United Nations Workshop on
the International Space Weather Initiative:
The Way Forward

26 – 30 June 2023
Vienna

Ionospheric modelling using machine learning towards space weather operational service

María Graciela Molina
gmolina@herrera.unt.edu.ar
FACET-UNT/CONICET/INGV

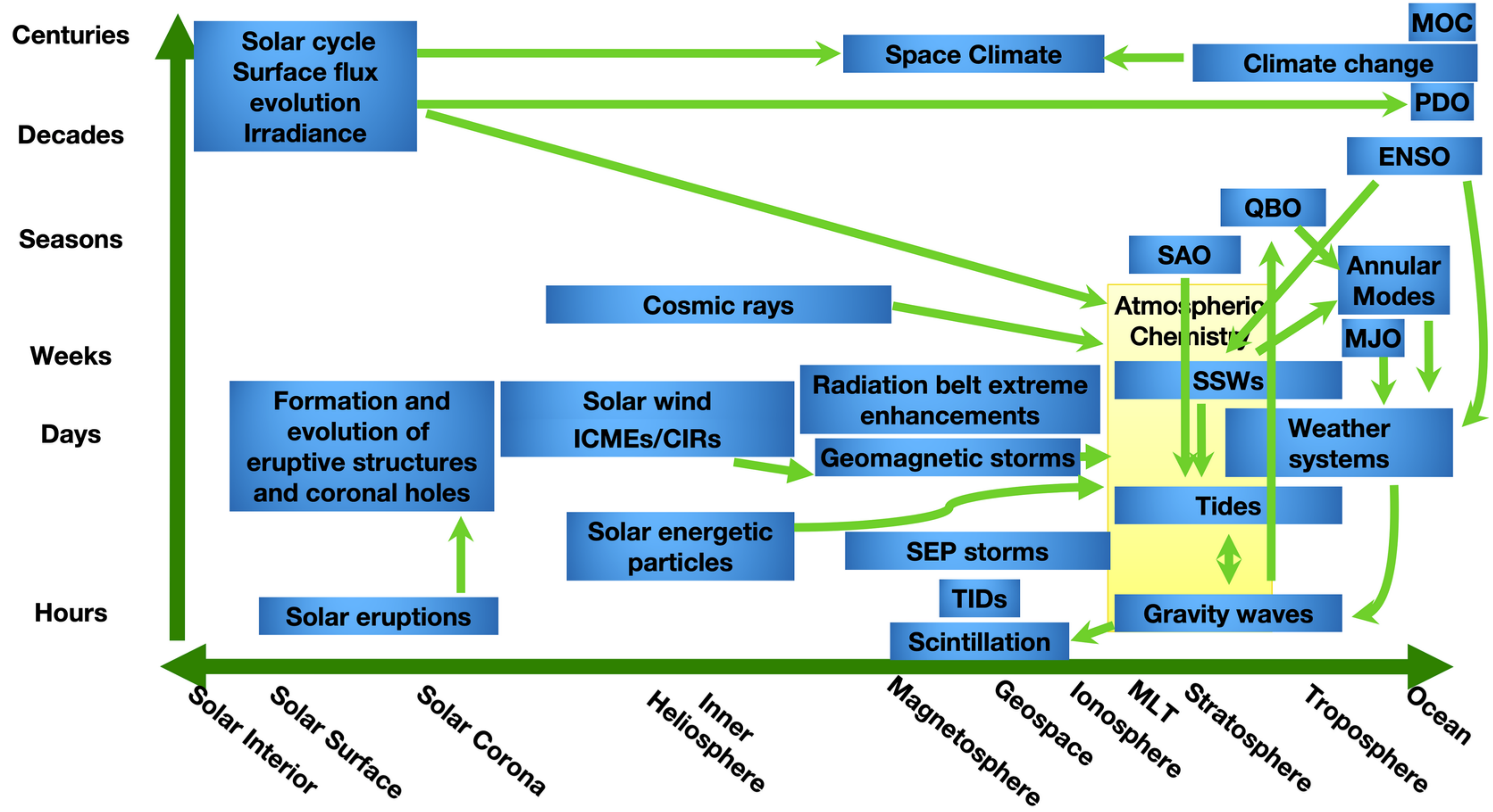
Outline

Disclaimer: I will go back to many concepts and remarks from yesterday's talks!

