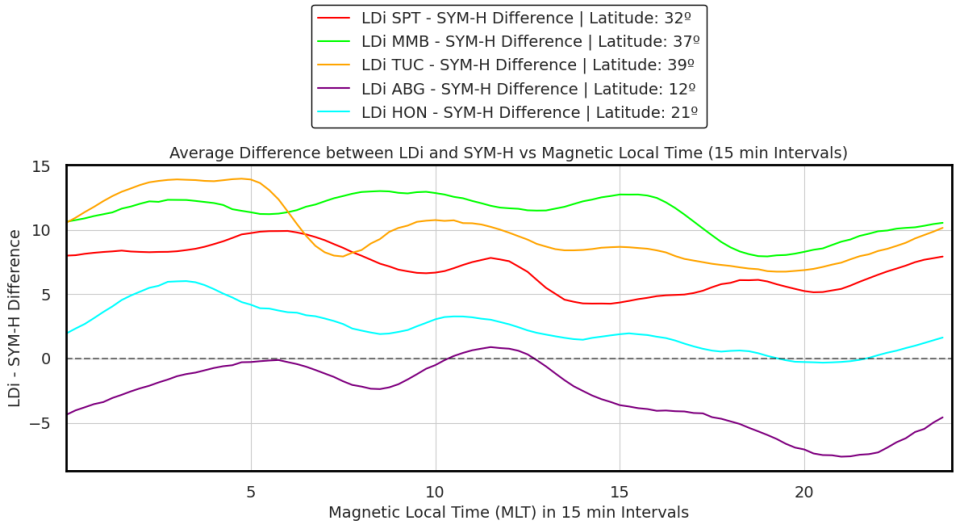


Figure 6.6: Average difference between the LDi at the selected observatories and the SYM-H during quiet periods (grouped by MLT in 15-minute intervals).

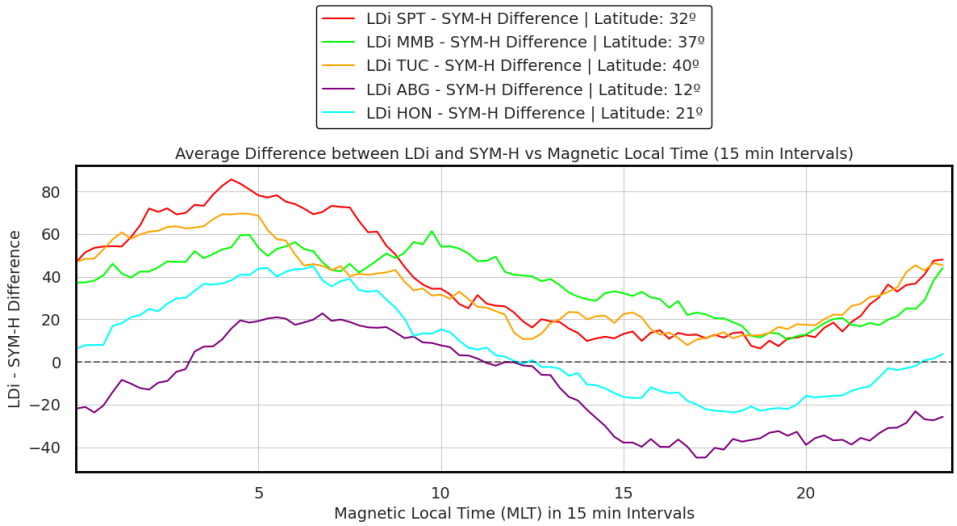


Figure 6.7: Average difference between the LDi at the selected observatories and the SYM-H during stormy periods (grouped by MLT in 15-minute intervals).

6.2 Neural network for local indices forecasting

To address the challenge of forecasting local geomagnetic indices, there are two primary modeling strategies:

- Developing an independent model for each station and training, validating, and testing it with the data specific to that station.
- Developing a compound model trained with data from multiple stations.

Each approach presents unique challenges, particularly when considering the complexities of geomagnetic activity and its interaction with the Earth's magnetic field at different latitudes, longitudes, and MLT.

6.2.1 Independent model for each station

Developing a model for each station offers the advantage of focusing specifically on the unique characteristics of that station's geomagnetic activity. However, this approach encounters a fundamental limitation: the scarcity of data. Even when forecasting the global SYM-H index we faced a shortage of samples, especially for intense and super-intense geomagnetic storms. By ML standards, the available data was already limited in quantity, and this issue becomes even more pronounced when trying to forecast local indices.

The primary challenge here is the dilution of storm samples across different MLT sectors (dawn, noon, dusk, and midnight) for each station. For instance, consider the geomagnetic storm of November 2004 (Figure 6.5). One of the four previous super-intense storms recorded in recent decades does not even qualify as an intense storm at a particular station when using the thresholds established in Section 4.1.1. This highlights the problem of data scarcity: for any given storm, we must also account for the specific

MLT sector during which the disturbance reaches the station, further reducing the already limited number of training samples.

If we had a sufficient number of samples from storms of all intensities, covering a range of conditions as they passed through different MLT sectors, training an independent model for each station would be a viable option. However, given the current data limitations, this approach is not feasible, as the available data is too sparse for robust model training.

6.2.2 Compound model for multiple stations

The alternative approach is to train a single, compound model using data from multiple stations. This strategy has the advantage of increasing the dataset size by pooling data from different observatories. However, this introduces additional complexity, as the NN must now account for spatial and temporal differences between stations.

To make this compound model effective, it is essential to provide the network with relevant context about each station's specific conditions. In particular, MLT must be included as an input feature, as we have already seen how strongly it influences the LDi across different stations. Additionally, the model must take into account the station's geographic location, including its magnetic longitude and latitude, because, as demonstrated by the differences in Figure 6.7, latitude also plays a crucial role in determining the intensity of geomagnetic disturbances.

The key challenge here lies in the increased complexity of the training process. Not only the model must learn to predict the geomagnetic disturbances caused by the IMF and solar wind parameters measured by ACE, but it must also learn to map those disturbances to each station's specific location and MLT. The same storm, as observed by ACE, can have very different LDi profiles depending on the station's position on Earth and the MLT at the time the disturbance reaches the Earth. Therefore, the NN needs to develop the capacity to generalize across various spatial and temporal conditions, which makes the training process more difficult.

6.2.3 Modeling the network

To address the challenge of forecasting local geomagnetic indices using data from multiple stations, we will proceed with the second option: training a compound model. This approach will enable us to utilize a broader dataset by incorporating observations from several stations, which helps to mitigate the limitations posed by the scarcity of data at individual observatories.

For simplicity and consistency with our previous work on forecasting the SYM-H index, we will use the same data subsets that were employed for SYM-H forecasting. These subsets have already been split into training, validation, and testing sets, ensuring that the results remain comparable. However, the addition of extra features, such as MLT, requires adapting the network that we have defined in Chapter 3 to properly process the extra features. But, even before adapting the network, the first step is to decide how to encode the new input features. While the LDi can be standardized in a similar manner as the SYM-H, the MLT has a completely different nature.

6.2.3.1 Encoding magnetic local time

Since the MLT is a cyclical feature, it must be encoded in a way that preserves its cyclical nature. To achieve this, we represent MLT using both sine and cosine transformations. This technique is commonly used in ML applications where cyclical features such as the time of day or seasonal information need to be encoded. By mapping MLT to sine and cosine components, we ensure that the cyclical pattern is preserved, meaning that the values for 23:59 MLT are close to those for 00:00 MLT, rather than being treated as distant points on a linear scale.

This encoding is calculated using the transformations as in Equations 6.2 and 6.3. This approach is preferred over standardization, since in that case, the maximum distance would happen between 23:59 (+1) to 0:00 (-1), while those two points should be contiguous.

$$\mathrm{MLT}_{\sin} = \sin\left(2\pi \times \frac{\mathrm{MLT}}{24}\right) \tag{6.2}$$

$$\mathrm{MLT}_{\cos} = \cos\left(2\pi \times \frac{\mathrm{MLT}}{24}\right) \tag{6.3}$$

6.2.3.2 Normalizing longitude and latitude

In addition to MLT, we also give the model the location of each station, specifically its magnetic longitude and latitude, as input features. Since the other features in the model have been normalized, we have also normalized the longitude and latitude features. For the latitude, since we are using working with from the Northern Hemisphere, we are taking the absolute value of the latitude, and then normalizing it between 0 and 1. Otherwise, if we were to try to forecast the LDi of a Southern Hemisphere station, the model would encounter a negative latitude that it has never seen before. For the longitude we have done the same normalization approach.

6.2.3.3 Neural network architecture

To accommodate these additional inputs, we have made adjustments to the base NN architecture used for SYM-H forecasting, introduced in Section 3.3.2 and expanded to include the quantile forecasts as defined in Section 5.2.2. The architecture is depicted in Figure 6.8. The primary differences compared to the architecture used for SYM-H forecasting lie in the handling of the inputs, particularly the position features (MLT, longitude, and latitude) and the LDi inputs.

In this updated architecture, we have split the inputs of the IMF and Solar Wind parameters from the LDi and positional information (MLT and geographic coordinates) at the start of the network. The reason for this separation is that the processing of IMF and Solar Wind data, measured by ACE, should be common across all stations. Therefore, these features are passed through the convolutional layers, which extract higher-level representations of the solar wind and IMF dynamics.

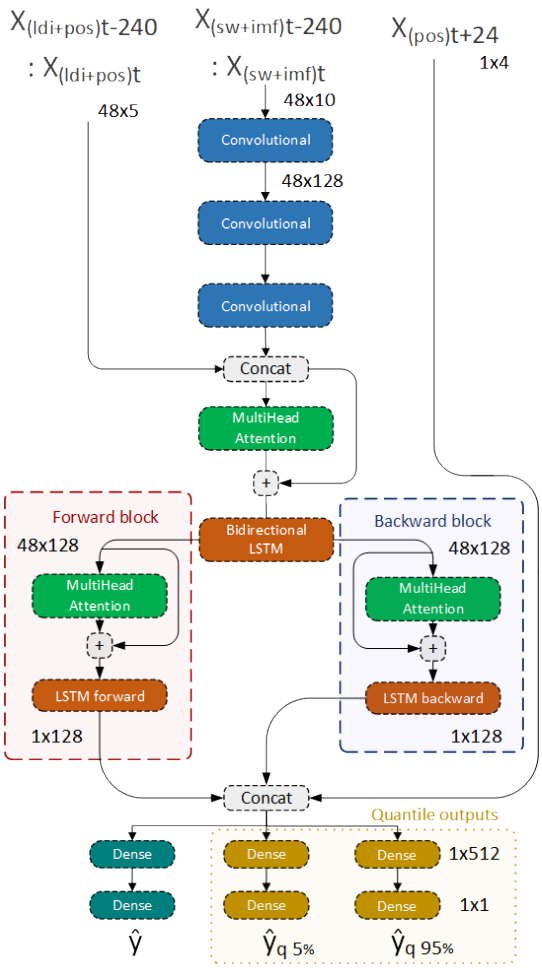


Figure 6.8: Neural network architecture for local geomagnetic index forecasting.

Once the IMF and solar wind data have been processed by the convolutional layers, the output is concatenated with the LDi, MLT, and positional information (longitude and latitude) of the station. This combined input is then fed into a multi-head attention layer, where the impact of the storm on the local indices is reweighted based on the additional spatial and temporal context provided by the MLT and station position. This layer models the influence of the MLT and positional information on the disturbance, allowing the network to account for local effects that vary across stations.

From this point, the architecture remains largely the same as the original SYM-H forecasting network, until the concatenation of the forward and backward blocks after the bidirectional LSTM network. In this version, we are also adding the future MLT and positional information of the station for the time we are forecasting. Since the future positions and MLT are known in advance, this information is provided to the network to ease the training process. Although, in theory, the network could learn these relationships on its own, the limited amount of training data makes it advantageous to directly supply this information and reduce the complexity of the task for the model.

6.2.4 Training the network

The NN model for local geomagnetic index forecasting is trained using the storm subsets presented in Section 4.2.1.1. Since the model produces multiple outputs (to provide the confidence interval along the index prediction), we apply different loss functions for each output. For the quantile outputs, corresponding to the 5th and 95th percentiles, we use the quantile loss function. This allows the network to predict the range of potential disturbances, providing uncertainty estimates along with the central forecast.

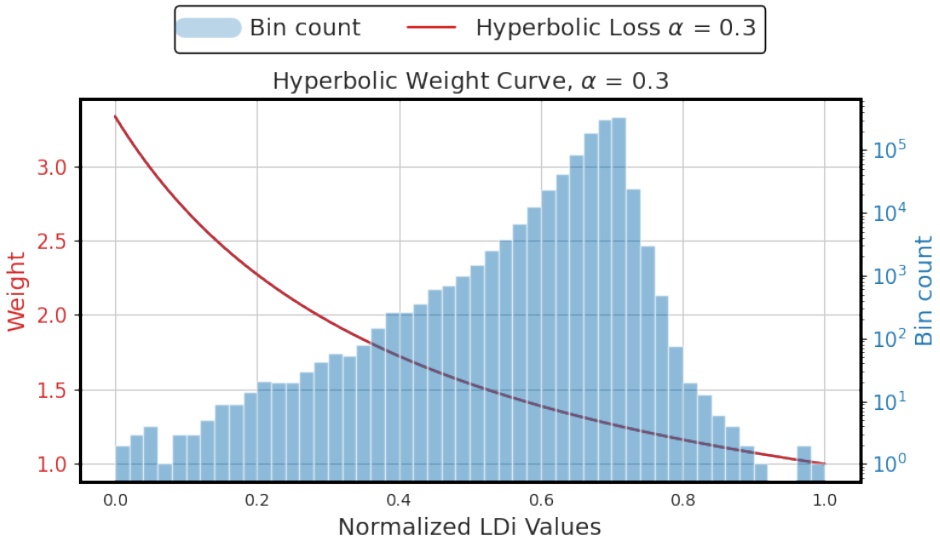


Figure 6.9: Example of the hyperbolic loss function weight curve with $\alpha = 0.3$, showing the increased weight for higher-intensity LDi values.

However, we introduce a significant change to the loss function used for the main forecast output. Instead of using the MSE loss, which tends to focus on minimizing the average error, we have opted for a hyperbolic loss function. As discussed in the BFE analysis, focusing solely on the mean error is not suitable for this type of problem, where extreme events, or outliers are of particular importance. The hyperbolic loss function adjusts the weight of each sample based on the intensity of the corresponding label, giving more importance to rarer, more intense geomagnetic disturbances, as illustrated in Figure 6.9.

The hyperbolic loss function is defined in Equation 6.4, where α controls the slope of the weighting function, y_{true} and y_{pred} represent the true and predicted values of the LDi, and $y_{\min}$ and $y_{\max}$ are the minimum and maximum values of the normalized LDi range. The weights applied to each sample increase as the intensity of y_{true} increases, ensuring that higher-intensity (less frequent) samples have a greater impact on the model's training. This loss function introduces the hyperparameter α to the training of the network, which controls the shape of the hyperbolic curve. As shown in Figure 6.9, smaller intensity values receive lower weights, while rarer and more intense values are weighted more heavily, reflecting their importance in geomagnetic forecasting.

$$L(y_{\text{true}}, y_{\text{pred}}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(y^i_{\text{true}} - y^i_{\text{pred}})^2}{\alpha(1-\hat{y}^i_{\text{true}}) + \hat{y}^i_{\text{true}}}\right] \tag{6.4}$$

$$\hat{y}^i_{\text{true}} = \frac{y^i_{\text{true}} - y_{\min}}{y_{\max} - y_{\min}}$$

We have selected the hyperbolic loss function based on the findings of Aguado et al. [143], who demonstrated that the recovery phase of the geomagnetic Dst index during storms is better represented by a hyperbolic decay function than an exponential one. Given the similarities between the Dst index and the SYM-H index, which we use for global storm forecasting, this hyperbolic function is a natural choice for improving the network's ability to forecast local disturbances. While using this loss function does not have a massive impact on the overall performance of the model, it slightly improves the performance on the most extreme cases, showing a higher improvement on the BFE over the RMSE. The proposed hyperbolic loss function enhances the gradients only for the critical samples, those representing the rare and intense events, while leaving the rest unchanged. This allows the model to prioritize important events without drastically altering the overall learning rate. Therefore, the hyperbolic loss function provides a balanced approach to gradient calculation, allowing for effective training without needing significant adjustments to the optimizer's hyperparameters.

For the validation loss, instead of using MSE or MAE as we have done in previous models, we have adopted the BFE as the validation metric. This approach allows us to evaluate the model based on how well it captures the different intensities of geomagnetic disturbances. The BFE, as a validation metric, works by first computing the bins and the number of samples in each bin of the validation data. During the validation process, the differences between the predicted values and the actual values are accumulated for each bin, and the average difference is computed. This allows us to track the model's performance across different intensity levels of the geomagnetic disturbances, calculating the metric incrementally during training.

While BFE is useful as a validation metric, it is not suitable as a direct replacement for MSE in the loss function during training. This is because the BFE metric reduces the contribution of each sample to the total loss based on how common its label is. If we applied this principle directly in the loss function, it would result in a significant reduction in the gradients, especially for the majority of samples with common labels. As a consequence, we would need to use a much larger learning rate to compensate for the diminished gradient magnitudes, which could lead to instability during training. Most optimizers are designed to work with relatively low learning rates, and increasing the learning rate to such an extent could cause issues like overshooting the optimal solution.

6.2.5 Model evaluation

We will evaluate the performance of the model in a similar manner to the evaluation performed for SYM-H forecasting. Specifically, we will test the model on two distinct datasets: the test storms and the test key storms. To simulate a real-time forecasting scenario, if there are missing values in the IMF or solar wind data from the ACE satellite, we will fill in the gaps by interpolation only when possible (i.e., when valid neighboring data points exist). If interpolation is not possible, we will propagate the last valid value forward to maintain continuity. Additionally, we will analyze each station separately.

Since this model also generates prediction intervals, we will calculate the metrics introduced in Section 5.2.2.1 to evaluate both the point predictions and the confidence intervals. For each storm, we will compute the following metrics:

- RMSE: to assess the overall error between predicted and true values.
- R^2 (Coefficient of Determination): to measure the proportion of variance explained by the model.
- BFE: to evaluate the model's ability to capture geomagnetic disturbances across intensity levels.
- PICP: to determine how well the prediction intervals cover the true values.
- PIAW: to measure the average width of the prediction intervals.
- PIBW: to measure the width of the prediction intervals across intensity levels.

We will present the results of these metrics separately for each observatory, allowing for a more detailed evaluation of the model's performance at different locations. Since there is no other model that forecasts this local disturbances, we will compare the performance of the model to a persistence one, that uses the last observed value as the forecast. Since the main focus of the project is the operational deployment of the models, we have chosen to forecast only 2 hours in advance, as the data is often provided in near-real-time. That means that 1 hour ahead forecasts results in less than 1 hour, reducing its meaningfulness.

It is important to note, that for some storms, there are missing or incorrect values in the magnetometer of a particular stations. In those cases, we have skipped those storms since we can not replace them using data from other stations, as is done to calculate the SYM-H.

6.2.5.1 San Pablo-Toledo results

Table 6.2 summarizes the evaluation metrics for each test storm at the SPT observatory, comparing the trained model to the persistence one for the two hours ahead forecast. The values for RMSE, R^2, BFE, PICP, PIAW and PIBW provide insight into how well the model predicts local geomagnetic disturbances, both in terms of accuracy (RMSE, R^2, BFE) and the quality of the prediction intervals (PICP, PIAW and PIBW).

Figure 6.10 presents the global BFE metric across all test storms for the SPT observatory. The global BFE is calculated by averaging the BFE values across all bins of intensity

Table 6.2: Metrics of the trained model compared to the persistence model for the test storms for the SPT station for the 2 hours ahead forecast.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
81	17.788	10.271	0.349	15.581	9.840	0.403	0.953	32.652	41.564
82	22.166	19.268	0.646	15.653	15.276	0.777	0.880	41.404	62.785
83	22.534	16.591	0.578	18.233	12.142	0.774	0.932	37.215	52.195
84	23.413	17.283	0.744	18.434	13.199	0.851	0.943	40.384	57.513
85	39.254	19.990	0.523	14.543	10.194	0.876	0.935	30.023	58.420
86	18.514	14.934	0.618	13.441	10.286	0.819	0.931	35.580	46.036
87	15.998	14.156	0.482	14.260	10.809	0.698	0.913	33.287	44.459
88	20.949	16.212	0.552	16.647	13.776	0.677	0.827	33.196	46.697
89	40.946	22.708	0.773	24.566	12.640	0.930	0.881	35.178	70.358
90	23.431	15.744	0.591	25.914	13.068	0.718	0.872	32.423	50.599
91	41.796	25.112	0.784	32.539	18.596	0.881	0.897	41.638	85.089
92	19.466	12.012	0.606	12.798	11.074	0.665	0.879	30.471	58.711
93	23.335	16.004	0.561	14.973	11.867	0.758	0.865	32.410	44.823
94	11.692	9.811	0.781	9.901	8.121	0.850	0.946	28.721	40.023
95	28.147	16.590	0.676	16.406	11.135	0.854	0.905	34.536	54.571
96	26.122	17.684	0.100	24.342	20.541	-0.221	0.881	44.537	66.896
97	35.581	16.911	0.381	29.651	14.847	0.523	0.808	37.290	47.105
98	26.417	18.480	0.312	22.161	14.765	0.561	0.856	36.767	48.406
99	34.953	22.961	0.371	27.139	16.685	0.668	0.873	43.055	50.016
100	28.839	17.384	0.445	20.358	12.581	0.709	0.877	33.615	50.639
101	38.775	30.089	0.716	30.643	24.491	0.812	0.891	43.627	69.555
102	58.187	20.032	-0.474	57.751	17.084	-0.072	0.902	35.119	62.372
103	32.524	18.100	0.303	20.210	14.072	0.579	0.859	33.817	53.751
104	19.956	12.230	0.441	24.468	11.012	0.547	0.891	26.484	40.804
105	18.153	14.961	0.181	15.856	12.419	0.436	0.879	31.680	47.611
106	17.209	13.730	-0.021	11.822	9.154	0.546	0.918	30.529	39.496
107	34.496	13.652	0.510	24.590	10.387	0.716	0.849	25.496	39.479
108	40.983	12.314	0.627	39.949	10.260	0.741	0.901	18.281	34.422
109	20.394	15.631	0.768	20.758	15.026	0.786	0.859	34.022	55.555
110	17.383	13.564	0.523	13.713	9.797	0.751	0.902	28.797	42.071
111	18.545	11.364	0.544	13.675	9.032	0.712	0.915	28.459	44.514
112	12.981	11.599	0.651	11.810	9.747	0.753	0.884	28.022	39.858
113	32.133	13.754	0.385	25.908	11.293	0.586	0.860	27.600	53.393
114	29.989	21.262	0.784	22.810	15.907	0.879	0.868	39.763	68.036
115	13.410	11.983	0.647	13.851	10.626	0.723	0.881	27.925	36.252
116	20.001	18.265	0.360	16.209	14.090	0.619	0.861	36.262	45.755
Mean:	26.290	16.462	0.495	20.877	12.940	0.663	0.888	33.618	51.384
Global:	42.545	17.295	0.677	33.023	13.666	0.798	0.861	36.262	45.755

for each storm, allowing us to assess the model's performance in capturing geomagnetic disturbances across different magnitudes.

Figure 6.11 shows the prediction interval coverage for SPT across all test storms. It provides a visual representation of how well the prediction intervals cover the true geomagnetic disturbance values, helping us evaluate the reliability and usefulness of the uncertainty estimates provided by the model.

On average, the trained model outperforms the persistence model across all key metrics, including the BFE, RMSE, and Coefficient of Determination (R^2). The global results show that the trained model has a considerably lower global BFE (33.023 compared to 42.545 for the persistence model) and improved RMSE (13.666 vs. 17.295). This reflects that

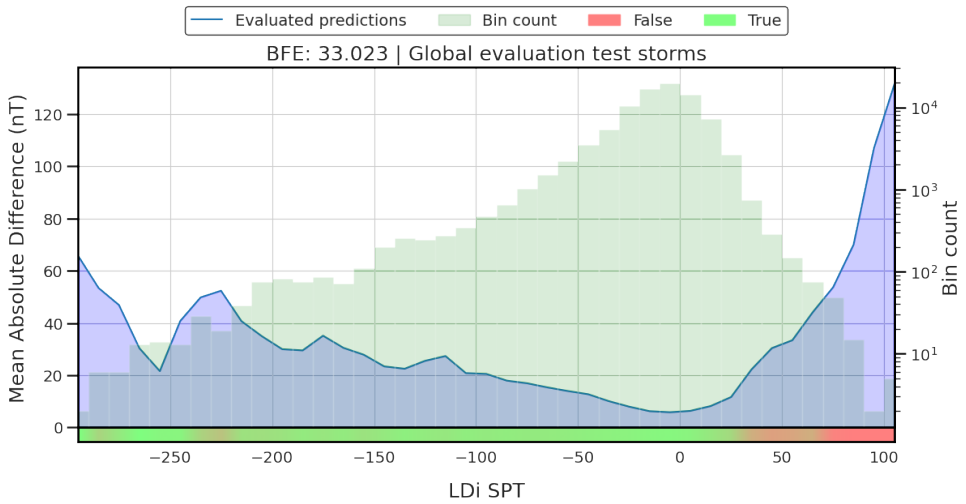


Figure 6.10: Global BFE for SPT on the test storms.

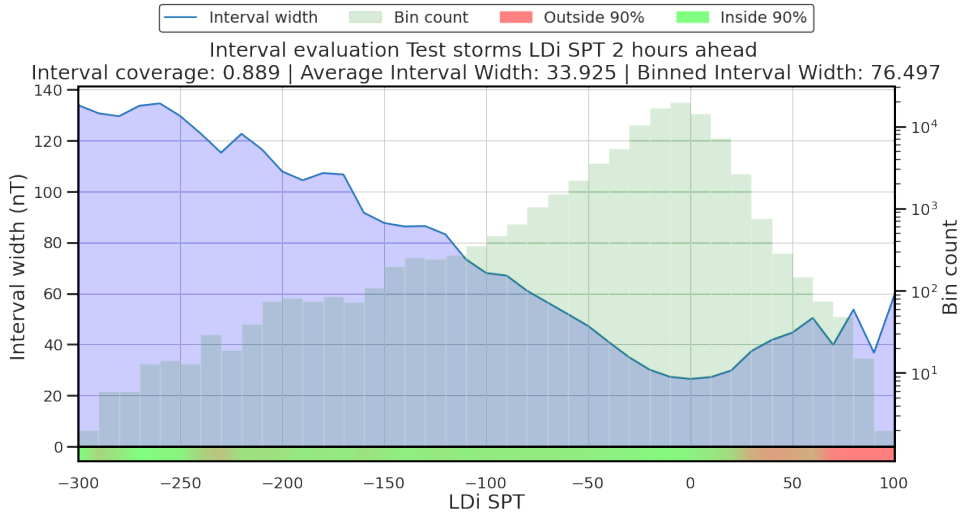


Figure 6.11: Interval Coverage for SPT on the test storms.

the trained model is more accurate in capturing the intensity of geomagnetic disturbances across a range of storm magnitudes.

Additionally, the PICP for the trained model remains high, averaging 0.888, indicating that the model's uncertainty estimates are reliable and the prediction intervals often cover the true values. The PIAW and PIBW metrics show that the trained model produces intervals that balance coverage with the average width, confirming that the prediction intervals are useful for operational SW forecasting.

An interesing example is presented in Figure 6.12; showcasing the forecast for storm 91. This storm was globally super-intense, driven by the successive arrival of two CMEs. In most stations and the overall environment, the first CME caused the most significant disturbance. However, the case of the SPT station is unique: due to its timing and position, the second CME produced a more intense local disturbance than the first, resulting in a lower LDi at that time, contrary to the global trend. Despite this particularity, the model successfully forecasts this complex event, demonstrating its robustness. In this case, the trained model shows a BFE of 32.539, which is significantly better than the

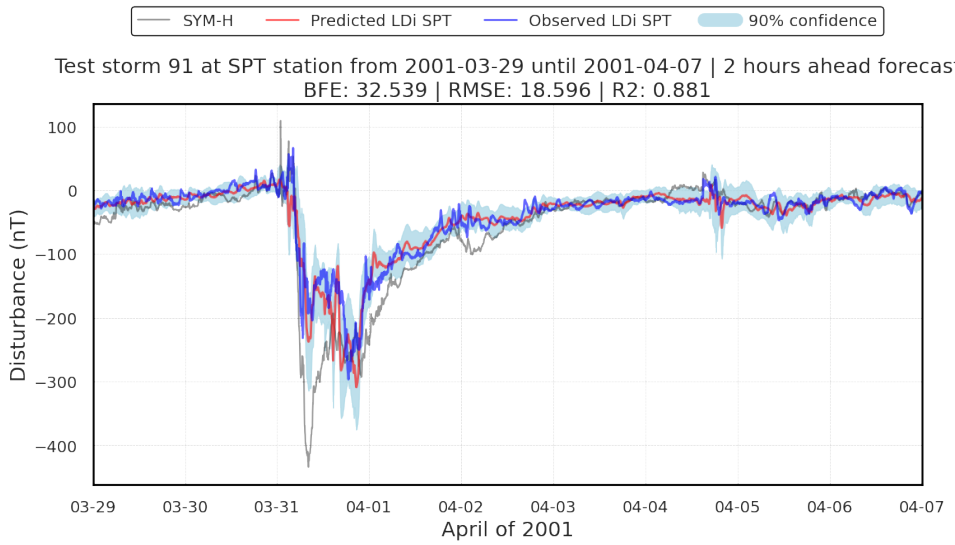


Figure 6.12: 2 hours ahead forecast of the test storm 91 of April 2001 for the LDi of SPT.

persistence model's BFE of 41.796. The RMSE of 18.596 and R^2 of 0.881 further confirm that the trained model can accurately capture the storm's evolution, especially during its most intense phases. The model effectively tracks the observed disturbances, and the 90% prediction interval covers the true values for most of the forecast period. The better performance during Storm 91 highlights the model's capability to handle globally intense geomagnetic events. This is particularly relevant when considering operational SW forecasting, as the model successfully identifies the critical phases of the storm while maintaining accurate forecasts of the LDi.

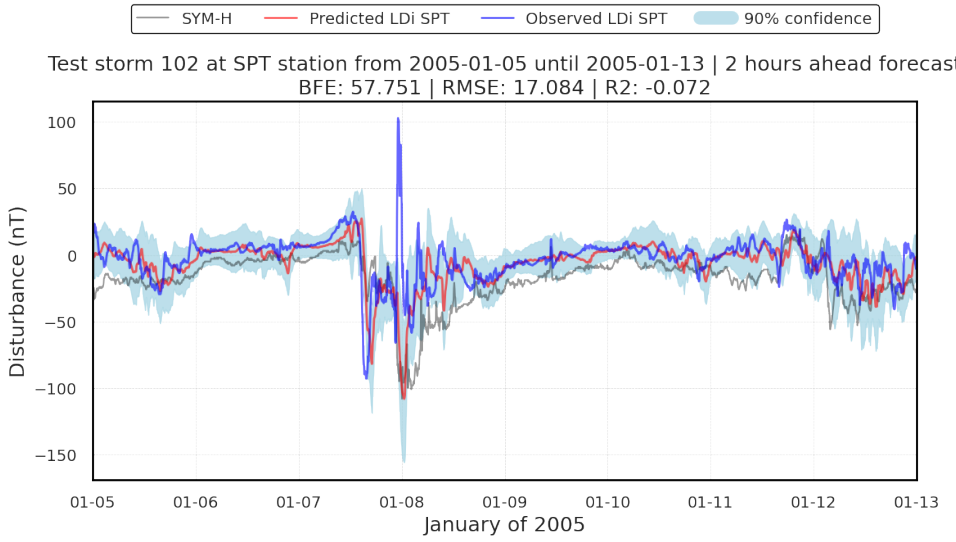


Figure 6.13: 2 hours ahead forecast of the test storm 102 of January 2005 for the LDi of SPT.

However, there are instances where the model struggles to make accurate predictions, as demonstrated by Storm 102. This particular storm, characterized by an atypical increase in LDi during the peak of the geomagnetic storm, poses significant challenges for the model. As shown in Figure 6.13, the forecast exhibits a large error, with a BFE of 57.751, which is comparable to the persistence model (BFE of 58.187). The RMSE of 17.084 and

negative R^2 (-0.072) highlight the model's inability to capture the abnormal behavior of the local disturbance.

Storm 102 is a prime example of why forecasting local geomagnetic indices is inherently difficult. The LDi during this storm behaves in an unusual manner, increasing rather than decreasing at the peak of the storm. This behavior contradicts the typical pattern observed during geomagnetic disturbances, where the local indices usually decrease in sync with the global storm intensity. As a result, this type of outlier significantly increases the forecast error and exposes the limitations of the current model in predicting rare, unexpected events.

Table 6.3 summarizes the evaluation metrics for the test key storm at the SPT observatory, comparing them to the persistence model for the 2 hours ahead forecast. In this case the trained model also outperforms the persistence one, having a BFE lower by around 6, from 28.253 to 22.264.

Table 6.3: Metrics of the trained model compared to the persistence model for the test key storms for the SPT station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
117	23.377	17.079	0.642	17.806	13.253	0.784	0.903	29.218	46.295
118	17.665	10.501	0.477	12.136	7.396	0.741	0.931	24.000	40.055
119	33.206	18.146	0.355	26.189	13.602	0.638	0.775	25.739	45.172
120	24.782	14.452	0.606	15.460	9.981	0.812	0.882	25.923	41.127
121	39.376	15.493	-0.148	27.013	10.703	0.452	0.875	23.217	44.484
122	26.171	12.415	0.658	21.138	9.148	0.814	0.936	23.326	37.963
Mean:	27.429	14.681	0.432	19.957	10.680	0.707	0.884	25.237	42.516
Global:	28.253	14.926	0.560	22.264	10.898	0.765	0.936	23.326	37.963

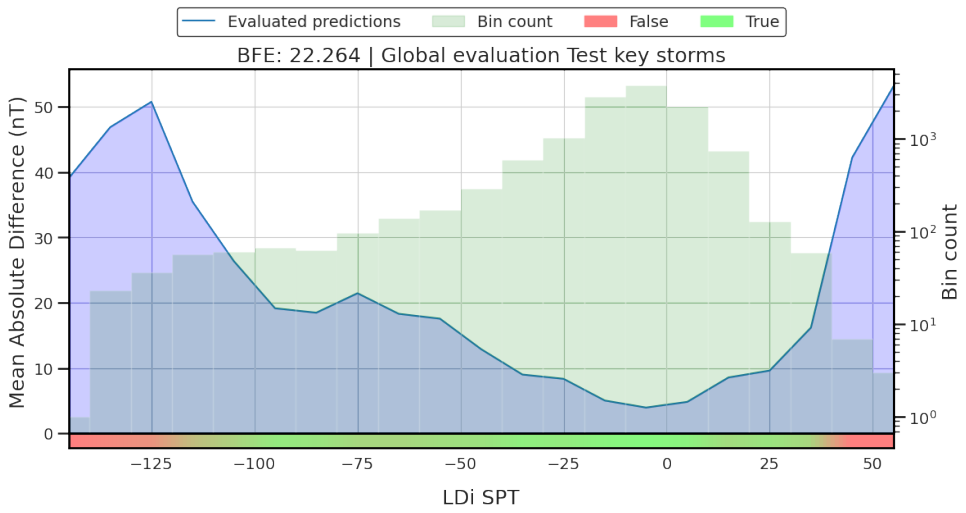


Figure 6.14: Global BFE for SPT on the test key storms.

Figures 6.14 and 6.15 depict the BFE and Interval Coverage of the test key storms for the SPT observatory. Another interesting result is shown in Figure 6.16. In this case, the forecast is considerably close to the observed values, but the interesting thing about the storm is that the local disturbance happened hours earlier than global disturbance (shown in black for the SYM-H). This is another problem of the local forecasting, when

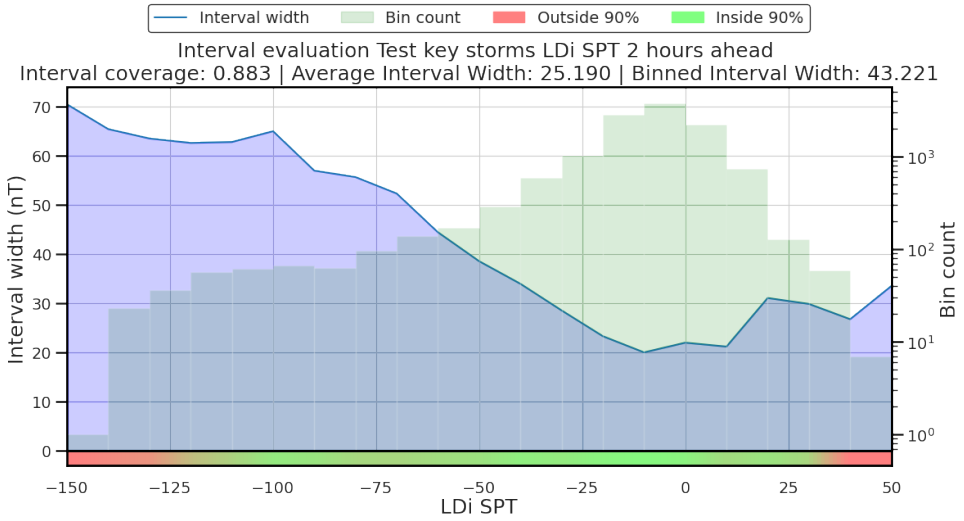


Figure 6.15: Interval Coverage for SPT on the test key storms.

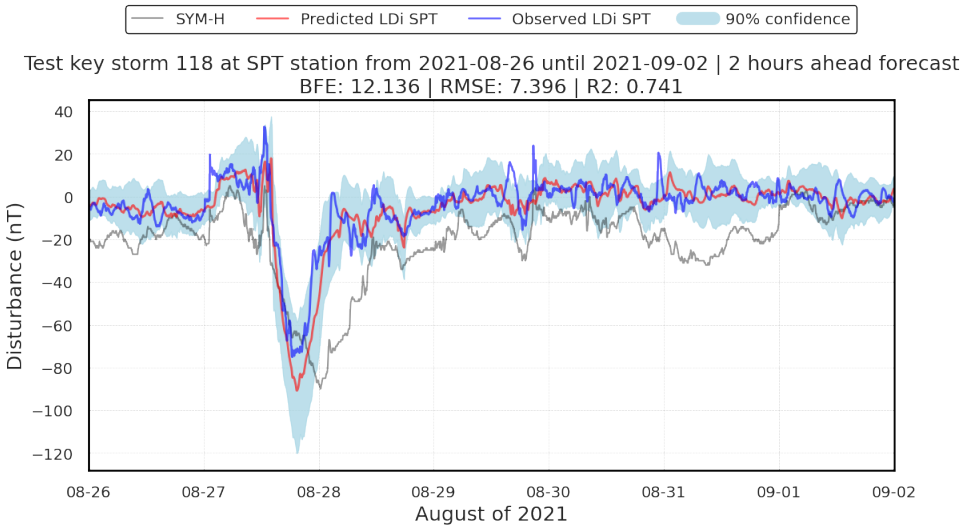


Figure 6.16: 2 hours ahead forecast of the test storm 118 of August 2021 for the LDi of SPT.

depending on the rotation, the disturbances can happen before, or after the global one. This makes the training of the model harder, as the same solar wind and IMF will impact each station at different times.

6.2.5.2 Memambetsu results

Table 6.4 summarizes the evaluation metrics for each test storm at the MMB observatory, comparing the trained model to the persistence one for the 2 hours ahead forecast. Similar to the SPT station, the trained model outperforms the persistence model on average across all metrics, including the BFE, RMSE, and R^2. The global BFE for the persistence model is 54.775, whereas the trained model achieves a significantly lower global BFE of 35.895, highlighting the trained model's ability to better capture the LDi across different storm intensities.

Moreover, the PICP metric for the trained model remains high, reflecting reliable prediction interval coverage. The average interval width (PIAW) also suggests that the

Table 6.4: Metrics of the trained model compared to the persistence model for the test storms for the MMB station.

#	**Persistence model**			**Trained model**					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
82	27.026	20.031	0.634	18.908	15.149	0.790	0.893	41.622	65.108
83	28.266	19.889	0.383	19.908	15.193	0.640	0.878	38.903	53.887
84	31.850	20.930	0.699	21.659	15.053	0.844	0.912	40.618	64.091
85	28.651	14.479	0.556	18.127	9.689	0.801	0.948	30.686	44.806
86	17.215	13.902	0.595	14.408	10.830	0.754	0.916	35.695	39.804
87	20.342	16.010	0.633	16.341	12.048	0.792	0.864	34.383	48.938
88	13.096	11.861	0.709	13.894	12.101	0.697	0.877	33.471	43.253
89	47.556	24.128	0.725	44.174	18.120	0.845	0.892	35.475	68.725
90	54.403	18.436	0.569	31.954	13.903	0.755	0.899	34.065	75.152
91	68.473	32.063	0.751	40.476	19.948	0.903	0.858	42.235	93.636
92	22.360	12.866	0.425	20.890	11.043	0.577	0.847	29.949	38.810
94	10.360	8.715	0.846	9.269	7.350	0.890	0.970	29.337	38.320
95	16.124	15.970	0.705	12.280	11.596	0.844	0.902	35.008	46.607
96	65.590	20.588	0.110	46.950	20.101	0.151	0.875	45.064	85.984
97	25.511	18.774	0.388	12.699	12.807	0.715	0.882	37.985	51.353
98	24.745	17.653	0.474	15.119	13.979	0.670	0.870	37.496	52.296
99	46.392	21.753	0.296	19.495	15.143	0.659	0.885	42.975	64.511
100	21.527	15.203	0.428	18.414	11.453	0.675	0.902	33.973	44.832
101	53.390	34.098	0.763	32.012	22.415	0.898	0.853	45.238	92.075
102	27.456	12.890	0.131	20.902	11.927	0.256	0.942	35.628	51.423
103	16.392	14.884	0.572	12.989	11.548	0.743	0.905	34.344	44.310
104	14.918	11.269	0.633	18.977	12.413	0.555	0.886	27.331	46.078
105	16.703	12.752	0.146	12.131	9.573	0.519	0.936	31.396	38.426
106	19.188	11.862	0.358	16.315	8.716	0.653	0.944	31.387	49.256
107	24.444	12.680	0.573	16.373	9.206	0.775	0.948	26.064	38.659
108	19.568	11.262	0.780	18.934	10.456	0.810	0.911	19.616	30.215
109	23.144	15.190	0.783	19.331	12.872	0.844	0.917	34.350	50.104
110	17.521	13.823	0.721	13.121	9.417	0.870	0.921	29.917	40.761
111	20.527	13.631	0.668	10.778	7.733	0.893	0.952	29.012	50.129
112	12.905	9.998	0.776	9.156	8.958	0.820	0.955	28.104	45.707
113	23.875	11.325	0.204	24.214	9.527	0.437	0.950	28.226	40.522
114	29.700	22.461	0.661	22.388	16.807	0.810	0.809	39.873	55.328
115	23.475	15.305	0.436	13.627	11.368	0.689	0.852	28.690	38.967
116	24.019	17.768	0.350	19.806	12.173	0.695	0.911	36.849	50.199
Mean:	27.550	16.601	0.544	19.883	12.665	0.714	0.902	34.264	52.420
Global:	54.775	18.038	0.702	35.895	13.445	0.834	0.911	36.849	50.199

model provides relatively narrow prediction intervals, balancing precision with robustness in the face of uncertainty.

Figure 6.17 presents the global BFE metric across all test storms for the MMB observatory. The global BFE is calculated by averaging the BFE values across all bins of intensity for each storm, allowing us to assess the model's performance in capturing geomagnetic disturbances across different magnitudes.

Figure 6.18 shows the prediction interval coverage for the MMB observatory across all test storms. It provides a visual representation of how well the prediction intervals cover the true geomagnetic disturbance values, helping us evaluate the reliability and usefulness of the uncertainty estimates provided by the model.