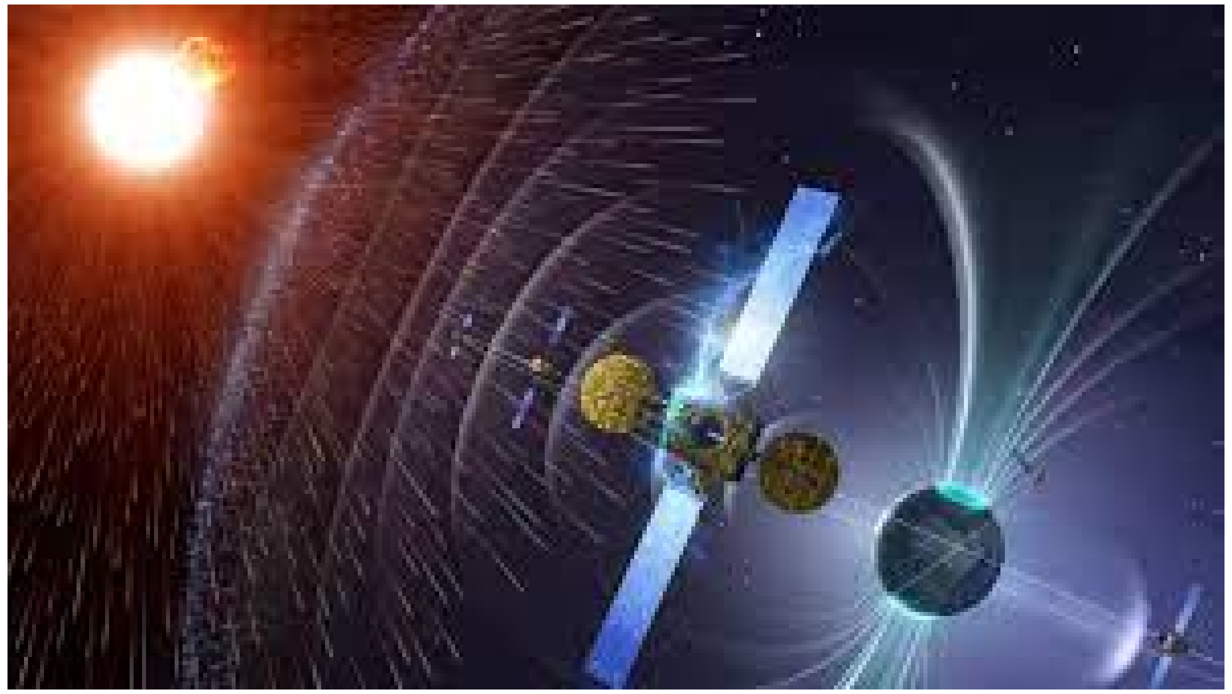
Introduction

SWx

An integrated view of solar-terrestrial prediction
Solar-Terrestrial phenomena in various spatial & temporal scales



PRESTO/SCOSTEP



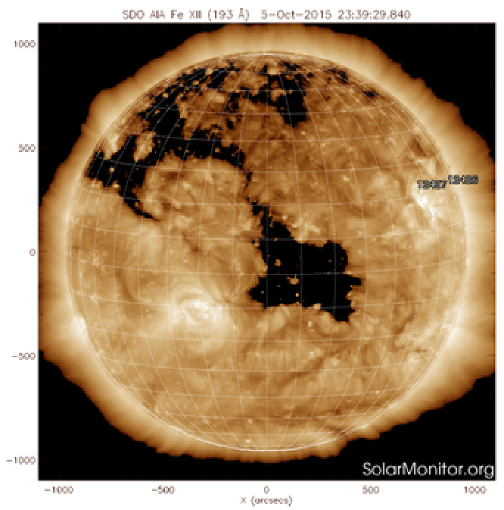
- Complex and highly coupled system
- Time/spatial scales
- Intrinsically unbalanced problem
- Difficult to model (e.g. too expensive to run physical models,)

. . .