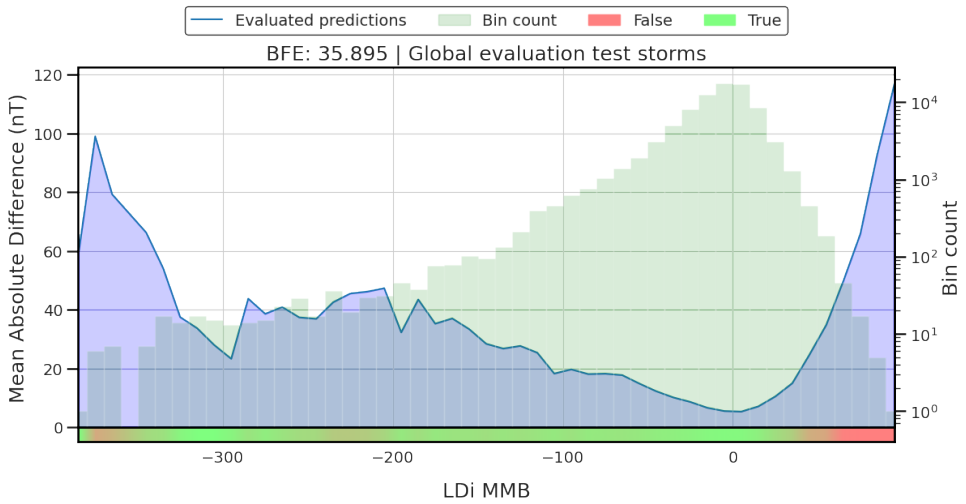


Figure 6.17: Global BFE for MMB on the test storms.

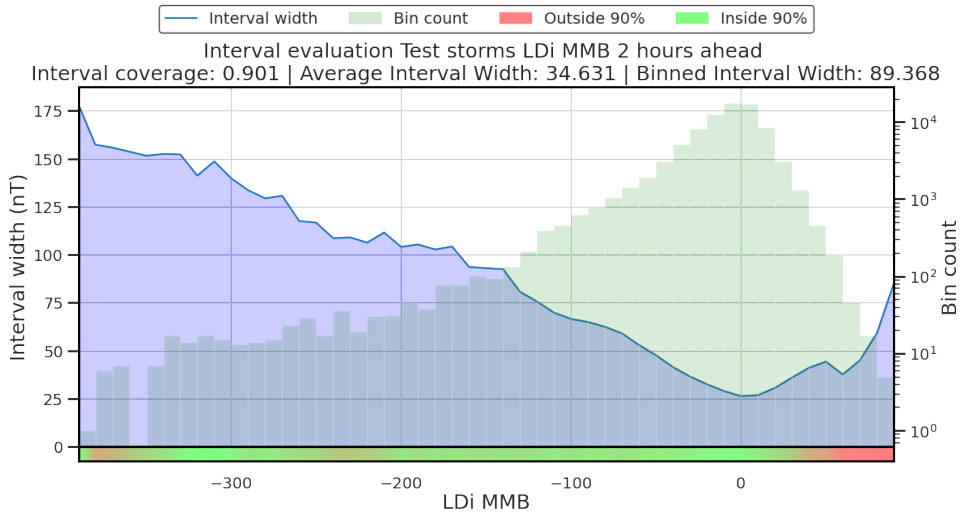


Figure 6.18: Interval Coverage for MMB on the test storms

For comparison, Figure 6.19 depicts the forecast for the Storm 91 at the MMB observatory shows that the behavior of the LDi is much more closely aligned with the global SYM-H index compared to the forecast at SPT. Both the predicted and observed LDi follow the general trend of SYM-H, indicating that the MMB station experienced a geomagnetic disturbance pattern that was more consistent with global averages. This is reflected in the relatively high R^2 of 0.903 and a reasonable BFE of 40.476. Although the RMSE (19.948) is slightly higher compared to SPT, the model successfully tracks the key phases of the storm, particularly during its peak intensity.

Another example, storm 101, a superintense geomagnetic event. In this case, the persistence model obtains a BFE of 53.390, whereas the trained model achieves a BFE of 32.012, a significant improvement. The RMSE for the trained model is 22.415 with an R^2 of 0.898, as seen in Figure 6.20. This result demonstrates that the trained model is capable of handling superintense storms more effectively, especially when the LDi behavior is aligned with the global storm patterns.

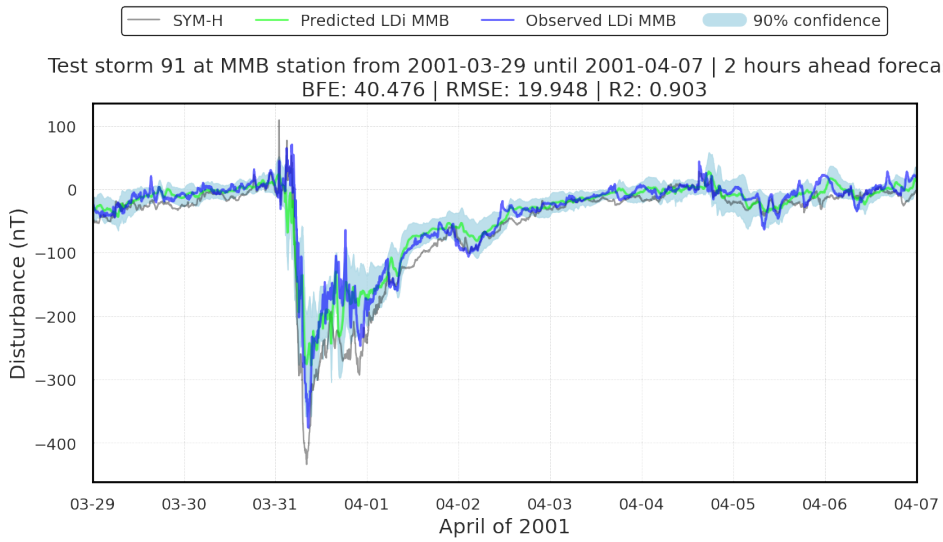


Figure 6.19: 2 hours ahead forecast of the test storm 91 of April 2001 for the LDi of MMB.

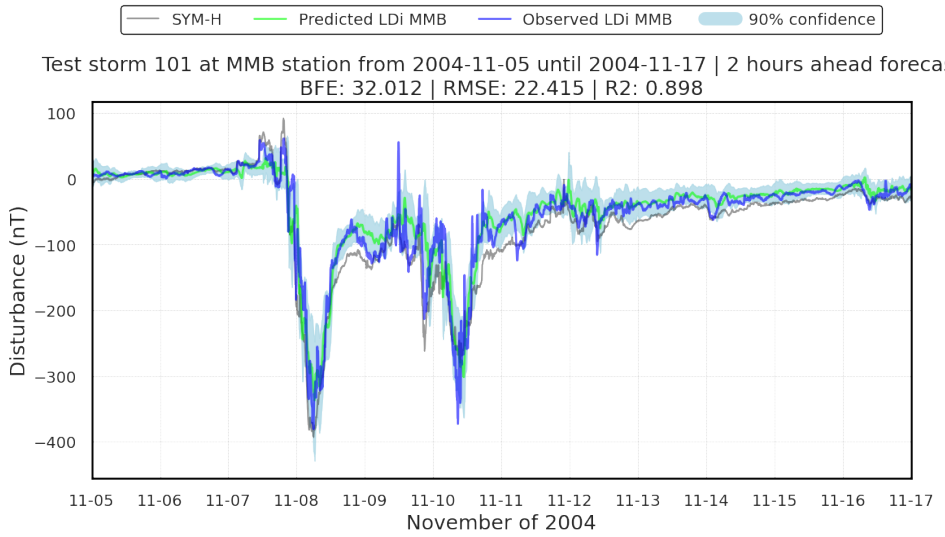


Figure 6.20: 2 hours ahead forecast of the test storm 101 of November 2004 for the LDi of MMB.

Overall, the trained model demonstrates superior performance compared to the persistence model at the MMB observatory, especially during globally intense storms. The high accuracy in predicting the LDi during these events reinforces the model's potential for operational SW forecasting.

Table 6.5 summarizes the evaluation metrics for the test key storm at the MMB observatory, comparing them to the persistence model for the 2 hours ahead forecast. This is an interesting case, despite the trained model having better performance on the individual storms, and better mean and global metrics in almost all of them; it has a higher global BFE, this is caused by the forecast of storm 117, where the model forecasted a disturbance lower than the one that actually happened. This can also be seen in the left side of the BFE plot of the test key storms, as shown in Figure 6.21.

Table 6.5: Metrics of the trained model compared to the persistence model for the test key storms for the MMB station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
117	28.348	16.320	0.776	25.422	12.734	0.864	0.869	29.486	53.480
118	12.063	8.203	0.568	8.485	6.580	0.722	0.959	25.234	31.903
119	31.053	17.356	0.614	28.495	13.838	0.755	0.834	27.443	48.108
120	9.361	8.147	0.673	8.526	7.296	0.738	0.940	26.423	30.882
121	21.422	10.461	0.198	17.122	10.086	0.255	0.885	23.725	42.549
122	12.845	11.167	0.677	15.393	10.947	0.689	0.895	25.265	31.145
Mean:	19.182	11.942	0.584	17.240	10.247	0.670	0.897	26.262	39.678
Global:	26.974	12.443	0.704	28.907	10.570	0.787	0.895	25.265	31.145

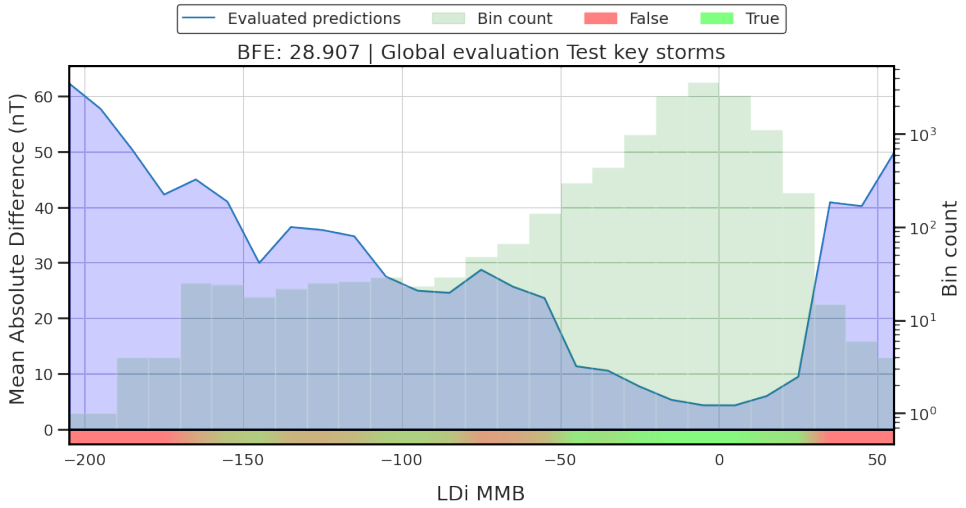


Figure 6.21: Global BFE for MMB on the test key storms.

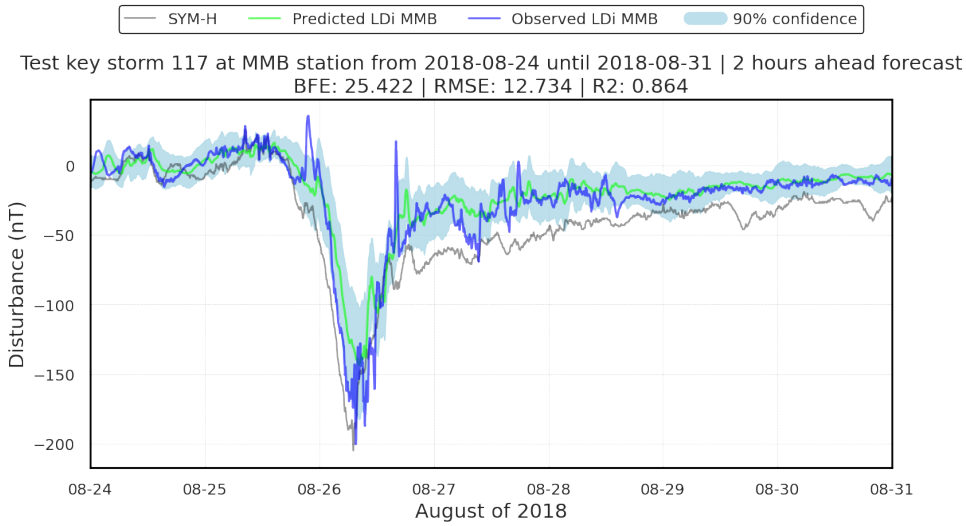


Figure 6.22: 2 hours ahead forecast of the test key storm 117 of August 2018 for the LDi of MMB.

6.2.5.3 Tucson results

The TUC observatory offers another example of the model's performance when forecasting local geomagnetic disturbances. As shown in Table 6.6, the trained model consistently outperforms the persistence model across all metrics. The global BFE for the persistence model is 63.051, whereas the trained model achieves a much lower global BFE of 37.076, reinforcing the effectiveness of the model for this observatory. The trained model's PICP metric also shows robust coverage for the prediction intervals, ensuring reliability in capturing uncertainty.

Table 6.6: Metrics of the trained model compared to the persistence model for the test storms for the TUC station.

#	**Persistence model**			**Trained model**					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
81	44.418	19.556	0.096	19.279	13.101	0.594	0.916	32.916	54.901
82	35.760	23.059	0.640	35.999	18.788	0.761	0.838	42.286	57.288
83	45.634	17.951	0.474	29.521	14.747	0.645	0.859	38.923	63.876
85	27.917	14.233	0.687	14.387	8.948	0.876	0.938	30.508	57.898
87	40.378	19.133	0.477	22.563	12.076	0.792	0.883	33.184	54.561
88	24.820	17.654	0.624	18.193	14.652	0.741	0.834	33.614	54.941
89	43.259	23.793	0.771	29.899	15.663	0.901	0.829	36.321	84.407
90	64.624	20.210	0.501	38.005	14.129	0.756	0.872	33.765	83.686
91	64.426	33.550	0.741	32.806	18.265	0.923	0.860	42.316	90.971
92	30.297	17.533	0.429	21.747	14.154	0.628	0.812	29.471	65.040
93	24.415	19.366	0.387	25.044	18.399	0.446	0.670	33.435	44.259
94	14.728	11.497	0.648	11.259	8.879	0.790	0.950	28.271	41.909
95	32.114	21.172	0.597	24.429	15.864	0.774	0.852	34.716	54.953
96	71.958	25.443	0.043	43.563	20.744	0.364	0.854	45.555	92.156
97	37.335	21.227	0.115	24.834	15.311	0.540	0.833	38.199	47.650
99	23.268	17.797	0.420	17.011	13.524	0.665	0.889	43.163	50.112
100	15.436	13.471	0.517	11.998	9.835	0.742	0.926	33.409	37.224
101	63.561	38.450	0.654	37.445	27.853	0.818	0.848	45.838	98.384
102	61.838	22.248	0.277	39.601	15.310	0.658	0.951	35.561	68.776
103	19.501	16.236	0.575	16.942	11.176	0.799	0.897	34.191	48.804
104	17.225	11.767	0.489	17.972	10.445	0.597	0.893	26.735	38.796
105	22.354	15.865	0.467	14.854	11.642	0.713	0.857	31.436	53.883
106	24.616	12.801	0.444	18.460	10.009	0.660	0.931	30.954	50.280
107	35.082	16.849	0.470	23.855	10.600	0.790	0.886	26.497	51.520
108	35.177	18.348	0.601	25.306	13.443	0.786	0.852	18.794	36.665
109	19.846	14.743	0.791	16.802	12.456	0.851	0.898	33.989	55.184
110	28.977	15.992	0.578	20.213	11.620	0.777	0.862	29.720	49.148
111	36.849	18.075	0.432	17.390	10.012	0.826	0.877	29.233	53.508
112	18.255	12.835	0.676	12.036	9.504	0.822	0.928	28.047	48.190
113	43.469	17.105	-0.013	21.433	10.132	0.644	0.918	27.505	72.121
114	31.515	24.080	0.719	28.160	17.570	0.851	0.812	40.086	61.865
115	27.346	15.092	0.480	18.079	12.330	0.653	0.779	28.341	41.438
116	25.428	19.761	0.360	20.125	13.922	0.682	0.857	36.728	49.232
Mean:	34.904	18.997	0.490	23.309	13.791	0.723	0.869	33.749	57.989
Global:	63.051	20.386	0.652	37.076	14.724	0.818	0.857	36.728	49.232

For some interesting results, Figure 6.25, the forecast for Storm 91 at the TUC station shows that the LDi is lower than the global SYM-H index. This deviation can be explained by the local magnetic conditions and the timing of the shock relative to the MLT of the station. In this case, the storm reached the Earth later in the day, affecting the

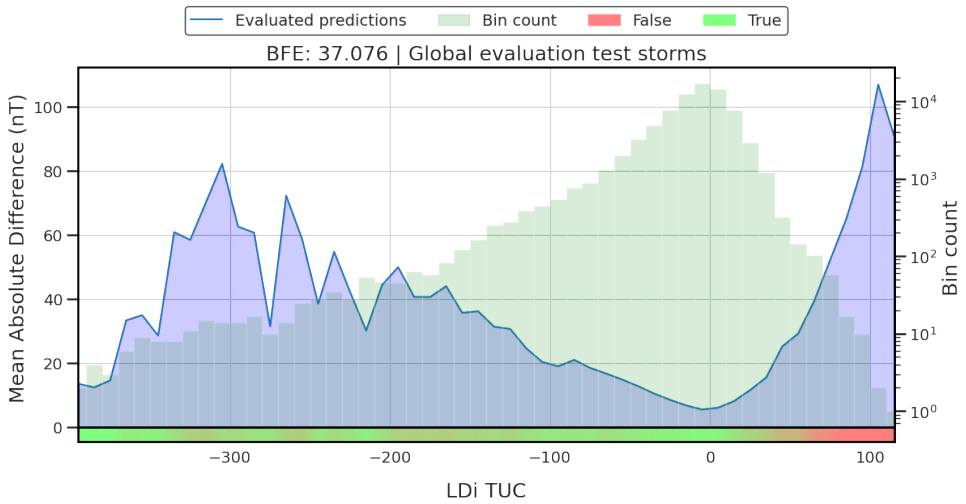


Figure 6.23: Global BFE for TUC on the test storms.

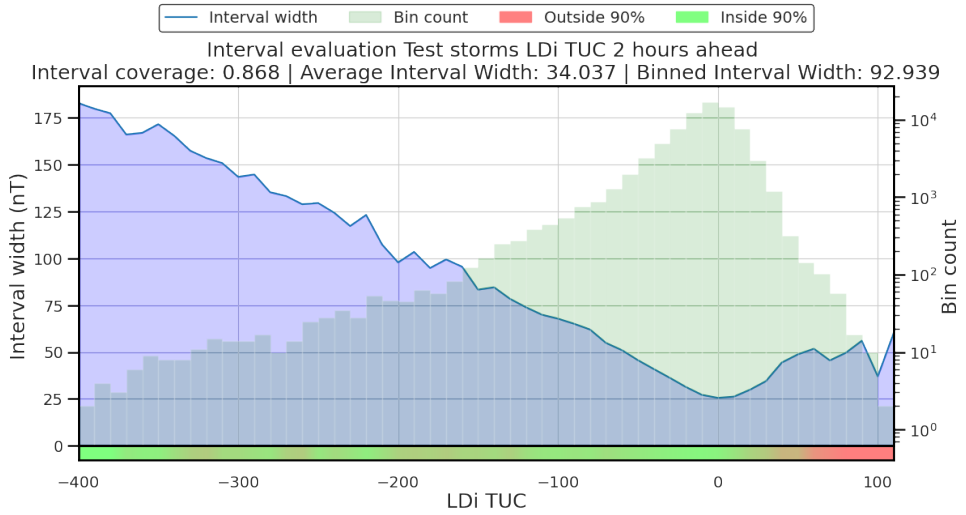


Figure 6.24: Interval Coverage for TUC on the test storms.

magnitude of the disturbance. The model achieves a BFE of 32.806 and a high R^2 of 0.923, demonstrating its ability to accurately capture the local geomagnetic activity. Although the RMSE is slightly higher at 18.265, the model still provides reliable forecasts during intense storm conditions.

Another example is the storm 101, the other superintense test geomagnetic storm. It presents a different behavior compared to Storm 91. As shown in Figure 6.26, the LDi and SYM-H behave much more similarly during this storm, with the local disturbance being almost even with the global index. This difference is due to the timing of the shock relative to the MLT, where the storm impacted closer to dawn at the TUC station. The trained model performs significantly better than the persistence model for this event, with a BFE of 37.445 compared to the persistence's 63.561. Although the RMSE is higher at 27.853, the model still captures key storm features, as indicated by an R^2 of 0.818.

The impact of MLT on the disturbance level at TUC highlights the importance of local factors in predicting geomagnetic indices, as even globally intense storms can have varying effects at different geographic locations.

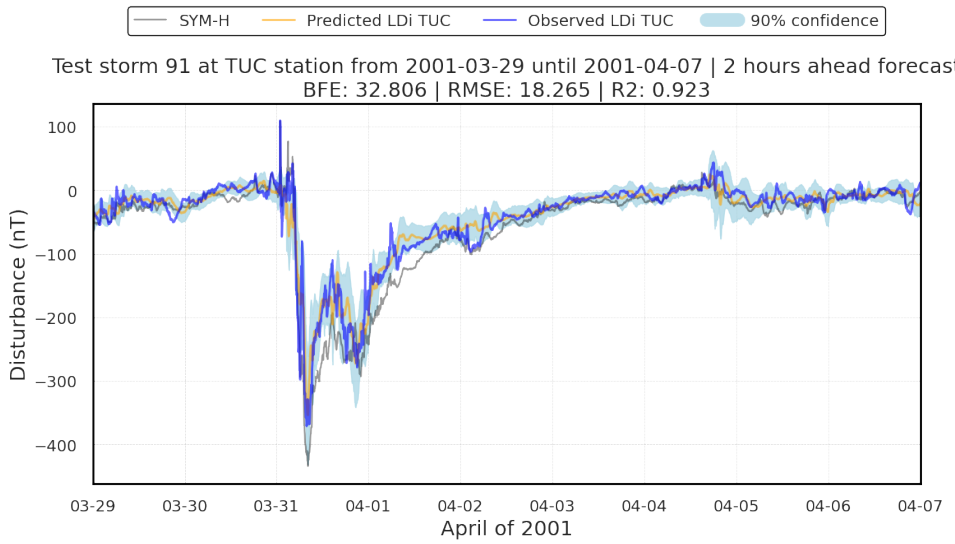


Figure 6.25: 2 hours ahead forecast of the test Storm 91 of April 2001 of the LDi of TUC.