(*) yesterday's talk by Sandro

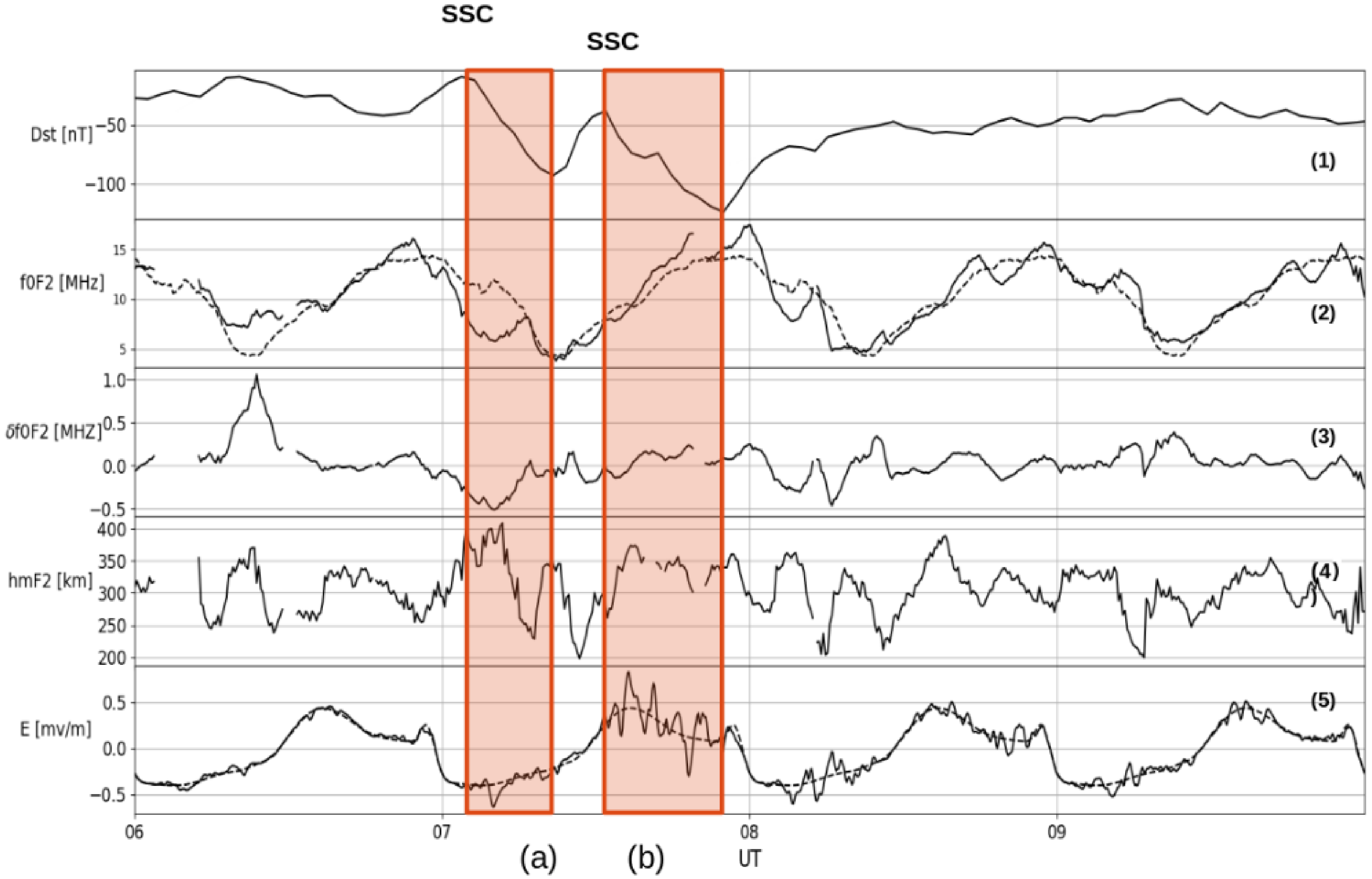
Introduction

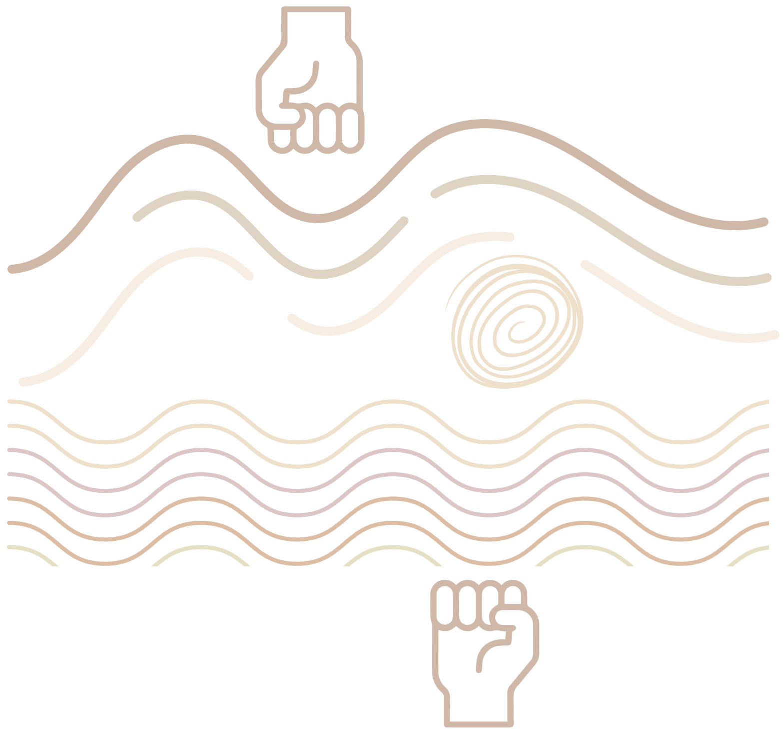
Iospheric response



HSS/CIR
Kp=7