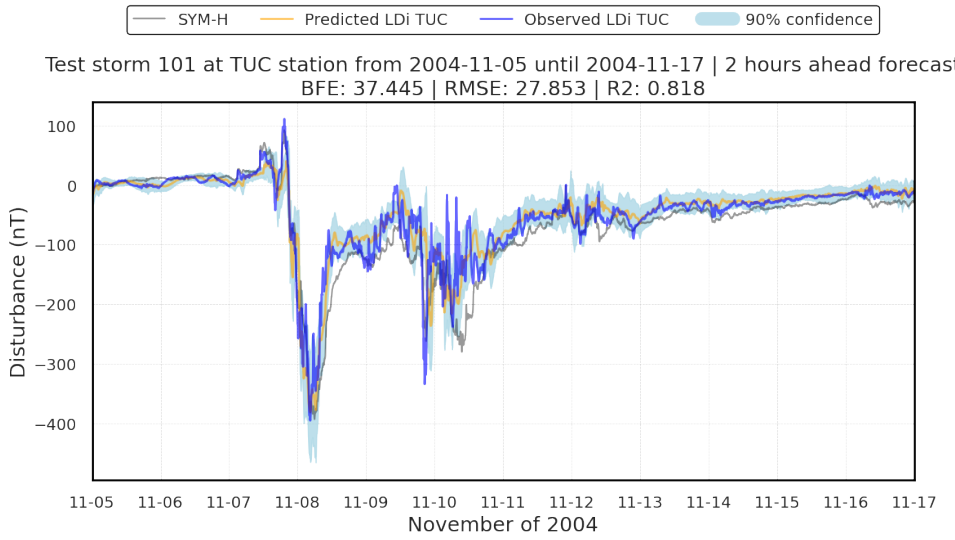


Figure 6.26: 2 hours ahead forecast of the test Storm 101 of November 2004 for the LDi of TUC.

Table 6.7 presents the performance of the model compared to the persistence model for key storms at the TUC station. One particularly interesting case is storm number 119, shown in Figure 6.28, which shows a very high BFE of 48.277. This large error is primarily caused by abnormal behavior, with sudden large positive and negative peaks that are difficult for the model to capture, leading to higher errors. Despite this, the model still outperforms the persistence model. On the other hand, for storm number 121, shown in Figure 6.29, which also presents significant challenges for the persistence model, the trained model demonstrates a much better performance. The predicted peak is nearly perfect, which significantly reduces the BFE to 23.773 compared to the persistence model's BFE of 48.190. This showcases the model's ability to handle complex dynamics in the local indices LDi and reflects its improved generalization for storms of varying intensities and behaviors.

Table 6.7: Metrics of the trained model compared to the persistence model for the test key storms for the TUC station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
117	26.855	16.855	0.747	20.820	12.974	0.850	0.845	29.444	53.191
118	16.063	10.160	0.581	11.616	8.173	0.729	0.864	24.400	36.468
119	52.821	19.457	0.130	48.277	16.024	0.410	0.765	27.347	40.139
120	17.741	11.483	0.697	11.007	8.761	0.823	0.937	25.938	37.088
121	48.190	17.520	0.233	23.773	12.921	0.583	0.873	23.437	44.646
122	12.068	11.137	0.666	11.750	9.770	0.743	0.806	24.289	31.365
Mean:	28.957	14.435	0.509	21.207	11.437	0.690	0.848	25.809	40.483
Global:	34.540	14.948	0.611	27.860	11.797	0.757	0.806	24.289	31.365

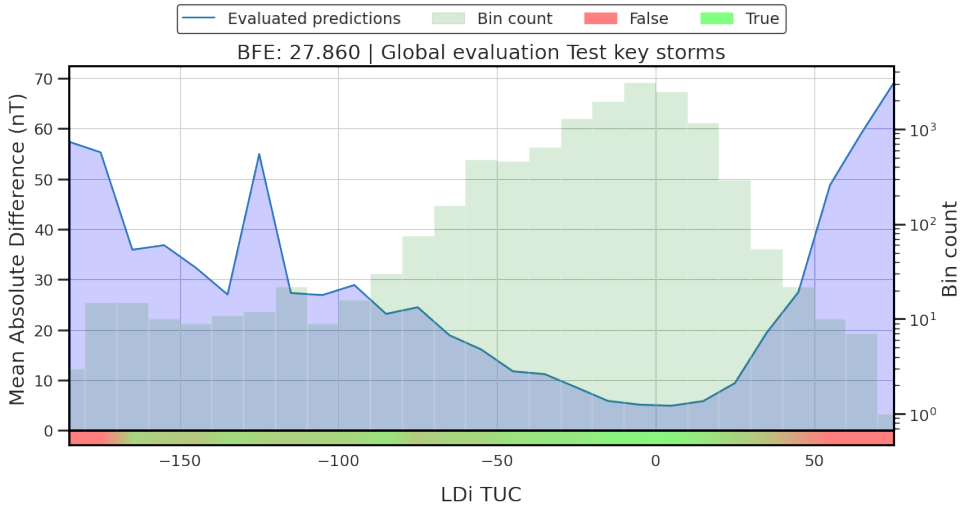


Figure 6.27: Global BFE for TUC on the test key storms.

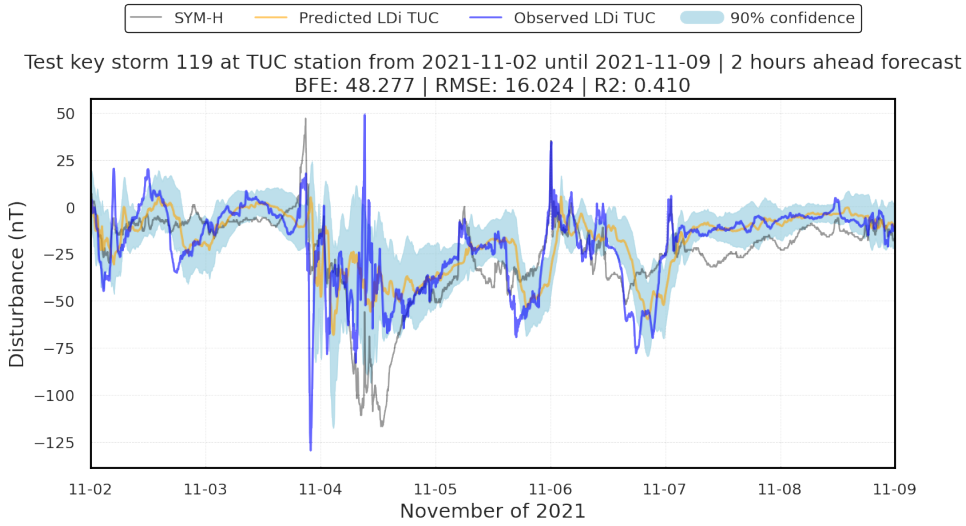


Figure 6.28: 2 hours ahead forecast for the test key storm 119 of March 2022 for the LDi at TUC.

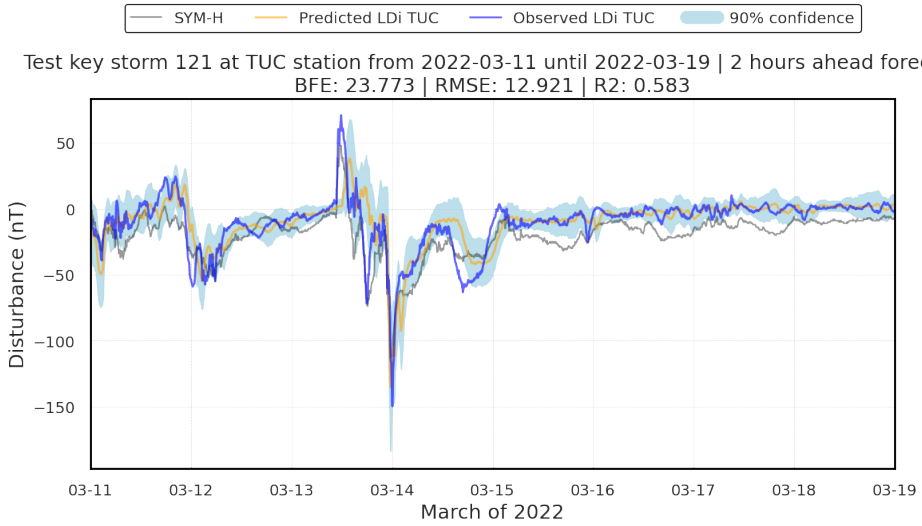


Figure 6.29: 2 hours ahead forecast for the test key storm 121 of March 2022 for the LDi at TUC.

6.2.5.4 Alibag results

Table 6.8 shows the comparison of the metrics between the trained model and the persistence one. Once again, the trained model clearly improves the persistence model. Similar to other stations, the BFE plot shown in Figure 6.30 shows that the model has issues forecasting the extreme values while it works reasonably well on the more common ones. The same is observed in the prediction interval evaluation in Figure 6.31 where the margin is considerably wider in the left side.

Table 6.8: Metrics of the trained model compared to the persistence model for the test storms for the ABG station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
81	20.385	12.305	0.512	13.002	11.383	0.582	0.928	32.250	45.446
82	28.325	21.957	0.801	18.226	16.230	0.891	0.861	42.742	66.420
83	23.594	18.554	0.704	17.455	13.435	0.845	0.903	37.620	57.574
84	23.656	21.198	0.791	14.870	15.105	0.894	0.934	41.181	62.134
85	27.854	17.892	0.653	11.785	8.981	0.913	0.938	29.477	49.814
86	16.053	17.741	0.633	11.785	12.658	0.813	0.855	35.563	42.901
87	20.209	15.034	0.691	16.288	11.132	0.830	0.898	33.622	48.179
88	14.136	12.363	0.792	13.033	11.346	0.825	0.908	32.888	38.872
89	41.302	24.003	0.827	26.555	13.197	0.948	0.892	36.396	73.330
90	20.093	16.022	0.768	17.686	12.577	0.857	0.879	32.708	53.332
91	48.889	34.911	0.813	40.036	25.193	0.902	0.755	43.527	84.195
92	29.016	13.677	0.697	24.090	11.381	0.790	0.905	29.679	53.996
94	14.848	13.304	0.841	11.950	10.384	0.903	0.850	29.088	38.885
95	33.234	20.411	0.731	17.354	14.706	0.860	0.833	35.734	54.089
96	25.907	18.817	0.391	19.538	19.010	0.378	0.868	43.724	69.809
99	50.833	25.914	0.388	37.339	19.771	0.644	0.866	43.921	60.259
100	26.952	16.148	0.675	13.128	11.288	0.841	0.857	35.041	51.317
101	46.001	32.540	0.841	28.543	20.679	0.936	0.884	45.101	83.713
102	30.581	15.943	0.165	27.661	15.002	0.261	0.878	34.588	48.781
104	15.895	11.183	0.810	15.002	9.695	0.857	0.897	27.195	45.423
105	18.229	13.614	0.502	14.329	12.280	0.595	0.907	30.793	40.747
106	16.152	14.459	0.306	11.016	10.169	0.657	0.909	29.927	36.858
107	16.765	11.392	0.795	14.896	8.760	0.879	0.927	25.509	33.836
108	21.530	11.428	0.846	20.740	10.738	0.864	0.868	19.777	28.366
109	21.891	16.365	0.876	18.066	11.969	0.934	0.938	34.666	49.896
110	19.116	14.156	0.771	12.067	9.594	0.895	0.880	29.510	43.595
111	24.561	14.705	0.682	17.108	9.539	0.866	0.944	27.992	46.278
112	14.001	10.922	0.856	11.251	9.214	0.898	0.949	28.604	45.670
113	22.191	13.252	0.561	26.346	14.056	0.506	0.878	28.021	42.742
114	37.510	26.901	0.782	21.979	18.208	0.900	0.808	41.794	69.807
115	14.122	12.157	0.753	9.932	8.957	0.866	0.913	28.160	37.644
116	18.813	19.048	0.586	13.118	13.710	0.785	0.879	36.639	48.653
Mean:	25.083	17.447	0.682	18.318	13.136	0.794	0.887	33.857	51.643
Global:	45.441	18.841	0.804	32.189	13.906	0.893	0.879	36.639	48.653

However, the ABG station is very particular; it has the lowest latitude of all the evaluated stations, this causes the LDi to have large deviations when compared to the SYM-H. One example is shown in Figure 6.32, for the storm of November 1998. In this case for some disturbances the difference between the SYM-H and the LDi is close to 100 nT, highlighting the problem of the local indices, where they can greatly differ from the

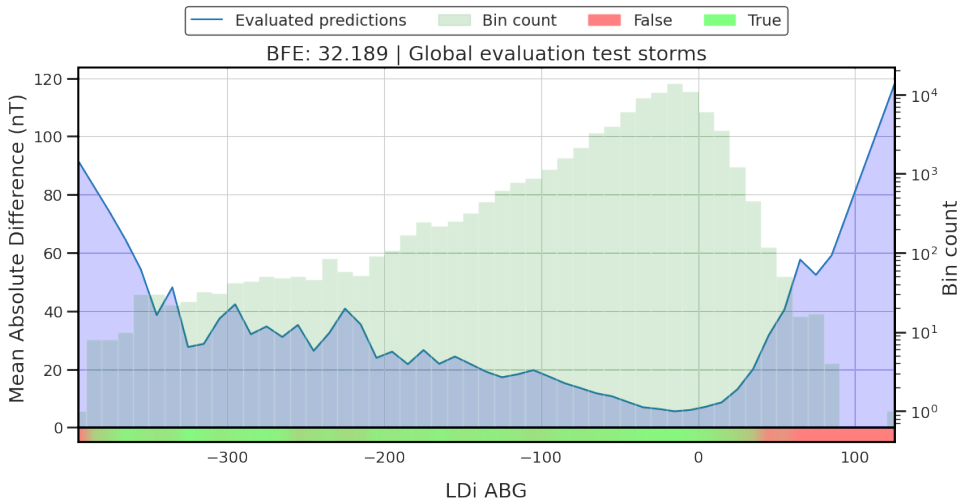


Figure 6.30: Global BFE for ABG on the test storms.

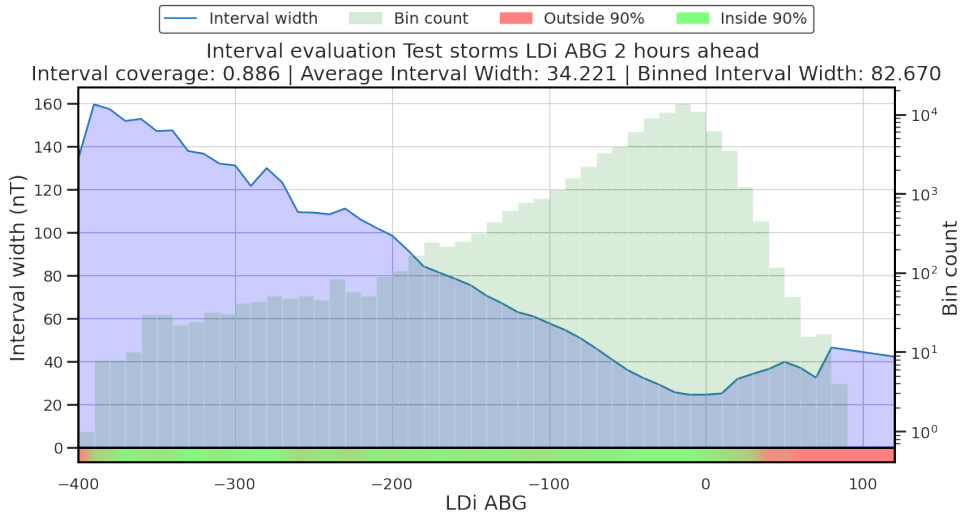


Figure 6.31: Interval Coverage for ABG on the test storms.

global average. Another noteworthy example is shown in Figure 6.33. In this case, once again, while the disturbance can barely be considered low from a global standpoint, it reaches more than -200 nT in the LDi, having more than 100 nT of difference.

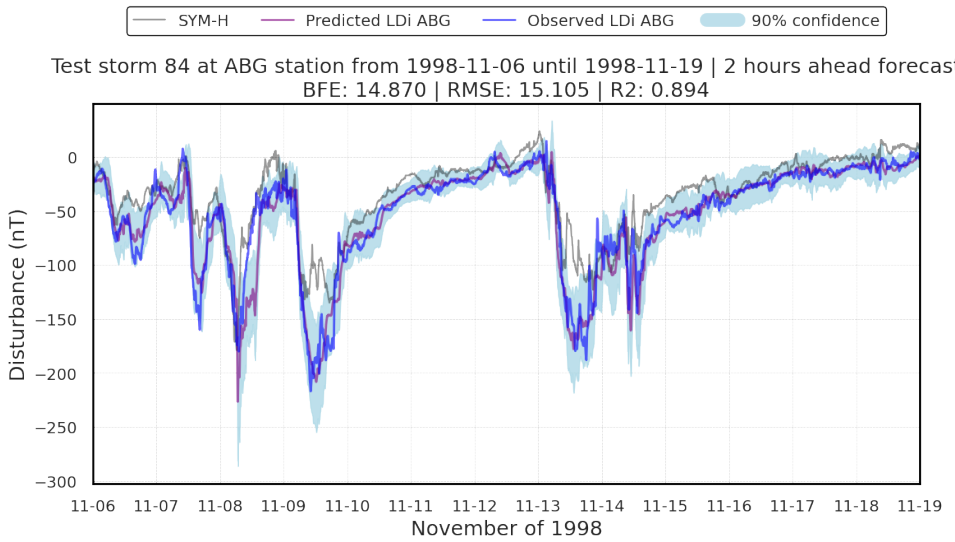


Figure 6.32: 2 hours ahead forecast of the test storm 84 of November 1998 for the LDi of ABG.

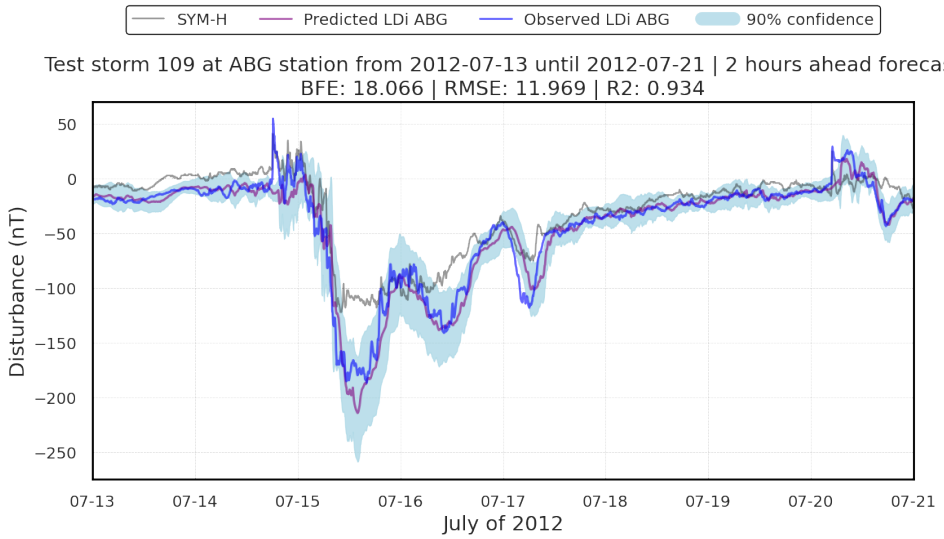


Figure 6.33: 2 hours ahead forecast of the test storm 109 of July 2012 for the LDi of ABG.

Table 6.9 presents the metrics of the test key storms for the ABG station, along with the plot of the BFE for all the test key storms in Figure 6.34. Similar to previous stations, the trained model performs better than a persistence model. From the storms we can highlight the test key storm 118 of August 2021, shown in Figure 6.35. In this case we can notice, once again, a misalignment of the SYM-H peak and the LDi peak, the LDi peak happens around 8 hours earlier than the SYM-H one, and during the SYM-H peak, the LDi increases instead of decreasing.

Table 6.9: Metrics of the trained model compared to the persistence model for the test key storms for the ABG station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
117	28.726	18.749	0.813	22.079	13.915	0.897	0.871	30.651	53.894
118	12.810	10.599	0.622	11.657	8.837	0.738	0.908	24.817	35.862
119	34.302	18.854	0.653	25.988	14.453	0.796	0.728	26.542	47.410
121	30.235	16.529	0.465	25.003	12.301	0.704	0.714	23.231	35.035
122	24.146	17.522	0.692	25.753	14.318	0.795	0.837	25.318	38.293
Mean:	26.044	16.451	0.649	22.096	12.765	0.786	0.812	26.112	42.099
Global:	28.722	16.725	0.737	24.836	12.920	0.843	0.837	25.318	38.293

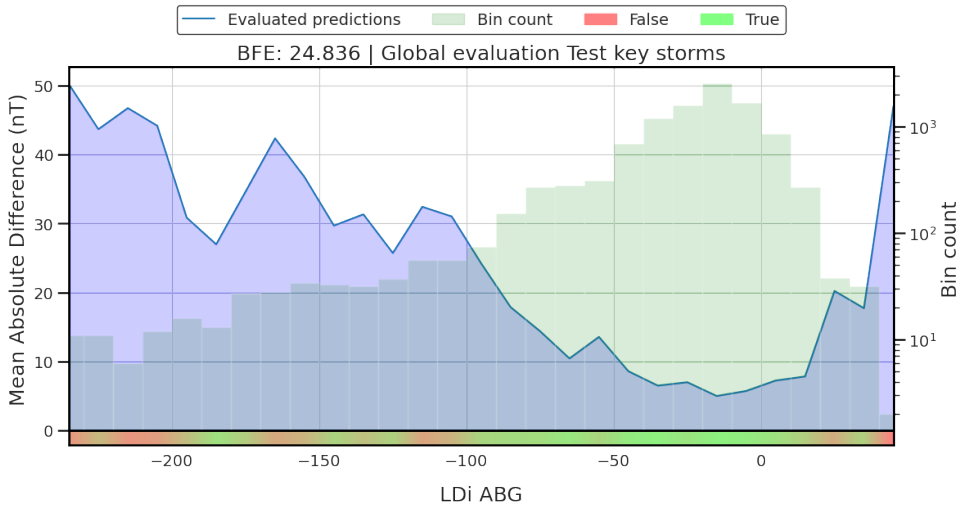


Figure 6.34: Global BFE for ABG on the test key storms.

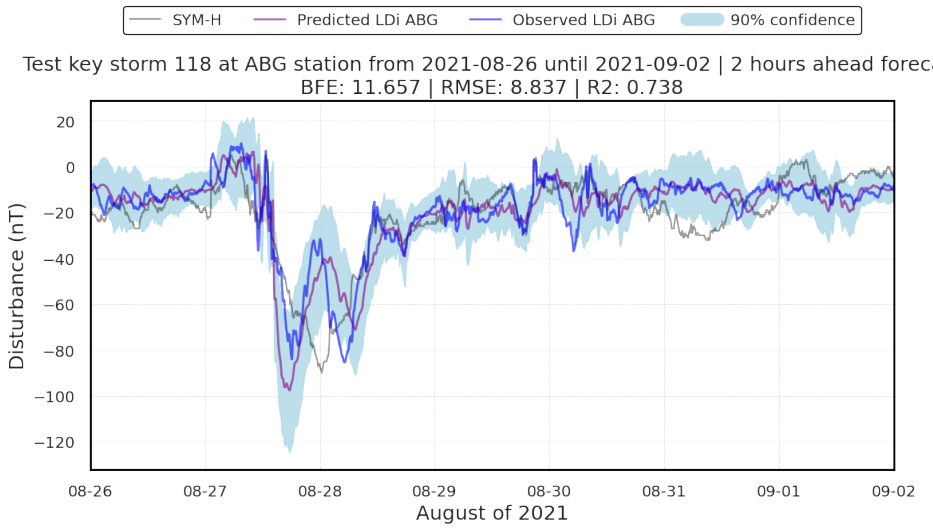


Figure 6.35: 2 hours ahead forecast of the test key storm 118 of August 2021 for the LDi of ABG.

6.2.5.5 Honolulu results

Table 6.10 shows the metrics comparison for the HON station, in this case the improvement compared to the persistence model is notable, as the BFE is reduced from 61.922 from the persistence model to 35.539 in the global computation of the test storms. Nevertheless, there are some key peaks in the most extreme values of the LDi as shown in the global plot of the BFE, as shown in Figure 6.36. This is caused by the errors when forecasting the superintense storms 91 and 101, shown in Figures 6.38 and 6.39, respectively. This station, while its latitude is not as low as ABG's, it is still fairly lower than the other stations, as HON is located at around 20º, while the others are around 40º. This causes the considerable difference around the peak for the storm 101, where the SYM-H does not even reach -400 nT, while the LDi surpasses the -450 mark.

Table 6.10: Metrics of the trained model compared to the persistence model for the test storms for the HON station.

#	**Persistence model**			**Trained model**					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
81	28.285	15.161	0.563	10.394	9.249	0.837	0.970	31.741	50.464
82	18.969	16.894	0.813	17.942	14.284	0.866	0.865	41.414	55.440
83	28.104	15.249	0.685	14.124	11.797	0.811	0.892	38.293	56.802
84	22.869	17.430	0.748	21.052	13.154	0.856	0.954	40.373	59.642
85	14.247	9.347	0.858	10.119	7.504	0.909	0.976	29.716	43.232
86	13.628	13.077	0.784	8.367	9.128	0.895	0.944	35.772	42.930
88	18.353	13.005	0.805	12.612	11.968	0.835	0.925	32.763	49.544
90	49.297	18.921	0.691	23.028	11.599	0.884	0.930	33.361	76.242
91	72.078	37.443	0.770	41.454	21.565	0.924	0.853	43.297	95.532
92	19.252	10.400	0.750	17.831	10.630	0.739	0.914	28.585	37.105
93	14.969	13.523	0.703	11.795	12.625	0.741	0.857	32.024	41.522
94	14.870	10.589	0.829	11.137	7.808	0.907	0.962	28.897	44.747
95	21.666	18.293	0.796	15.259	12.215	0.909	0.868	35.570	48.197
96	36.299	17.848	0.586	21.779	15.184	0.701	0.926	44.433	86.692
97	26.190	16.321	0.729	13.435	10.520	0.887	0.929	37.934	56.032
98	26.494	15.174	0.677	13.036	11.341	0.819	0.928	36.799	54.944
99	17.560	14.966	0.575	11.624	12.076	0.723	0.936	42.419	49.636
100	8.611	9.354	0.752	7.853	8.250	0.807	0.949	32.383	39.170
101	52.815	31.757	0.842	30.887	19.675	0.940	0.890	46.641	104.175
102	29.405	15.296	0.454	19.500	11.301	0.702	0.957	34.398	54.421
103	15.542	12.816	0.795	11.216	7.557	0.929	0.967	33.233	49.148
104	14.459	11.272	0.777	15.183	11.068	0.785	0.918	26.245	46.522
105	13.884	10.291	0.739	11.512	7.431	0.864	0.976	30.585	47.330
106	26.237	12.546	0.549	15.079	8.493	0.793	0.941	29.776	53.120
107	21.613	13.362	0.736	14.730	7.230	0.923	0.961	25.467	42.681
108	23.583	14.113	0.783	16.677	10.936	0.870	0.882	19.202	33.306
110	19.213	13.382	0.795	13.618	9.234	0.902	0.934	29.614	45.621
111	29.634	17.063	0.704	10.052	7.921	0.936	0.963	28.176	51.018
112	13.334	9.938	0.904	9.273	7.785	0.941	0.960	27.762	46.258
113	32.659	13.142	0.497	16.538	8.797	0.774	0.916	27.479	54.951
114	26.299	19.771	0.822	19.395	14.350	0.906	0.908	40.216	55.519
115	20.890	11.128	0.756	10.640	7.263	0.896	0.926	27.162	41.634
116	17.546	14.403	0.644	11.978	9.118	0.857	0.961	35.796	47.711
Mean:	24.511	15.251	0.725	15.428	10.880	0.851	0.928	33.561	53.372
Global:	61.922	16.879	0.806	35.539	11.666	0.907	0.961	35.796	47.711

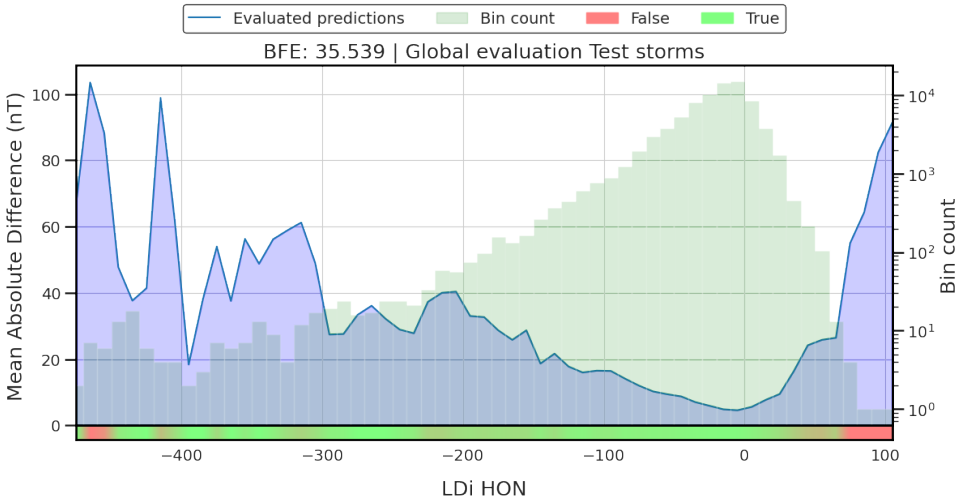


Figure 6.36: Global BFE for HON on test storms.

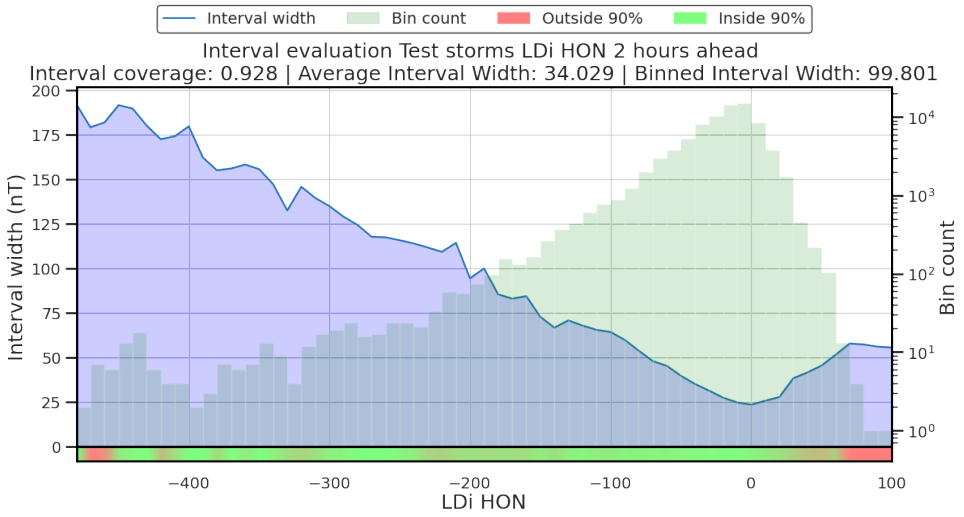


Figure 6.37: Interval Coverage for HON on test storms.

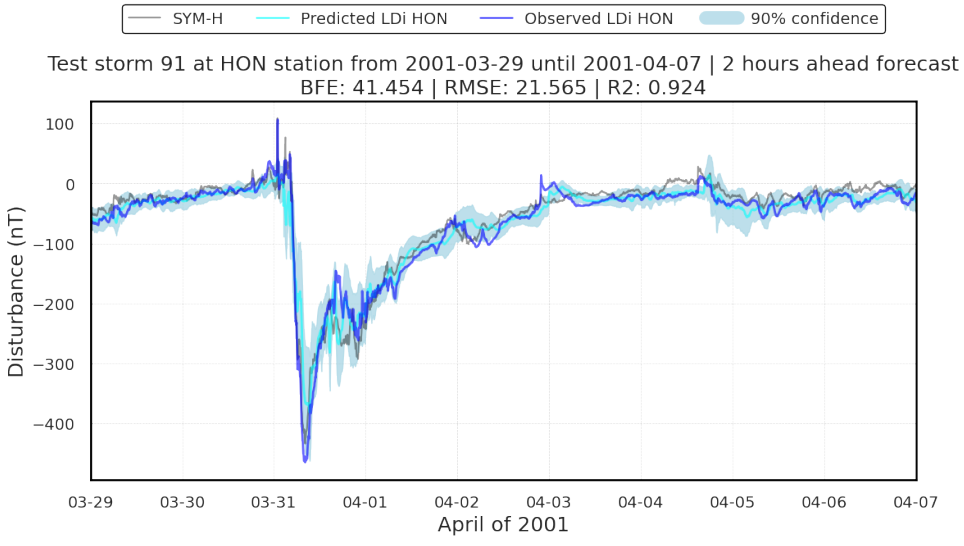


Figure 6.38: 2 hours ahead forecast of the test storm 91 of April 2001 for the LDi of HON.

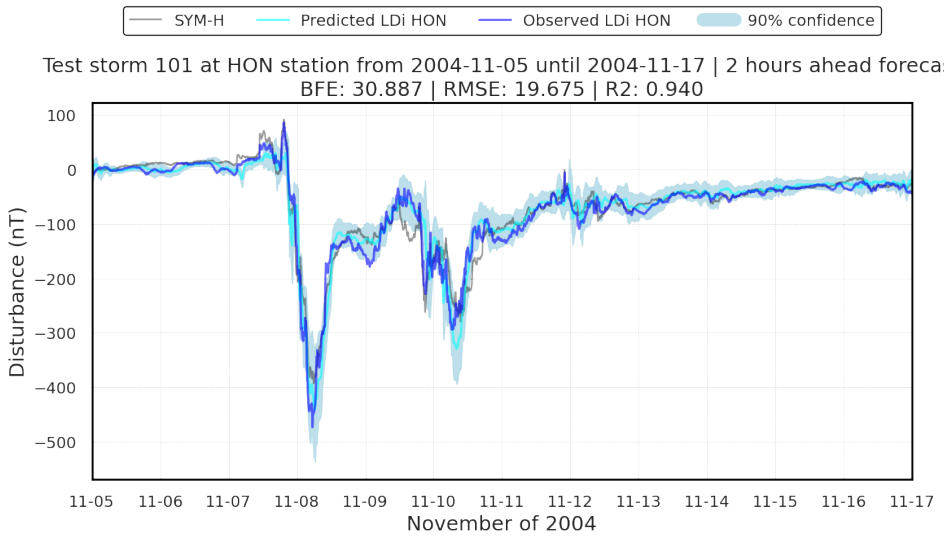


Figure 6.39: 2 hours ahead forecast of the test storm 101 of November 2004 for the LDi of HON.

Table 6.11 shows the metrics comparison for the test key storms. In this case, once again the improvement is notable, from 39.009 to 27.162 on the global computation of the BFE compared to the persistence model. The location of this station makes the LDi very unstable, making it very hard for a persistence model to work properly.

Table 6.11: Metrics of the trained model compared to the persistence model for the test key storms for the HON station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
117	39.545	18.486	0.789	25.327	12.674	0.901	0.904	28.729	54.972
118	13.472	8.279	0.795	8.145	6.469	0.875	0.936	24.120	37.687
119	27.815	14.533	0.485	25.177	13.532	0.553	0.855	26.202	41.564
120	11.911	9.400	0.813	8.353	6.972	0.897	0.977	25.347	33.137
121	29.833	12.433	0.496	13.750	9.184	0.725	0.932	23.233	41.637
122	7.363	8.207	0.719	8.808	7.554	0.762	0.895	23.849	25.579
Mean:	21.656	11.890	0.683	14.927	9.397	0.785	0.917	25.246	39.096
Global:	39.009	12.461	0.747	27.162	9.781	0.844	0.895	23.849	25.579

In this station, we can highlight the test key storm 122, shown in Figure 6.41. This storm is another example of why the forecast of local indices can be extremely difficult, while the disturbance is not that great on the global scale, as the SYM-H reached a peak close to -120 nT, the disturbance on the LDi on HON can't even be considered a storm, as it barely reached -60 nT. This underscores the difficulty of the problem and the high impact of the MLT and location of the station on the LDi.

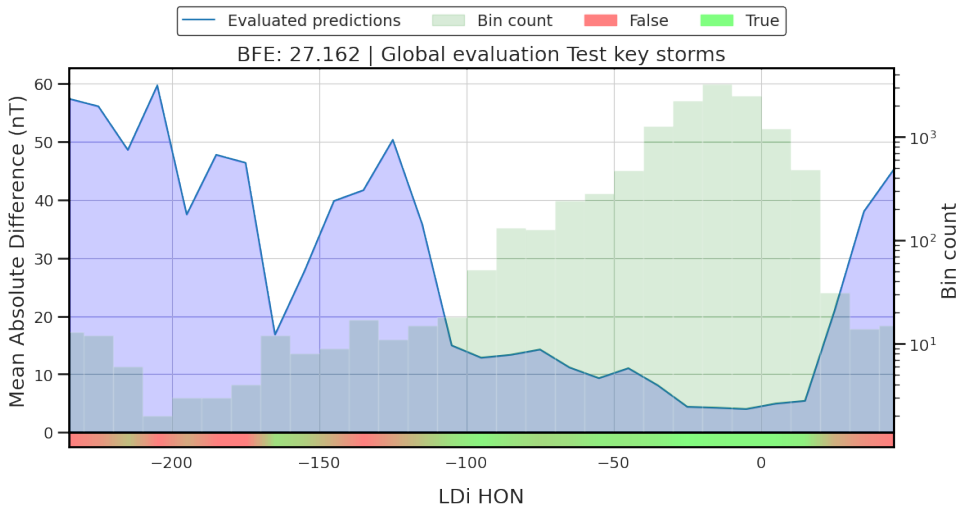


Figure 6.40: Global BFE for HON on the test key storms.

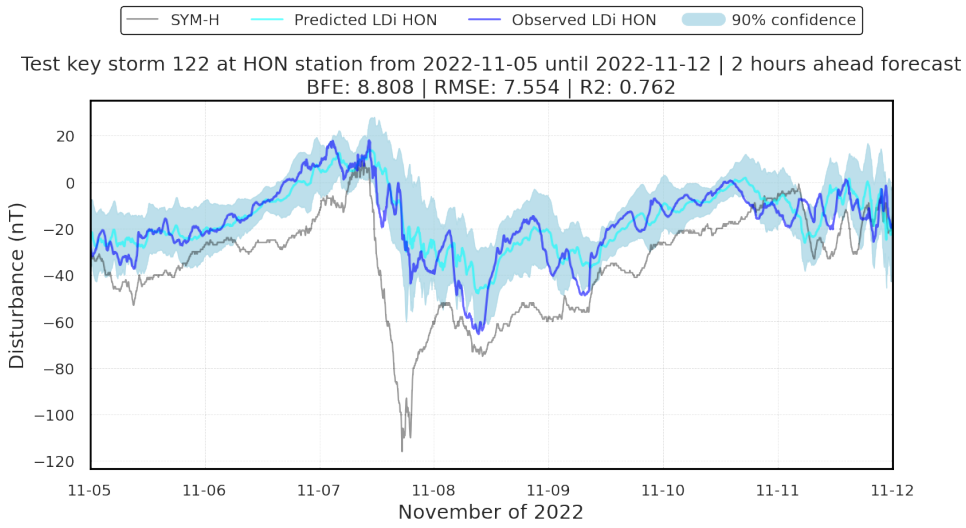


Figure 6.41: 2 hours ahead forecast of the test key storm 122 of November 2022 for the LDi of HON.

6.2.6 Testing the network on unseen stations

Given the architecture of our model, which processes input data from the ACE satellite and incorporates the LDi alongside the geographical coordinates of the stations, we hypothesize that the model is capable of forecasting the LDi for stations that it has not been explicitly trained on. The key aspect that enables this generalization is that the methodology for calculating the LDi remains consistent across different magnetometer stations, independent of their location. As long as the baseline data for a new station can be computed accurately, the model should be able to generate reliable forecasts for previously unseen stations.

To evaluate this capability, we conducted an experiment using data from the Coimbra geomagnetic observatory in Portugal, which is located at a latitude of 40.21º and a longitude of -8.417º. This location is relatively close to the SPT station in Spain. However, one significant limitation is that the Coimbra magnetometer was installed very recently, with the earliest data available from September 1st, 2021. This limited the amount of historical data that could be used to compute the station's baseline and calculate the LDi accurately.

Given these constraints, the test storm sets presented in Section 4.2.1.1 are not applicable for Coimbra. As a result, we selected storms that occurred between January 2023 and June 2024. Table 6.12 outlines the storms used for testing at the Coimbra station, detailing their storm index (following the last test key storm, number 122), the start and end dates, and the SYM-H peak values observed during each event.

Table 6.12: Selected storms for testing at the Coimbra station.

Storm index	Start date	End date	SYM-H peak (nT)
123	2023-02-23	2023-03-05	-161
124	2023-03-19	2023-03-29	-170
125	2023-04-19	2023-04-29	-231
126	2023-08-01	2023-08-10	-103
127	2023-09-08	2023-09-17	-83
128	2023-09-15	2023-09-24	-95
129	2023-10-17	2023-10-26	-107
130	2023-11-01	2023-11-11	-188
131	2023-11-21	2023-12-07	-132
132	2024-02-28	2024-03-09	-127
133	2024-03-20	2024-03-30	-170
134	2024-04-15	2024-04-24	-138
135	2024-04-28	2024-05-07	-97
136	2024-05-08	2024-05-16	-497

It is important to emphasize that this test presents several additional challenges. First, the solar wind and IMF parameters from ACE used for these storms are provisional, similar to the test key storms, which are of lower quality than the definitive parameters that the network was trained on. Furthermore, the model was trained with LDi, longitude, and latitude values from the previous five stations stations, and had no exposure to the Coimbra station during training. This combination of factors contributes to a decline in performance, which can be observed in the results. Notably, Storm 136, which occurred in May 2024, was an extremely intense event that contributed significantly to the model's challenges.

Table 6.13 summarizes the metrics for the selected storms at Coimbra, showing the model's performance relative to the persistence model. The figures demonstrate the broader performance trends and challenges faced by the network in this unseen environment. Additionally, Figures 6.42 and 6.43 present the plots for the computation of the BFE for all the evaluated storms and the interval coverage analysis, respectively.

Table 6.13: Metrics of the trained model compared to the persistence model for the evaluated storms for the Coimbra (COI) station.