7 October 2015

Molina +, 2020





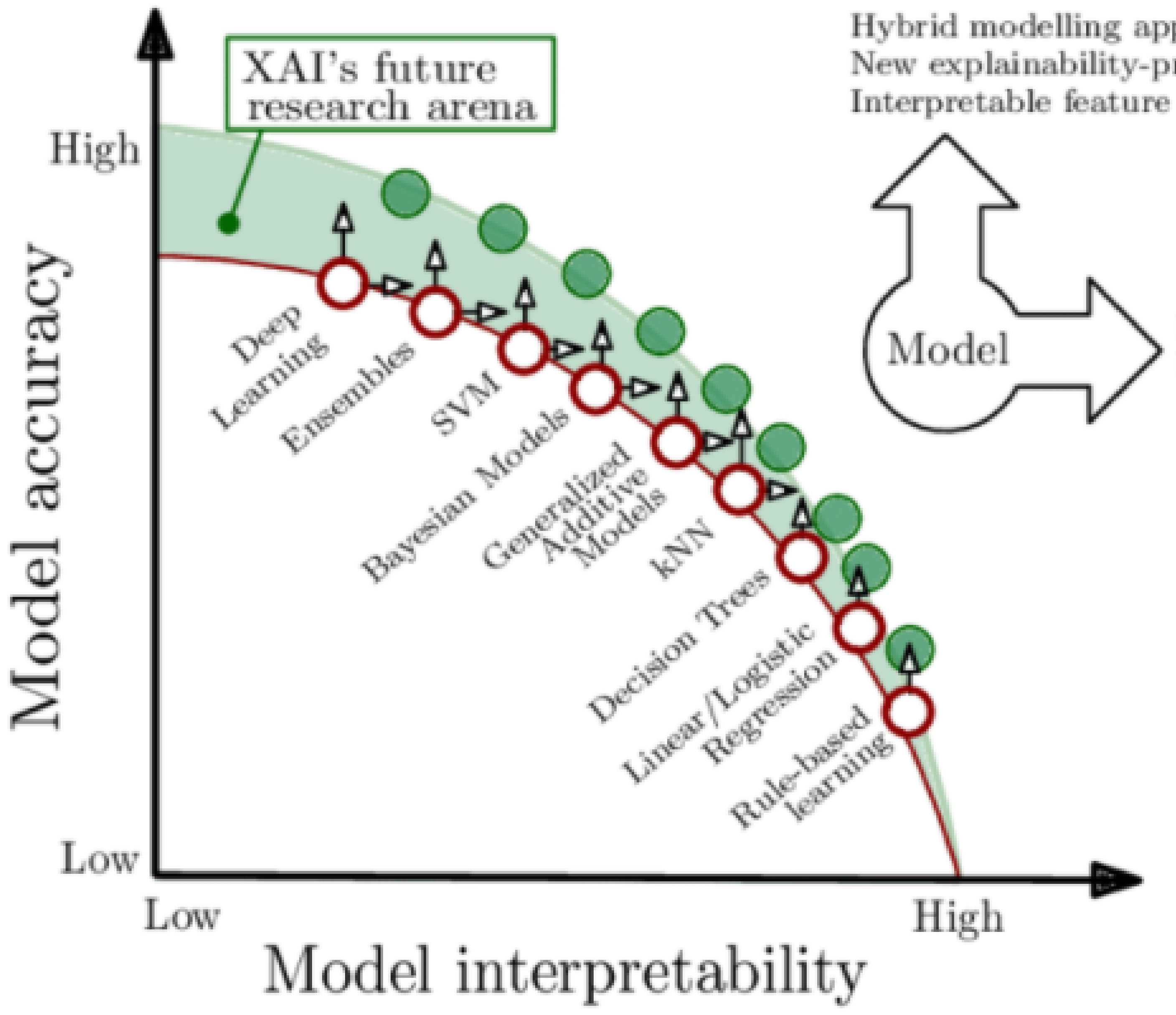
- Difficult to forecast the impact!
- Regular variability (solar cycles, daily, etc) + Irregularities (e.g. SWx)
- Global - Regional - Local (different scales, different problems). -> systemic view: all together!
- Instruments deployment (ground and space-based)