#	Persistence model			Trained model					
	BFE	RMSE	R2	BFE	RMSE	R2	PICP	PIAW	PIBW
123	27.323	20.449	0.540	22.730	15.317	0.742	0.873	33.531	49.934
124	26.125	16.936	0.687	19.318	13.327	0.806	0.825	28.484	43.700
125	36.896	17.805	0.681	23.028	13.588	0.814	0.815	30.135	59.235
126	24.450	12.073	0.525	20.999	10.977	0.608	0.884	25.599	32.780
127	21.941	15.177	0.492	18.743	12.147	0.675	0.821	25.961	33.460
128	17.651	14.499	0.499	17.507	13.308	0.578	0.787	26.255	35.043
129	13.014	10.051	0.703	12.441	8.912	0.766	0.830	21.693	28.188
130	47.456	23.074	0.631	30.346	15.496	0.834	0.911	36.234	66.728
131	23.797	14.234	0.648	24.167	10.846	0.795	0.909	26.619	41.811
132	29.057	15.394	0.619	20.046	9.938	0.841	0.913	24.944	43.308
133	47.403	18.736	0.388	40.755	15.221	0.596	0.863	28.887	54.862
134	14.207	10.829	0.719	13.020	9.987	0.761	0.846	25.393	38.870
135	26.671	14.816	0.584	16.862	11.431	0.752	0.816	26.432	39.729
136	76.407	50.432	0.555	62.308	33.615	0.802	0.804	51.668	79.545
Mean:	30.886	18.179	0.591	24.448	13.865	0.741	0.850	29.417	46.228
Global:	73.540	20.061	0.650	58.790	14.718	0.812	0.855	29.273	72.897

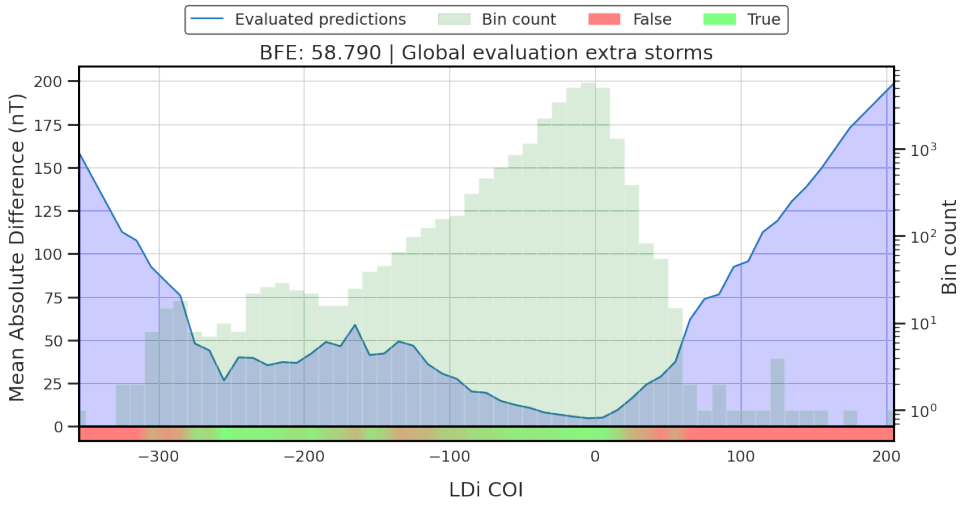


Figure 6.42: Global BFE for Coimbra on the evaluated storms.

Despite these challenges, the model demonstrates a reasonable ability to generalize, although there is a clear reduction in performance compared to its behavior on stations it was trained on. We can see that the trained model consistently outperforms the persistence model in terms of BFE, RMSE, and R^2 values. The mean BFE for the persistence model across the storms is 30.886, while the trained model achieves a lower average BFE of 24.448, indicating a notable improvement. The trained model also shows a lower RMSE on average, reducing the error from 18.179 to 13.865, and a higher average R^2 of 0.741,

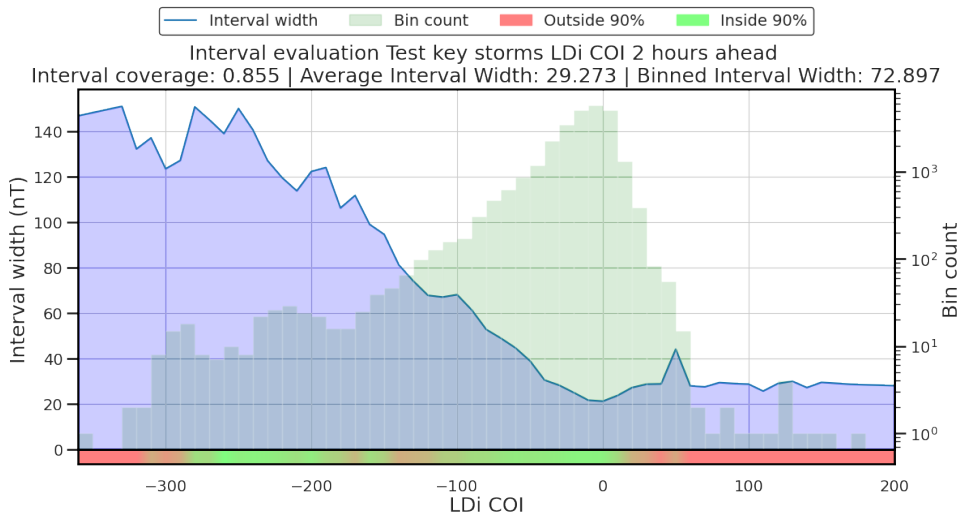


Figure 6.43: Interval coverage for Coimbra on evaluated storms.

indicating the model captures a significant portion of the variance in the data. For most storms, the trained model shows solid accuracy in forecasting the LDi. This is also reflected in the global evaluation, as the persistence model has a BFE of 73.540, while the trained model achieves a lower BFE of 58.790.

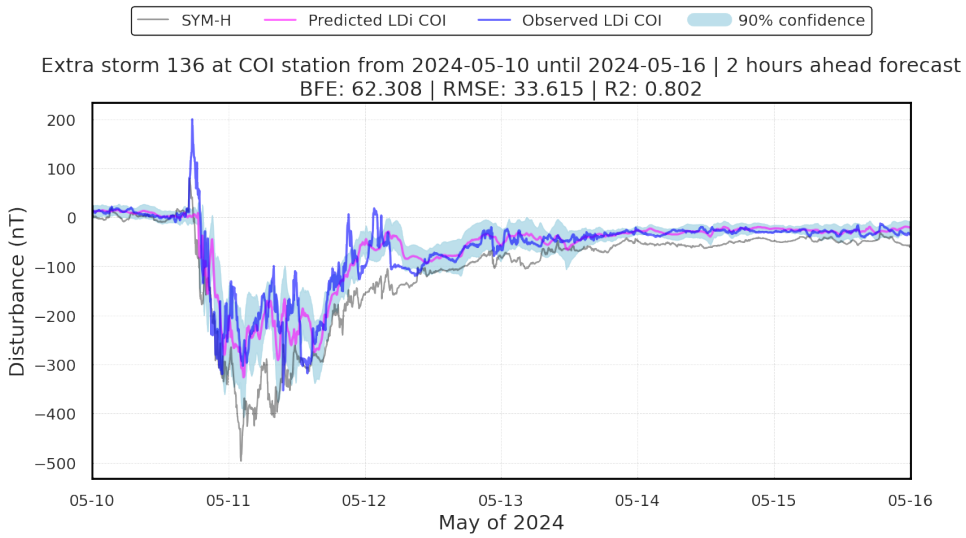


Figure 6.44: 2 hours ahead forecast of the storm 136 of May 2024 for the LDi of COI.

However, certain storms present more significant challenges for the model. For example, Storm 136, which is the superintense storm of May 2024, shown in Figure 6.44, shows the highest BFE and RMSE values across all storms, with a BFE of 62.308 for the trained model compared to 76.407 for the persistence model. Although the model still shows an improvement, the extreme intensity of this storm, combined with the provisional data and unfamiliar station, makes it more difficult for the model to maintain the level of accuracy observed in other storms. Most of the error in the BFE is located in the extreme positive peak at the start of the storm, which the model totally overlooks, this explains the error of close to 200 nT in the right side of the BFE plot. Additionally, the highest peak in the

LDi that happened during May 11th was forecasted slightly earlier, also contributing to a high BFE error.

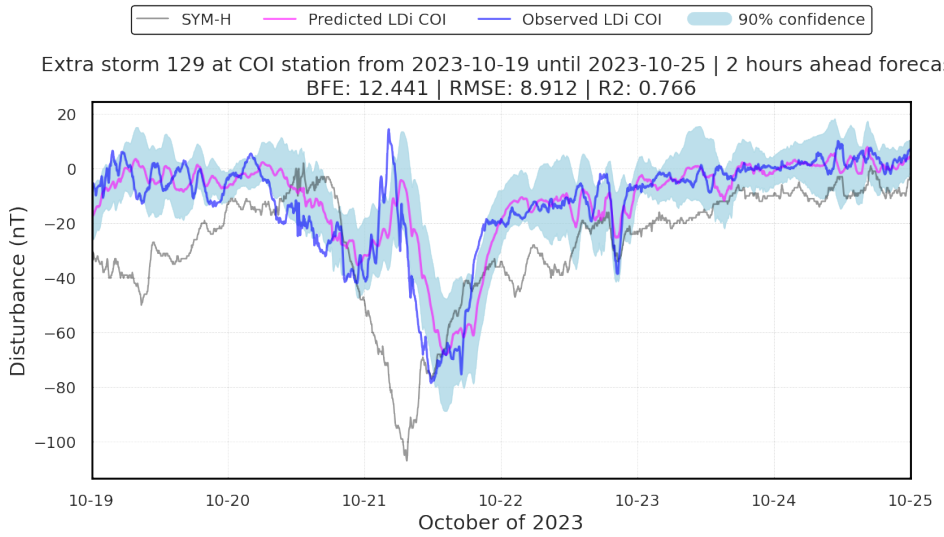


Figure 6.45: 2 hours ahead forecast of the storm 129 of October 2023 for the LDi of COI.

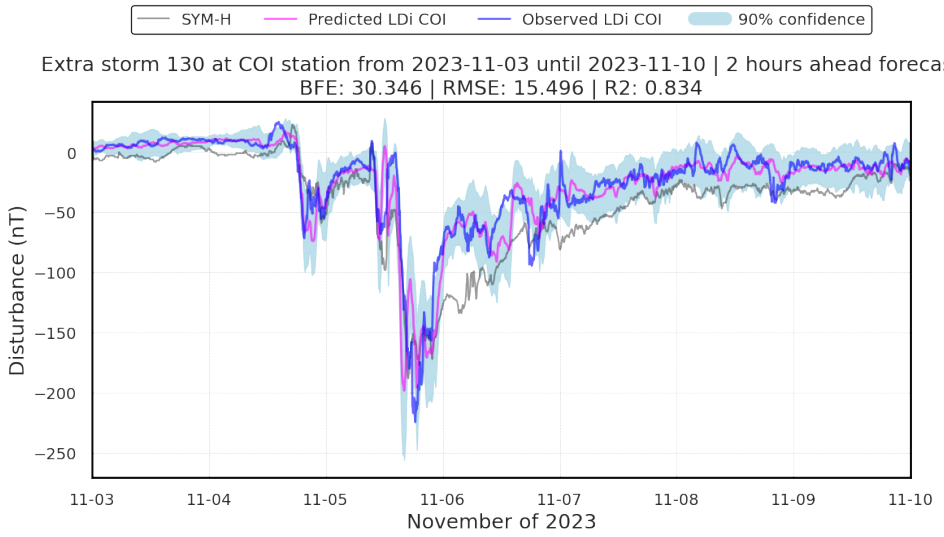


Figure 6.46: 2 hours ahead forecast of the storm 130 of November 2023 for the LDi of COI.

Aside from that particular storm, we can also highlight storms 129 and 130, shown in Figures 6.45 and 6.46. For storm 129, the model fails to timely forecast the disturbance, as it is slightly late; nevertheless this particular storm also showcases the behavior that we have seen before, of the LDi presenting a delay compared to the SYM-H. The other example, on storm 130, shows a better forecast, in this case the model forecasted the disturbance slightly earlier, but with considerably accuracy.

Aside from that, the interval evaluation plot shown in Figure 6.43 for the forecast uncertainty provides useful context. The interval coverage achieved by the model is 0.855, indicating that 85.5% of the true values fall within the 90% confidence interval predicted by the model. This is relatively consistent with the model's performance on trained stations, but the average interval width is wider in the Coimbra station test, with an average of 29.273 nT and a binned interval width of 72.897 nT. This suggests that the

model is more uncertain about its predictions in this unseen environment, as reflected in the broader intervals, especially for more extreme values of LDi. The increase in uncertainty is likely due to the combined impact of provisional solar wind parameters and the lack of direct LDi exposure during training.

In summary, the model demonstrates a robust ability to generalize to unseen stations, but its performance is understandably reduced under more challenging conditions, such as extreme storms and when provisional data is used. Despite this, the trained model consistently outperforms the persistence model, making it a valuable tool for forecasting geomagnetic disturbances, even in stations like Coimbra that it has not encountered before.

6.3 Conclusions

The results of our analysis suggest significant potential for further improving the model's performance by expanding the training dataset to include more stations from the INTERMAGNET network; some of them are shown in Figure 6.47. With a wider variety of geomagnetic observatories available worldwide, incorporating data from additional stations covering a broader range of geographical areas and magnetic conditions would likely enhance the model's generalization capabilities. By training the model with a more diverse set of stations, we expect that its overall performance would increase, even when tested on previously unseen stations. A more comprehensive training set would allow the model to better capture the variability of geomagnetic activity across different locations, leading to more accurate forecasts globally.

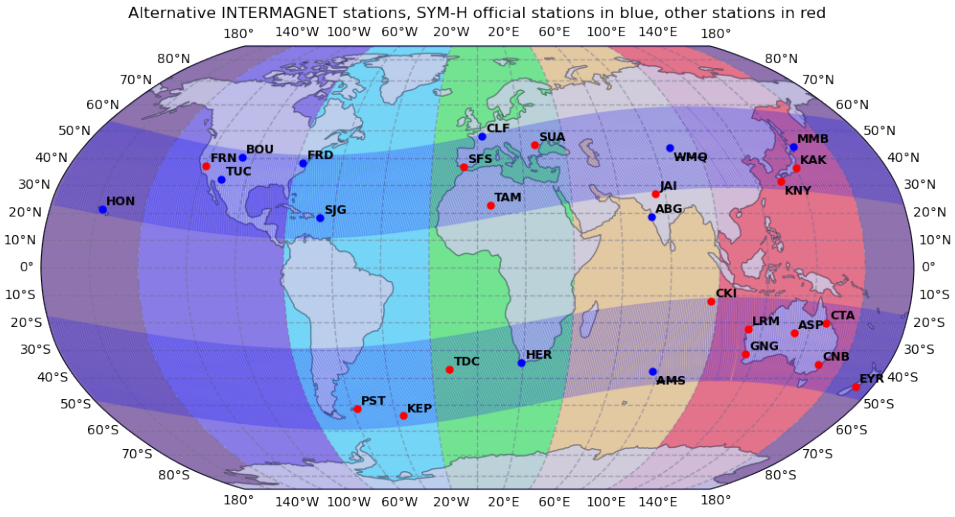


Figure 6.47: Other INTERMAGNET stations that could be used to train the model.

Additionally, these results underscore the feasibility of developing effective forecasting models for new stations, even when there is limited training data available. Although the model's performance on stations it has not been trained with, such as Coimbra, does not match its accuracy on trained stations, the forecasts remain reasonably reliable and useful. This capability opens up the possibility of deploying forecasting models at newly established or less-monitored stations, providing valuable geomagnetic forecasts that can

aid in mitigating the effects of geomagnetic storms, even with the inherent challenges of working with provisional or unseen data.

Chapter 7

Conclusions

"Predicting space weather: helping satellites avoid cosmic sunburns since forever."

This last chapter presents the main conclusions derived from this dissertation. The focus was on using DNNs to forecast geomagnetic indices, specifically the SYM-H and ASY-H global indices, along with local indices. The outcome shows that ML models can be used to forecast geomagnetic indices with notable accuracy, enabling more effective operational SW monitoring and mitigating potential technological disruptions caused by geomagnetic storms.

The conclusions are divided into three parts, corresponding to the main aspects covered in this thesis: (i) the training and evaluation framework, (ii) the global indices, covering the SYM-H and ASY-H indices, and (iii) the local indices. The dissertation ends with a list of future research lines.

7.1 Training and evaluation framework

Data selection is critical for any ML project. In forecasting geomagnetic indices with DNNs, ensuring a statistically valid storm selection and separation is essential. Previous works, notably Siciliano et al. [36], established a foundation for storm selection; however, not all relevant phenomena were covered, and many storms were excluded. Since that selection, numerous additional storms require incorporation into the training and testing subsets.

To address these limitations, we proposed a statistically backed classification system based on the cumulative distribution function of both indices. By applying percentiles from industry-wide practices, we classified storms according to intensity and occurrence probability, ensuring a more objective and transparent storm classification system. This comprehensive method captures a broader range of SW events, providing a more complete representation of different storm types. As new storms occur, they can be incorporated into the training data for the models, further enriching the training environment with a diverse and robust dataset. Additionally, the classification system is adaptable to other

geomagnetic indices, offering a flexible framework that can be readily applied beyond the SYM-H and ASY-H indices.

An equally important aspect of this research was to improve the way we evaluate the models' performance. While traditional regression metrics like RMSE, MAE, and R^2 are widely used, they fail to capture the specific particularities of forecasting geomagnetic storms, particularly during periods of high activity. To overcome this issue, we introduced the Binned Forecasting Error metric, which evaluates model performance across different storm intensities. This metric allows for a more nuanced and comprehensive assessment, facilitating a more accurate comparison of models' predictive capabilities, especially when storms vary in duration and intensity.

Finally, we emphasized the importance of recreating an operational environment to ensure that models are ready for real-time deployment. Forecasting geomagnetic indices in real-time requires the use of provisional solar wind and IMF data, often with minimal post-processing. To meet this need, we established a secondary test set using preliminary solar wind parameters, providing an environment where the model could be evaluated under conditions similar to real-time operations.

7.2 Global indices

In recent years, ML has become a powerful tool for forecasting geomagnetic indices. Unlike traditional physics-based models, ML models provide faster computation and great accuracy, processing large volumes of data to identify complex patterns not apparent in physical models. This approach has become the current trend in SW forecasting, with many researchers adopting ML techniques to enhance both the timeliness and reliability of predictions. As the field continues to evolve rapidly, ML shows great potential for revolutionizing geomagnetic forecasting by providing real-time, accurate predictions that are crucial for mitigating the adverse impacts of SW events.

We developed a DNN to forecast global geomagnetic indices using solar wind and IMF data from the ACE mission as inputs, along with the geomagnetic indices themselves. Our model not only outperformed previous models developed for the same task, but also introduced the capability to forecast prediction intervals. These intervals provide a range of values likely to contain the forecast with a specified confidence (90% in our case), giving users more actionable information for operational SW forecasting.

A significant achievement of this work is the successful deployment of the model in a real-time environment. The model is actively working in real-time, providing forecasts for geomagnetic storms as they occur. Since its deployment, the model has successfully predicted several storms, demonstrating its robustness and practical utility in operational SW forecasting scenarios.

The SYM-H index has attracted the most attention in recent geomagnetic storm forecasting research. Numerous studies have developed models to predict SYM-H with varying degrees of success, but our model currently achieves the highest level of accuracy. A major contributing factor for its popularity is the fact that SYM-H is a high-resolution version of the Dst index, which has long been used as a proxy for overall geomagnetic activity. Given the advancements in data resolution and quality, SYM-H provides a more detailed

representation of geomagnetic storms, enabling our DNN model to capture the underlying dynamics more effectively.

The results for the SYM-H index showed significant improvements in forecasting accuracy, particularly for short-term predictions. DNNs models demonstrated robust performance in predicting SYM-H values during both calm and storm conditions. Compared to other approaches, the proposed NN architecture reduced the RMSE, BFE and improved the FSS for most test storms. The integration of plasma data from instruments such as SWEPAM and SWICS was crucial in achieving these results, allowing for a more complete and accurate representation of the solar wind conditions driving geomagnetic disturbances. The operational evaluation, which included real-time deployment of the model, confirmed its practical applicability for SW forecasting.

Unlike SYM-H, ASY-H has received considerably less attention in the research community. Our model showed substantial improvement over previous attempts at forecasting ASY-H, particularly in capturing the recovery phases of geomagnetic storms.

Although the results for ASY-H were promising, its asymmetric nature and the lack of extensive historical data continue to pose challenges. Our model managed to improve the forecast accuracy, but further refinements will be necessary to address these difficulties, especially during periods of extreme geomagnetic activity.

7.3 Local indices

The development of local geomagnetic indices forecasting models has revealed valuable insights into the dynamics of geomagnetic storms at regional levels. By focusing on specific observatories, we demonstrated that local indices can capture crucial information that global indices, such as SYM-H, may overlook. These localized disturbances are significantly influenced by the geographical position of the observatories and the MLT when the storm reaches Earth, leading to varying levels of geomagnetic activity across different regions.

We successfully developed a DNN trained using data from several stations distributed across different longitudes. The model effectively captured the unique dynamics of local geomagnetic disturbances, demonstrating strong generalization capabilities across different geographic regions. By accounting for station-specific variability, the model was able to predict local disturbances accurately, offering more precise insights into the impact of geomagnetic storms on specific locations.

An important advantage of this approach is the ability to use the common model to forecast the LDi for stations that were not part of the training set. While the model's performance at these new locations does not match the accuracy of stations included in the training data, the forecasts remain reasonably reliable. This capability highlights the robustness and flexibility of the model, making it a valuable tool for forecasting local geomagnetic activity, even in regions where data might be limited for training.

In summary, the advancements made in forecasting local geomagnetic indices underscore the importance of local disturbances, which can differ significantly from global averages. Our work demonstrates that DNNs can provide reliable and useful predictions across a range of locations, further reinforcing the potential of ML in SW forecasting.

7.4 Future research lines

The work described in this thesis opens several extensions and directions for possible future works. Here, we review some of them.

- **Extending the forecast horizon:** While the current forecast lead time is constrained by the time it takes for solar wind disturbances to travel from the ACE satellite to the Earth, there is potential to extend the forecast horizon. However, as demonstrated throughout this dissertation, increasing the lead time significantly raises both the forecast error and uncertainty. Further research is required to strike a balance between forecast lead time and accuracy. It would be beneficial to engage with the end users of these forecasts, such as power grid operators and satellite companies, to better understand their needs regarding lead time, forecast precision, and acceptable levels of uncertainty. This user feedback could guide future improvements and customization of the forecasting models.
- **Using other indices:** Our work has successfully shown that the developed DNN can be adapted to forecast the LDi instead of the global SYM-H index. Building on this success, future research could expand the network to forecast other indices of operational interest, such as the recently developed Hp30 index. This index has been gaining attention for its application in SW forecasting and risk assessment. Extending the model to forecast Hp30 or other geomagnetic indices would enhance the utility of the model for different SW forecasting applications and provide a broader understanding of geomagnetic disturbances.
- **Training the network with additional local data:** The current model was trained and tested on data from a limited number of stations. To improve the model's generalization and robustness, future work should focus on training the network with data from a wider array of geomagnetic observatories. Additionally, it will be important to evaluate how well the model performs when forecasting for stations that were not included in the training dataset. This will give insights into the model's adaptability and its potential use for stations where data availability is limited because they have been recently installed or where the model has not been specifically trained.
- **Using solar imagery to increase lead time:** One promising avenue for increasing the forecast lead time is the use of solar imagery to predict geomagnetic disturbances before they reach the Earth. Historical data from the SOHO observatory provides an extensive dataset of solar imagery that could be used to train models in forecasting disturbances. In the near future, ESA's Vigil mission, which will provide imagery from the L5 point, will offer a prime vantage point for observing solar activity earlier than current missions. Incorporating these imagery datasets into the forecasting pipeline could extend the warning times for geomagnetic storms, giving industries and governments more time to prepare for potential disruptions. Further research should focus on integrating solar imagery with existing data-driven models to assess its feasibility and effectiveness in extending forecast horizons.

Bibliography

[1] J. E. Borovsky and Y. Y. Shprits, "Is the Dst Index Sufficient to Define All Geospace Storms?", *Journal of Geophysical Research: Space Physics*, vol. 122, no. 11, Nov. 2017, ISSN: 2169-9402. DOI: 10.1002/2017ja024679.

[2] K. P. Macpherson, A. J. Conway, and J. C. Brown, "Prediction of solar and geomagnetic activity data using neural networks", *Journal of Geophysical Research: Space Physics*, vol. 100, no. A11, pp. 21 735–21 744, Nov. 1995, ISSN: 0148-0227. DOI: 10.1029/95ja02283.

[3] A. W. P. Thomson, "Non-linear predictions of Ap by activity class and numerical value", *pure and applied geophysics*, vol. 146, no. 1, pp. 163–193, Feb. 1996, ISSN: 1420-9136. DOI: 10.1007/bf00876675.

[4] A. Nagai, *Prediction of magnetospheric parameters using artificial neural networks*. USA: Rice University, 1995.

[5] V. Sreeja, "Impact and mitigation of Space Weather effects on GNSS receiver performance", *Geoscience Letters*, Aug. 2016. DOI: 10.1186/s40562-016-0057-0.

[6] R. Pirjola, "Space weather effects on power grids", in *Space Weather- Physics and Effects*, Springer Praxis Books, 2007, pp. 269–288. DOI: 10.1007/978-3-540-34578-7_10.

[7] J. Kappernman and V. Albertson, "Bracing for the geomagnetic storms", *IEEE Spectrum*, vol. 27, no. 3, pp. 27–33, Mar. 1990. DOI: 10.1109/6.48847.

[8] R. Pirjola, A. Pulkkinen, and A. Viljanen, "Studies of Space Weather effects on the finnish natural gas pipeline and on the finnish high-voltage power system", *Advances in Space Research*, Jan. 2003. DOI: 10.1016/s0273-1177(02)00781-0.

[9] R. Pirjola, "Effects of Space Weather on high-latitude ground systems", *Advances in Space Research*, Jan. 2005. DOI: 10.1016/j.asr.2003.04.074.

[10] R. Gummow and P. Eng, "GIC effects on pipeline corrosion and corrosion control systems", *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 64, no. 16, pp. 1755–1764, Nov. 2002. DOI: 10.1016/s1364-6826(02)00125-6.

[11] W. H. Campbell, "Induction of auroral zone electric currents within the Alaska pipeline", *Pure and Applied Geophysics PAGEOPH*, vol. 116, no. 6, pp. 1143–1173, 1978. DOI: 10.1007/bf00874677.

[12] S. A. Jyothi, "Solar superstorms", in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, Aug. 2021, pp. 692–704. DOI: 10.1145/3452296.3472916.

[13] D. J. Knipp, B. J. Fraser, M. A. Shea, and D. F. Smart, "On the Little-Known Consequences of the 4 August 1972 Ultra-Fast Coronal Mass Ejecta: Facts, Commentary, and Call to Action", *Space Weather*, Nov. 2018. DOI: 10.1029/2018sw002024.

[14] A. L. S. A. Center, *True Impacts of Space Weather on a Ground Force*, Available at https://www.alsa.mil/News/Article/2532178/true-impacts-of-space-weather-on-a-ground-force/, 2017.

[15] SpaceX, "Geomagnetic storm and recently deployed Starlink satellites", SpaceX, Tech. Rep., Jan. 2022. [Online]. Available: https://www.spacex.com/updates/.

[16] D. Smart, M. Shea, and K. McCracken, "The Carrington event: Possible solar proton intensity–time profile", *Advances in Space Research*, vol. 38, no. 2, pp. 215–225, Jan. 2006. DOI: 10.1016/j.asr.2005.04.116.

[17] J. P. Eastwood, E. Biffis, M. A. Hapgood, *et al.*, "The Economic Impact of Space Weather: Where Do We Stand?", *Risk Analysis*, Feb. 2017. DOI: 10.1111/risa.12765.

[18] M. Hapgood, M. J. Angling, G. Attrill, *et al.*, "Development of Space Weather Reasonable Worst-Case Scenarios for the UK National Risk Assessment", *Space Weather*, Apr. 2021. DOI: 10.1029/2020sw002593.

[19] J. H. King, "Solar Wind Spatial Scales in and Comparisons of Hourly Wind and ACE Plasma and Magnetic Field Data", *Journal of Geophysical Research*, vol. 110, no. A2, 2005. DOI: 10.1029/2004ja010649.

[20] C. Larrodera and C. Cid, "Bimodal distribution of the solar wind at 1 AU", *Astronomy & Astrophysics*, vol. 635, A44, 2020. DOI: 10.1051/0004-6361/201937307.

[21] J. Bartels, N. H. Heck, and H. F. Johnston, "The Three-Hour-Range Index Measuring Geomagnetic Activity", *Journal of Geophysical Research*, 1939. DOI: 10.1029/te044i004p00411.

[22] M. Sugiura, "Hourly Values of Equatorial Dst for the Igy", *Ann. Int. Geophys. Yr.*, Jan. 1964. [Online]. Available: https://www.osti.gov/biblio/4554034.

[23] T. N. Davis and M. Sugiura, "Auroral Electrojet Activity index and Its Universal Time Variations", *Journal of Geophysical Research*, 1966. DOI: 10.1029/jz071i003p00785.

[24] P. N. Mayaud, "Indices Kn, Ks et Km", Tech. Rep., 1968. [Online]. Available: http://isgi.unistra.fr/Documents/References/Mayaud%7B%5C_%7DCNRS%7B%5C_%7D1968.pdf.

[25] P.-N. Mayaud, "The aa indices: A 100-year series characterizing the magnetic activity", *Journal of Geophysical Research*, vol. 77, no. 34, pp. 6870–6874, Dec. 1972. DOI: 10.1029/ja077i034p06870.

[26] O. Troshichev, N. Dmitrieva, and B. Kuznetsov, "Polar cap magnetic activity as a signature of substorm development", *Planetary and Space Science*, vol. 27, no. 3, pp. 217–221, 1979, ISSN: 0032-0633. DOI: 10.1016/0032-0633(79)90063-1.

[27] S. Vennerstrom, E. Friis-Christensen, O. A. Troshichev, and V. G. Andersen, "Geomagnetic Polar Cap (PC) Index, 1975-1993", Tech. Rep., 1994, p. 1975. [Online]. Available: https://ui.adsabs.harvard.edu/abs/1994gpcp.rept.1975V.

[28] M. Sugiura and S. Chapman, "The average morphology of geomagnetic storms with sudden commencement", High Altitude Observatory Boulder Colo, Tech. Rep., 1961. [Online]. Available: https://cir.nii.ac.jp/crid/1571698599066035712.

[29] S.-I. Akasofu and S. Chapman, "On the asymmetric development of magnetic storm fields in low and middle latitudes", *Planetary and Space Science*, vol. 12, no. 6, pp. 607–626, Jun. 1964. DOI: 10.1016/0032-0633(64)90008-x.

[30] J. Palacios, A. Guerrero, C. Cid, E. Saiz, and Y. Cerrato, "Defining Scale Thresholds for Geomagnetic Storms Through Statistics (Preprint)", *Natural Hazards and Earth System Sciences Discussions*, vol. 2018, pp. 1–17, 2018. DOI: 10.5194/nhess-2018-92.

[31] L. A. Dremukhina, Y. I. Yermolaev, and I. G. Lodkina, "Differences in the Dynamics of the Asymmetrical Part of the Magnetic Disturbance during the Periods of Magnetic Storms Induced by Different Interplanetary Sources", *Geomagnetism and Aeronomy*, vol. 60, no. 6, pp. 714–726, Nov. 2020, ISSN: 1555-645X. DOI: 10.1134/s0016793220060031.

[32] J. Bartels, *The standardized index Ks and the planetary index Kp, IATME Bulletin 12b*. 1949.

[33] T. Iyemori, M. Takeda, M. Nose, Y. Odagi, and H. Toh, "Mid-latitude geomagnetic indices asy and sym for 2009 (provisional)", Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Japan, Tech. Rep., 2010, https://wdc.kugi.kyoto-u.ac.jp/aeasy/asy.pdf.

[34] Y. Yamazaki, J. Matzka, C. Stolle, *et al.*, "Geomagnetic Activity Index Hpo", *Geophysical Research Letters*, vol. 49, no. 10, May 2022, ISSN: 1944-8007. DOI: 10.1029/2022gl098860.

[35] C. Cid, A. Guerrero, E. Saiz, A. Halford, and A. Kellerman, "Developing the LDi and LCi geomagnetic indices, an example of application of the AULs framework", *Space Weather*, vol. 18, no. 1, 2020. DOI: 10.1029/2019SW002171.

[36] F. Siciliano, G. Consolini, R. Tozzi, M. Gentili, F. Giannattasio, and P. D. Michelis, "Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks", *Space Weather*, Feb. 2021. DOI: 10.1029/2020sw002589.

[37] A. Collado-Villaverde, P. Muñoz, and C. Cid, "Deep Neural Networks with Convolutional and LSTM layers for SYM-H and ASY-H forecasting", *Space Weather*, Jun. 2021. DOI: 10.1029/2021sw002748.

[38] D. Iong, Y. Chen, G. Toth, *et al.*, "New Findings From Explainable SYM-H Forecasting Using Gradient Boosting Machines", *Space Weather*, vol. 20, no. 8, Aug. 2022, ISSN: 1542-7390. DOI: 10.1029/2021sw002928.

[39] J. M. Zurada, *Introduction to artificial neural systems*. Mumbai, India: Jaico Publishing House, Aug. 2006, ISBN: 8172246501.

[40] A. Geron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow*. Sebastopol, CA: O'Reilly Media, Mar. 2017, ISBN: 1492032646.

[41] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language Models are Few-Shot Learners*, 2020. DOI: 10.48550/ARXIV.2005.14165.

[42] K. Umapavankumar, S. V. N. Srinivasu, S. S. N. Rao, and S. N. T. Rao, "Machine Learning Usage in Facebook, Twitter and Google Along with the Other Tools", in *Emerging Research in Data Engineering Systems and Computer Communications*, Springer Singapore, 2020, pp. 465–471. DOI: 10.1007/978-981-15-0135-7_43.

[43] H. Erdinc Kocer and K. Kursat Cevik, "Artificial Neural Networks based vehicle license plate recognition", *Procedia Computer Science*, vol. 3, pp. 1033–1037, 2011, World Conference on Information Technology, ISSN: 1877-0509. DOI: 10.1016/j.procs.2010.12.169.

[44] R. Uhrig, "Introduction to artificial Neural Networks", in *Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics*, ser. IECON-95, vol. 1, IEEE, pp. 33–37. DOI: 10.1109/iecon.1995.483329.

[45] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. [Online]. Available: https://www.tensorflow.org/.

[46] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[47] G. Cramer, R. Ford, and R. Hall, "Estimation of toxic hazard—a decision tree approach", *Food and cosmetics toxicology*, vol. 16, no. 3, pp. 255–276, 1976. DOI: 10.1016/S0015-6264(76)80522-6.

[48] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on Neural Network interpretability", *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. DOI: 10.1109/TETCI.2021.3100641.

[49] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks", *CoRR*, 2015. DOI: 10.48550/arXiv.1511.08458. arXiv: 1511.08458.

[50] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[51] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1706.03762.

[52] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey", *ACM Computing Surveys*, Jan. 2022. DOI: 10.1145/3505244.

[53] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting", *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021. DOI: 10.1016/j.ijforecast.2021.03.012.

[54] M. D. Zeiler, "ADADELTA: an adaptive learning rate method", *CoRR*, 2012. DOI: 10.48550/arXiv.1212.5701. arXiv: 1212.5701.

[55] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "RMSProp and equilibrated adaptive learning rates for non-convex optimization.", *CoRR*, 2015. DOI: 10.48550/arXiv.1502.04390.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", Tech. Rep., 2014. DOI: 10.48550/arxiv.1412.6980.

[57] E. Camporeale, "The challenge of Machine Learning in Space Weather: Nowcasting and forecasting", *Space weather*, vol. 17, no. 8, pp. 1166–1207, 2019. DOI: 10.1029/2018sw002061.

[58] N. C. Joshi, N. S. Bankoti, S. Pande, B. Pande, and K. Pandey, "Relationship between interplanetary field/plasma parameters with geomagnetic indices and their behavior during intense geomagnetic storms", *New Astronomy*, vol. 16, no. 6, pp. 366–385, 2011, ISSN: 1384-1076. DOI: 10.1016/j.newast.2011.01.004.

[59] M. W. Liemohn, J. P. McCollough, V. K. Jordanova, *et al.*, "Model Evaluation Guidelines for Geomagnetic Index Predictions", *Space Weather*, 2018. DOI: 10.1029/2018sw002067.

[60] R. K. Burton, R. McPherron, and C. Russell, "An Empirical Relationship Between Interplanetary Conditions and Dst", *Journal of geophysical research*, vol. 80, no. 31, pp. 4204–4214, 1975. DOI: 10.1029/JA080i031p04204.

[61] T. P. O'Brien and R. L. McPherron, "An Empirical Phase Space Analysis of Ring Current Dynamics: Solar Wind Control of Injection and Decay", *Journal of Geophysical Research: Space Physics*, vol. 105, no. A4, pp. 7707–7719, 2000. DOI: 10.1029/1998JA000437.

[62] L. Rastätter, M. M. Kuznetsova, A. Glocer, *et al.*, "Geospace Environment Modeling 2008–2009 Challenge: Dst Index", *Space Weather*, 2013. DOI: 10.1002/swe.20036.

[63] S. Wing, J. R. Johnson, J. Jen, *et al.*, "Kp forecast models", *Journal of Geophysical Research: Space Physics*, vol. 110, no. A4, Apr. 2005. DOI: 10.1029/2004ja010500.

[64] P. Wintoft, M. Wik, J. Matzka, and Y. Shprits, "Forecasting kp from solar wind data: Input parameter study using 3-hour averages and 3-hour range values", *Journal of Space Weather and Space Climate*, 2017. DOI: 10.1051/swsc/2017027.

[65] I. S. Zhelavskaya, R. Vasile, Y. Y. Shprits, C. Stolle, and J. Matzka, "Systematic Analysis of Machine Learning and Feature Selection Techniques for Prediction of the Kp Index", *Space Weather*, 2019. DOI: 10.1029/2019sw002271.

[66] J. V. Hernandez, T. Tajima, and W. Horton, "Neural Net Forecasting For Geomagnetic Activity", *Geophysical Research Letters*, vol. 20, no. 23, pp. 2707–2710, 1993. DOI: 10.1029/93gl02848.

[67] G. Tóth, B. van der Holst, I. V. Sokolov, *et al.*, "Adaptive numerical algorithms in Space Weather modeling", *Journal of Computational Physics*, Feb. 2012. DOI: 10.1016/j.jcp.2011.02.006.

[68] Y. Tan, Q. Hu, Z. Wang, and Q. Zhong, "Geomagnetic index kp forecasting with LSTM", *Space Weather*, Apr. 2018. DOI: 10.1002/2017sw001764.

[69] M. Maimaiti, B. Kunduri, J. M. Ruohoniemi, J. B. H. Baker, and L. L. House, "A Deep Learning-Based Approach to Forecast the Onset of Magnetic Substorms", *Space Weather*, 2019. DOI: 10.1029/2019sw002251.

[70] R. J. Licata, W. K. Tobiska, and P. M. Mehta, "Benchmarking Forecasting Models for Space Weather Drivers", *Space Weather*, vol. 18, no. 10, Oct. 2020, ISSN: 1542-7390. DOI: 10.1029/2020sw002496.

[71] S. Hu, A. Bhattacharjee, J. Hou, *et al.*, "Ionospheric Storm Forecast for High-Frequency Communications", *Radio Science*, vol. 33, no. 5, pp. 1413–1428, 1998. DOI: 10.1029/98RS02219.

[72] H. Wei, D. Zhu, S. Billings, and M. Balikhin, "Forecasting the Geomagnetic Activity of the Dst Index Using Multiscale Radial Basis Function Networks", *Advances in Space Research*, vol. 40, no. 12, pp. 1863–1870, Jan. 2007. DOI: 10.1016/j.asr.2007.02.080.

[73] E.-Y. Ji, Y.-J. Moon, N. Gopalswamy, and D.-H. Lee, "Comparison of Dst Forecast Models for Intense Geomagnetic Storms", *Journal of Geophysical Research: Space Physics*, 2012. DOI: 10.1029/2011ja016872.

[74] G. Pallocchia, E. Amata, G. Consolini, M. F. Marcucci, and I. Bertello, "Geomagnetic Dst index forecast based on IMF data only", *Annales Geophysicae*, vol. 24, no. 3, pp. 989–999, 2006. DOI: 10.5194/angeo-24-989-2006.

[75] M. A. Gruet, M. Chandorkar, A. Sicard, and E. Camporeale, "Multiple-Hour-Ahead Forecast of the Dst Index Using a Combination of Long Short-Term Memory Neural Network and Gaussian Process", *Space Weather*, 2018. DOI: 10.1029/2018sw001898.

[76] J. A. Lazzús, P. Vega-Jorquera, L. Palma-Chilla, M. V. Stepanova, and N. V. Romanova, "Dst Index Forecast Based on Ground-Level Data Aided by Bio-Inspired Algorithms", *Space Weather*, 2019. DOI: 10.1029/2019sw002215.

[77] C. Forsyth, C. E. J. Watt, M. K. Mooney, I. J. Rae, S. D. Walton, and R. B. Horne, "Forecasting GOES 15 >2mev electron fluxes from solar wind data and geomagnetic indices", *Space Weather*, Jul. 2020. DOI: 10.1029/2019sw002416.

[78] M. Tshisaphungo, J. B. Habarulema, and L.-A. McKinnell, "Modeling ionospheric foF2 response during geomagnetic storms using Neural Network and linear regression techniques", *Advances in Space Research*, vol. 61, no. 12, pp. 2891–2903, Jun. 2018. DOI: 10.1016/j.asr.2018.03.025.

[79] T. Iyemori, "Storm-Time Magnetospheric Currents Inferred From Mid-Latitude Geomagnetic Field Variations", *Journal of geomagnetism and geoelectricity*, 1990. DOI: 10.5636/jgg.42.1249.

[80] A. Bhaskar and G. Vichare, "Forecasting of SYM-H and ASY-H indices for geomagnetic storms of solar cycle 24 including St. Patrick's day, 2015 storm using NARX neural network", *Journal of Space Weather and Space Climate*, 2019. DOI: 10.1051/swsc/2019007.

[81] S. Nijman, A. Leeuwenberg, I. Beekers, *et al.*, "Missing data is poorly handled and reported in prediction model studies using Machine Learning: a literature review", *Journal of clinical epidemiology*, vol. 142, pp. 218–229, 2022. DOI: 10.1016/j.jclinepi.2021.11.023.

[82] F. Muharemi, D. Logofătu, and F. Leon, "Review on General Techniques and Packages for Data Imputation in R on a Real World Dataset", in *Computational Collective Intelligence*, Springer International Publishing, 2018, pp. 386–395. DOI: 10.1007/978-3-319-98446-9_36.

[83] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction", *BMC Medical Research Methodology*, vol. 20, no. 1, Jul. 2020. DOI: 10.1186/s12874-020-01080-1.

[84] M. Temerin and X. Li, "A new model for the prediction of Dst on the basis of the solar wind", *Journal of Geophysical Research: Space Physics*, vol. 107, no. A12, SMP 31–1–SMP 31–8, Dec. 2002. DOI: 10.1029/2001ja007532.

[85] A. M. Keesee, V. Pinto, M. Coughlan, C. Lennox, M. S. Mahmud, and H. K. Connor, "Comparison of Deep Learning Techniques to Model Connections Between Solar Wind and Ground Magnetic Perturbations", *Frontiers in Astronomy and Space Science*, vol. 7, Oct. 2020. DOI: 10.3389/fspas.2020.550874.

[86] R. C. Wang, D. Li, T. Sun, X. Peng, Z. Yang, and J. Q. Wang, "A 3D Magnetospheric CT Reconstruction Method Based on 3D GAN and Supplementary Limited-Angle 2D Soft X-Ray Images", *Journal of Geophysical Research: Space Physics*, vol. 128, no. 1, Jan. 2023, ISSN: 2169-9402. DOI: 10.1029/2022ja030424.

[87] Y. Yang and F. Shen, "Modeling the global distribution of solar wind parameters on the source surface using multiple observations and the Artificial Neural Network technique", *Solar Physics*, vol. 294, no. 8, Aug. 2019. DOI: 10.1007/s11207-019-1496-5.

[88] Y. Yang, F. Shen, Z. Yang, and X. Feng, "Prediction of solar wind speed at 1 AU using an Artificial Neural Network", *Space Weather*, vol. 16, no. 9, pp. 1227–1244, 2018. DOI: 10.1029/2018SW001955.

[89] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1705.07874.

[90] P. Newell, T. Sotirelis, K. Liou, C.-I. Meng, and F. Rich, "A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables", *Journal of Geophysical Research: Space Physics*, vol. 112, no. A1, 2007. DOI: 10.1029/2006JA012015.

[91] P. Poudel, S. Simkhada, B. Adhikari, D. Sharma, and J. J. Nakarmi, "Variation of solar wind parameters along with the understanding of energy dynamics within the magnetospheric system during geomagnetic disturbances", *Earth and Space Science*, vol. 6, no. 2, pp. 276–293, Feb. 2019, ISSN: 2333-5084. DOI: 10.1029/2018ea000495.

[92] B. S. Rathore, D. C. Gupta, and K. K. Parashar, "Relation between solar wind parameter and geomagnetic storm condition during cycle-23", *International Journal of Geosciences*, vol. 05, no. 13, pp. 1602–1608, 2014, ISSN: 2156-8367. DOI: 10.4236/ijg.2014.513131.

[93] M. Alves, E. Echer, and W. Gonzalez, "Geoeffectiveness of corotating interaction regions as measured by Dst index", *Journal of Geophysical Research: Space Physics*, vol. 111, no. A7, 2006. DOI: 10.1029/2005JA011379.

[94] R. Kane, "Relationship between interplanetary plasma parameters and geomagnetic Dst", *Journal of Geophysical Research*, vol. 79, no. 1, pp. 64–72, 1974. DOI: 10.1029/JA079i001p00064.

[95] R. Kane, "How good is the relationship of solar and interplanetary plasma parameters with geomagnetic storms?", *Journal of Geophysical Research: Space Physics*, vol. 110, no. A2, 2005. DOI: 10.1029/2004JA010799.

[96] E. Nahayo, A. Guerrero, S. Lotz, C. Cid, M. Tshisaphungo, and E. Saiz, "Validating the LDi and LCi indices in the southern hemisphere", *Space Weather*, vol. 20, no. 10, Oct. 2022. DOI: 10.1029/2022sw003092.

[97] J. Curto, T. Araki, and L. Alberca, "Evolution of the concept of sudden storm commencements and their operative identification", *Earth, planets and space*, vol. 59, no. 11, pp. i–xii, 2007. DOI: 10.1186/BF03352059.

[98] R. M. Skoug, J. T. Gosling, J. T. Steinberg, *et al.*, "Extremely high speed solar wind: 29–30 October 2003", *Journal of Geophysical Research: Space Physics*, vol. 109, no. A9, Sep. 2004, ISSN: 0148-0227. DOI: 10.1029/2004ja010494.

[99] L. Rosenqvist, H. Opgenoorth, S. Buchert, I. McCrea, O. Amm, and C. Lathuillere, "Extreme solar-terrestrial events of october 2003: High-latitude and cluster observations of the large geomagnetic disturbances on 30 october", *Journal of Geophysical Research: Space Physics*, vol. 110, no. A9, Sep. 2005, ISSN: 0148-0227. DOI: 10.1029/2004ja010927.

[100] Ace Science Center, *SWEPAM/SWICS Level 3 Merged Solar Wind Proton Data Documentation*, https://izw1.caltech.edu/ACE/ASC/level2/sweswi_l2desc.html, Feb. 2010.

[101] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[102] M. Gardner and S. Dorling, "Artificial Neural Networks (The Multilayer Perceptron)—a Review of Applications in the Atmospheric Sciences", *Atmospheric Environment*, vol. 32, no. 14, pp. 2627–2636, 1998, ISSN: 1352-2310. DOI: 10.1016/S1352-2310(97)00447-0.

[103] I. Koprinska, D. Wu, and Z. Wang, "Convolutional Neural Networks for Energy Time Series Forecasting", in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2018. DOI: 10.1109/ijcnn.2018.8489399.

[104] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification", *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017. DOI: 10.21629/JSEE.2017.01.18.

[105] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network", *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017. DOI: 10.1109/TSG.2017.2753802.

[106] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification", *Neural Networks*, vol. 116, pp. 237–245, 2019. DOI: 10.1016/j.neunet.2019.04.014.

[107] M. Shi, K. Wang, and C. Li, "A C-LSTM with Word Embedding Model for News Text Classification", in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, IEEE, Jun. 2019. DOI: 10.1109/icis46139.2019.8940289.

[108] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *International Conference on Learning Representations*, 2014. DOI: 10.48550/ARXIV.1409.0473.

[109] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016. DOI: 10.1109/cvpr.2016.90.

[110] S. Reza, M. C. Ferreira, J. Machado, and J. M. R. Tavares, "A Multi-Head Attention-based Transformer model for Traffic flow forecasting with a comparative analysis to Recurrent Neural Networks", *Expert Systems with Applications*, vol. 202, p. 117 275, Sep. 2022, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.117275.

[111] X. Du, H. Zhang, H. V. Nguyen, and Z. Han, "Stacked LSTM Deep Learning Model for Traffic Prediction in Vehicle-to-Vehicle Communication", in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, IEEE, Sep. 2017. DOI: 10.1109/vtcfall.2017.8288312.

[112] S. O. Ojo, P. A. Owolawi, M. Mphahlele, and J. A. Adisa, "Stock Market Behaviour Prediction using Stacked LSTM Networks", in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, IEEE, Nov. 2019. DOI: 10.1109/imitec45504.2019.9015840.

[113] E. Oh, T. Kim, Y. Ji, and S. Khyalia, "STING: Self-attention based Time-series Imputation Networks using GAN", in *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2021, pp. 1264–1269. DOI: 10.1109/ICDM51629.2021.00155.

[114] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, pp. 107–116, Apr. 1998. DOI: 10.1142/S0218488598000094.

[115] J. Zhuang, T. Tang, Y. Ding, *et al.*, "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients", *Advances in neural information processing systems*, vol. 33, pp. 18 795–18 806, 2020. DOI: 10.48550/arXiv.2010.07468.

[116] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[117] T. O'Brien and R. L. McPherron, "Forecasting the ring current index Dst in real time", *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 62, no. 14, pp. 1295–1299, Sep. 2000, ISSN: 1364-6826. DOI: 10.1016/s1364-6826(00)00072-9.

[118] A. H. Murphy, "Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient", *Monthly Weather Review*, vol. 116, no. 12, pp. 2417–2424, Dec. 1988. DOI: 10.1175/1520-0493(1988)116<2417:ssbotm>2.0.co;2.

[119] E. Amata, G. Pallocchia, G. Consolini, M. Marcucci, and I. Bertello, "Comparison between three algorithms for Dst predictions over the 2003–2005 period", *Journal of atmospheric and solar-terrestrial physics*, vol. 70, no. 2-4, pp. 496–502, 2008. DOI: 10.1016/j.jastp.2007.08.041.

[120] Y. Kamide, N. Yokoyama, W. Gonzalez, *et al.*, "Two-step development of geomagnetic storms", *Journal of Geophysical Research: Space Physics*, vol. 103, no. A4, pp. 6917–6921, 1998. DOI: 10.1029/97JA03337.

[121] G. Brueckner, J.-P. Delaboudiniere, R. Howard, *et al.*, "Geomagnetic storms caused by coronal mass ejections (CMEs): March 1996 through june 1997", *Geophysical Research Letters*, vol. 25, no. 15, pp. 3019–3022, 1998. DOI: 10.1029/98GL00704.

[122] I. G. Richardson and H. V. Cane, "Solar wind drivers of geomagnetic storms during more than four solar cycles", *J Space Weather Space Clim*, vol. 2, A01, 2012. DOI: 10.1051/swsc/2012001.

[123] J. C. Uwamahoro and J. B. Habarulema, "Modelling total electron content during geomagnetic storm conditions using empirical orthogonal functions and Neural Networks", *Journal of Geophysical Research: Space Physics*, vol. 120, no. 12, 2015. DOI: 10.1002/2015JA021961.

[124] W. D. Gonzalez, J. A. Joselyn, Y. Kamide, *et al.*, "What is a Geomagnetic Storm?", *Journal of Geophysical Research: Space Physics*, vol. 99, no. A4, pp. 5771–5792, 1994. DOI: 10.1029/93JA02867.

[125] E. Doornbos and H. Klinkrad, "Modelling of Space Weather effects on satellite drag", *Adv Space Res*, vol. 37, no. 6, pp. 1229–1239, 2006. DOI: 10.1016/j.asr.2005.04.097.

[126] D. M. Oliveira, E. Zesta, H. Hayakawa, and A. Bhaskar, "Estimating satellite orbital drag during historical magnetic superstorms", *Space Weather*, vol. 18, no. 11, Nov. 2020, ISSN: 1542-7390. DOI: 10.1029/2020sw002472.

[127] G. Ma and T. Maruyama, "A super bubble detected by dense GPS network at east asian longitudes", *Geophysical Research Letters*, vol. 33, no. 21, 2006. DOI: 10.1029/2006GL027512.

[128] E. Astafyeva, Y. Yasyukevich, A. Maksikov, and I. Zhivetiev, "Geomagnetic storms, super-storms, and their impacts on GPS-based navigation systems", *Space Weather*, vol. 12, no. 7, pp. 508–525, 2014. DOI: 10.1002/2014SW001072.

[129] J. Kappenman, "Geomagnetic storms and their impacts on the US power grid", Metatech Corporation, Tech. Rep., 2010. [Online]. Available: https://irp.fas.org/eprint/geomag.pdf.

[130] W. Gonzalez, B. Tsurutani, R. Lepping, and R. Schwenn, "Interplanetary phenomena associated with very intense geomagnetic storms", *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 64, no. 2, pp. 173–181, 2002. DOI: 10.1016/S1364-6826(01)00082-7.

[131] E. Echer, W. Gonzalez, and B. Tsurutani, "Interplanetary conditions leading to superintense geomagnetic storms (Dst $\leq$ 250 nT) during solar cycle 23", *Geophysical Research Letters*, vol. 35, no. 6, 2008. DOI: 10.1029/2007GL031755.

[132] R. Rawat, S. Alex, and G. Lakhina, "Storm-time characteristics of intense geomagnetic storms (Dst$\leq$- 200nT) at low-latitudes and associated energetics", *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 72, no. 18, pp. 1364–1371, 2010. DOI: 10.1016/j.jastp.2010.09.029.

[133] Q. Li, Y. Gao, P. Zhu, H. Chen, and X. Zhang, "Statistical study on great geomagnetic storms during solar cycle 23", *Earthquake Science*, vol. 24, pp. 365–372, 2011. DOI: 10.1007/s11589-011-0799-x.

[134] C. Loewe and G. Prölss, "Classification and mean behavior of magnetic storms", *Journal of Geophysical Research: Space Physics*, vol. 102, no. A7, pp. 14 209–14 213, 1997. DOI: 10.1029/96JA04020.

[135] J. Gosling, D. McComas, J. Phillips, and S. Bame, "Geomagnetic activity associated with Earth passage of interplanetary shock disturbances and coronal mass ejections", *Journal of Geophysical Research: Space Physics*, vol. 96, no. A5, pp. 7831–7839, 1991. DOI: 10.1029/91JA00316.

[136] E. Saiz, A. Guerrero, and C. Cid, "Mid-Latitude Double H-Spikes: Their Properties and Signatures in Different Geomagnetic Indices", *Space Weather*, vol. 21, no. 6, 2023. DOI: 10.1029/2023SW003453.

[137] J. A. Wanliss and K. M. Showalter, "High-resolution global storm index: Dst versus SYM-H", *Journal of Geophysical Research: Space Physics*, vol. 111, no. A2, 2006. DOI: 10.1029/2005JA011034.

[138] J. A. Hutchinson, D. Wright, and S. Milan, "Geomagnetic storms over the last solar cycle: A superposed epoch analysis", *Journal of Geophysical Research: Space Physics*, vol. 116, no. A9, 2011. DOI: 10.1029/2011JA016463.

[139] Y. Y. Shprits, R. Vasile, and I. S. Zhelavskaya, "Nowcasting and Predicting the Kp Index Using Historical Values and Real-Time Observations", *Space Weather*, 2019. DOI: 10.1029/2018sw002141.

[140] L. Cai, S. Ma, H. Cai, Y. Zhou, and R. Liu, "Prediction of SYM-H index by NARX Neural Network from IMF and solar wind data", *Sci China Technol Sci*, vol. 52, pp. 2877–2885, 2009. DOI: 10.1007/s11431-009-0296-9.

[141] N. Jiang, "IFRS 17: Risk adjustment a numerical example", Society of Actuaries, Tech. Rep., 2020. [Online]. Available: https://www.soa.org/globalassets/assets/library/newsletters/financial-reporter/2020/may/fr-2020-iss-05-20-jiang.pdf.

[142] K. R. Murphy, C. Watt, I. R. Mann, *et al.*, "The global statistical response of the outer radiation belt during geomagnetic storms", *Geophysical Research Letters*, vol. 45, no. 9, pp. 3783–3792, 2018. DOI: 10.1002/2017GL076674.

[143] J. Aguado, C. Cid, E. Saiz, and Y. Cerrato, "Hyperbolic decay of the Dst Index during the recovery phase of intense geomagnetic storms", *Journal of Geophysical Research: Space Physics*, vol. 115, no. A7, Jul. 2010, ISSN: 0148-0227. DOI: 10.1029/2009ja014658.

[144] A. Mannucci, B. Tsurutani, M. Abdu, *et al.*, "Superposed epoch analysis of the dayside ionospheric response to four intense geomagnetic storms", *Journal of Geophysical Research: Space Physics*, vol. 113, no. A3, 2008. DOI: 10.1029/2007JA012732.

[145] S. Wharton, I. Rae, J. Sandhu, M.-T. Walach, D. Wright, and T. Yeoman, "The changing eigenfrequency continuum during geomagnetic storms: Implications for plasma mass dynamics and ulf wave coupling", *Journal of Geophysical Research: Space Physics*, vol. 125, no. 6, e2019JA027648, 2020. DOI: 10.1029/2019JA027648.

[146] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost", *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012. DOI: 10.1080/01621459.2012.737745.

[147] J. P. Thayer, J. Lei, J. M. Forbes, E. K. Sutton, and R. S. Nerem, "Thermospheric density oscillations due to periodic solar wind high-speed streams", *Journal of Geophysical Research: Space Physics*, vol. 113, no. A6, 2008. DOI: 10.1029/2008JA013190.

[148] M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, and A. Mukhopadhyay, "Rmse is not enough: Guidelines to robust data-model comparisons for magnetospheric physics", *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 218, 2021. DOI: 10.1016/j.jastp.2021.105624.

[149] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, Available: http://www.deeplearningbook.org.

[150] J. E. Borovsky and M. H. Denton, "Differences between CME-driven storms and CIR-driven storms", *Journal of Geophysical Research: Space Physics*, vol. 111, no. A7, 2006. DOI: 10.1029/2005JA011447.

[151] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy", *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006. DOI: 10.1016/j.ijforecast.2006.03.001.

[152] M. Fernandes, R. Mendes, S. M. Vieira, *et al.*, "Predicting Intensive Care Unit admission among patients presenting to the emergency department using Machine Learning and Natural Language Processing", *PloS one*, vol. 15, no. 3, 2020. DOI: 10.1371/journal.pone.0229331.

[153] J.-E. Lee, J. H. Kim, J.-H. Bae, I. Song, and J.-Y. Shin, "Detecting early safety signals of infliximab using Machine Learning algorithms in the Korea adverse event reporting system", *Scientific Reports*, vol. 12, no. 1, p. 14 869, 2022. DOI: 10.1038/s41598-022-18522-z.

[154] C. Larrodera and M. Temmer, *Interplanetary coronal mass ejections multi-catalog*, 2023. DOI: 10.21950/XGUIYX.

[155] L. Cai, S. Y. Ma, and Y. L. Zhou, "Prediction of SYM-H index during large storms by NARX neural network from IMF and solar wind data", *Annales Geophysicae*, vol. 28, no. 2, pp. 381–393, Feb. 2010. DOI: 10.5194/angeo-28-381-2010.

[156] R. Koenker and G. Bassett, "Regression quantiles", *Econometrica*, vol. 46, no. 1, p. 33, Jan. 1978, ISSN: 0012-9682. DOI: 10.2307/1913643.

[157] Z. Cai, "Regression quantiles for time series", *Econometric theory*, vol. 18, no. 1, pp. 169–192, 2002. [Online]. Available: https://www.jstor.org/stable/3533031.

[158] C. Granger, H. White, and M. Kamstra, "Interval forecasting", *Journal of Econometrics*, vol. 40, no. 1, pp. 87–96, Jan. 1989, ISSN: 0304-4076. DOI: 10.1016/0304-4076(89)90031-6.

[159] N. Meinshausen, "Quantile regression forests", *Journal of Machine Learning Research*, vol. 7, no. 35, pp. 983–999, 2006. [Online]. Available: http://jmlr.org/papers/v7/meinshausen06a.html.

[160] J. W. Taylor, "A quantile regression Neural Network approach to estimating the conditional density of multiperiod returns", *Journal of Forecasting*, vol. 19, no. 4, pp. 299–311, 2000, ISSN: 1099-131X. DOI: 10.1002/1099-131x(200007)19:4<299::aid-for775>3.0.co;2-v.

[161] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, *A Multi-Horizon Quantile Recurrent Forecaster*, 2017. DOI: 10.48550/ARXIV.1711.11053.

[162] C. Wan, J. Lin, J. Wang, Y. Song, and Z. Y. Dong, "Direct quantile regression for nonparametric probabilistic forecasting of wind power generation", *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2767–2778, Jul. 2017, ISSN: 1558-0679. DOI: 10.1109/tpwrs.2016.2625101.

[163] G. Papacharalampous and A. Langousis, "Probabilistic Water Demand Forecasting Using Quantile Regression Algorithms", *Water Resources Research*, vol. 58, no. 6, Jun. 2022, ISSN: 1944-7973. DOI: 10.1029/2021wr030216.

[164] D. Pradeepkumar and V. Ravi, "Forecasting financial time series volatility using Particle Swarm Optimization trained Quantile Regression Neural Network", *Applied Soft Computing*, vol. 58, pp. 35–52, Sep. 2017, ISSN: 1568-4946. DOI: 10.1016/j.asoc.2017.04.014.

[165] A. Zarnani, S. Karimi, and P. Musilek, "Quantile Regression and Clustering Models of Prediction Intervals for Weather Forecasts: A Comparative Study", *Forecasting*, vol. 1, no. 1, pp. 169–188, Oct. 2019, ISSN: 2571-9394. DOI: 10.3390/forecast1010012.

[166] H. Afifi, M. Elmahdy, M. E. Saban, and M. Abu-Elkheir, "Probabilistic Time Series Forecasting for Unconventional Oil and Gas Producing Wells", in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference*, IEEE, Oct. 2020. DOI: 10.1109/niles50944.2020.9257962.

[167] S. Elvidge, H. C. Godinez, and M. J. Angling, "Improved forecasting of thermospheric densities using multi-model ensembles", *Geoscientific Model Development*, vol. 9, no. 6, pp. 2279–2292, 2016. DOI: 10.5194/gmd-9-2279-2016.

[168] I. Steinwart and A. Christmann, "Estimating conditional quantiles with the help of the pinball loss", *Bernoulli*, vol. 17, no. 1, pp. 211–225, 2011. DOI: 10.3150/10-BEJ267.

[169] H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts", *Wind Energy*, vol. 9, no. 1-2, pp. 95–108, Dec. 2005, ISSN: 1099-1824. DOI: 10.1002/we.180.

[170] J. R. Trapero, M. Cardós, and N. Kourentzes, "Quantile forecast optimal combination to enhance safety stock estimation", *International Journal of Forecasting*, vol. 35, no. 1, pp. 239–250, Jan. 2019, ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2018.05.009.

[171] M. Abdar, F. Pourpanah, S. Hussain, *et al.*, "A review of uncertainty quantification in Deep Learning: Techniques, applications and challenges", *Information Fusion*, vol. 76, pp. 243–297, Dec. 2021, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.05.008.

[172] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach", in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 4075–4084. [Online]. Available: https://proceedings.mlr.press/v80/pearce18a.html.

[173] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C. van Lint, "Prediction intervals to account for uncertainties in travel time prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 537–547, Jun. 2011, ISSN: 1558-0016. DOI: 10.1109/tits.2011.2106209.

[174] J. Pang, D. Liu, Y. Peng, and X. Peng, "Optimize the coverage probability of prediction interval for anomaly detection of sensor-based monitoring series", *Sensors*, vol. 18, no. 4, p. 967, Mar. 2018, ISSN: 1424-8220. DOI: 10.3390/s18040967.

[175] Y. Wu, Y. Ye, A. Zeb, J. J. Yu, and Z. Wang, "Adaptive modeling of uncertainties for traffic forecasting", *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2024, ISSN: 1558-0016. DOI: 10.1109/tits.2023.3327100.

[176] ESA SSA Team, "Space Situational Awareness - Space Weather PSD", European Space Agency, Tech. Rep., Jul. 2013, https://swe.ssa.esa.int/DOCS/SSA-SWE/SSA-SWE-RS-SSD-0001_i1r3.pdf.

[177] M. Madelaire, K. M. Laundal, J. P. Reistad, S. M. Hatch, A. Ohma, and S. Haaland, "Geomagnetic response to rapid increases in solar wind dynamic pressure: Event detection and large scale response", *Frontiers in Astronomy and Space Sciences*, vol. 9, May 2022, ISSN: 2296-987X. DOI: 10.3389/fspas.2022.904620.

[178] K. K. Hashimoto, T. Kikuchi, I. Tomizawa, *et al.*, "Penetration electric fields observed at middle and low latitudes during the 22 June 2015 geomagnetic storm", *Earth, Planets and Space*, vol. 72, no. 1, May 2020, ISSN: 1880-5981. DOI: 10.1186/s40623-020-01196-0.

[179] S.-I. Akasofu, "A review of studies of geomagnetic storms and auroral/magnetospheric substorms based on the electric current approach", *Frontiers in Astronomy and Space Sciences*, vol. 7, Jan. 2021, ISSN: 2296-987X. DOI: 10.3389/fspas.2020.604750.

[180] I. Michaelis, K. Styp-Rekowski, J. Rauberg, C. Stolle, and M. Korte, "Geomagnetic data from the GOCE satellite mission", *Earth, Planets and Space*, vol. 74, no. 1, Sep. 2022, ISSN: 1880-5981. DOI: 10.1186/s40623-022-01691-6.

[181] S.-i. Oyama, A. Aikio, T. Sakanoi, *et al.*, "Geomagnetic activity dependence and dawn-dusk asymmetry of thermospheric winds from 9-year measurements with a fabry-perot interferometer in tromsø, norway", *Earth, Planets and Space*, vol. 75, no. 1, May 2023, ISSN: 1880-5981. DOI: 10.1186/s40623-023-01829-0.

[182] G. S. Lakhina and B. T. Tsurutani, "Geomagnetic storms: Historical perspective to modern view", *Geoscience Letters*, vol. 3, no. 1, Feb. 2016, ISSN: 2196-4092. DOI: 10.1186/s40562-016-0037-4.

[183] E. M. H. Takla and A. A. Khashaba, "Investigation of low-latitude nighttime geomagnetic pulsation events occurred under variable IMF and solar conditions", *Terrestrial, Atmospheric and Oceanic Sciences*, vol. 35, no. 1, Aug. 2024, ISSN: 2311-7680. DOI: 10.1007/s44195-024-00071-9.