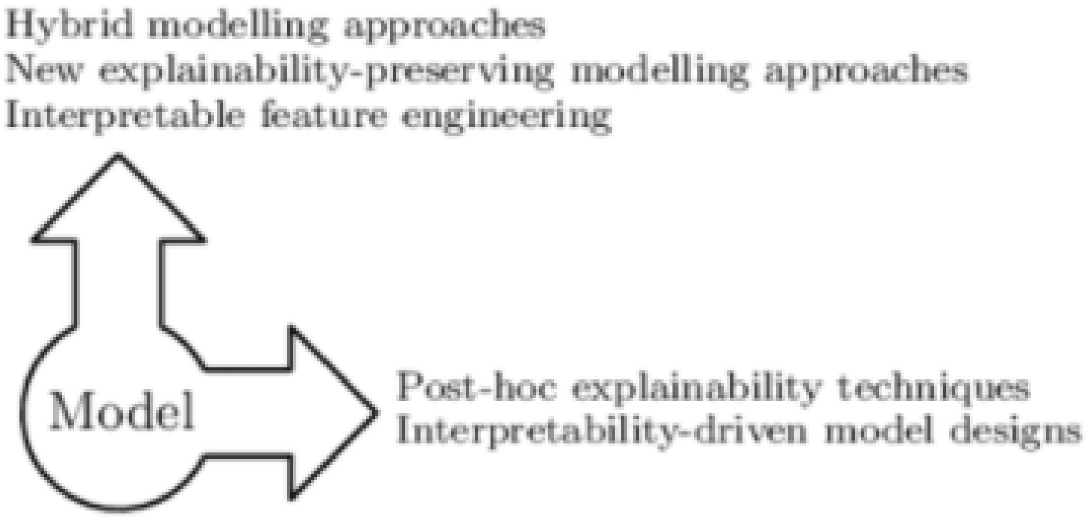
What about the data

- Huge amount of heterogeneous data
- Data availability (?) -> particularly in R2O
- Data quality:
  - high quality = science; less quality = operations; levels of pre-processing
  - In ML: Better no data than bad data (*).
  - Understand the data-> e.g. calibrated TEC derived from GNSS (*)
- Data covers partially the domain
- Integration & interoperability:
  - Formating madness! resolution madness!
  - Produced by instruments, interpreters/forecasters, simulations or models, metadata (No standard data model)
- Not straightforward to understand (learn your physics!)
- Data preparation is expensive

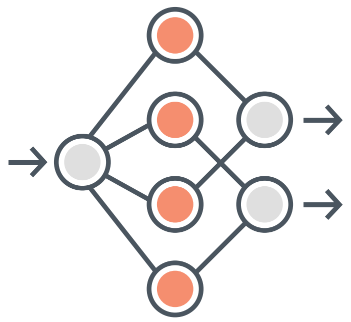
What about ML

data-driven modelling





Barredo +, 2019



Output

Linear combination of inputs

$$\hat{y} = g\left(w_0 + \sum_{i=1}^{m} x_i\, w_i\right)$$

Non-linear activation function

Bias

- No generalization
- Easy to implement (+ toolboxes, better hw) - > not easy to adapt
- White - grey - black box
- More predictive capabilities, less interpretability (DL) - > XAI methods
- We need + robust/mature algorithms

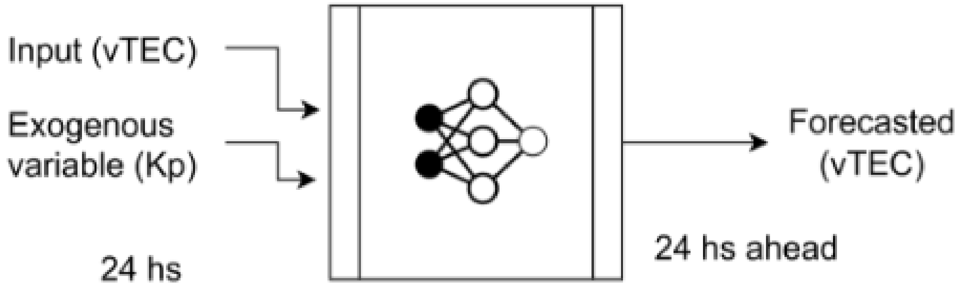
An application

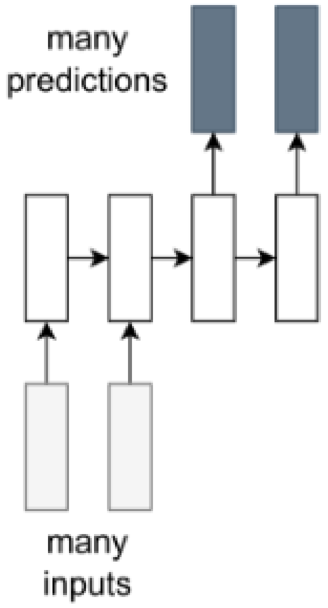
Molina +(submitted)

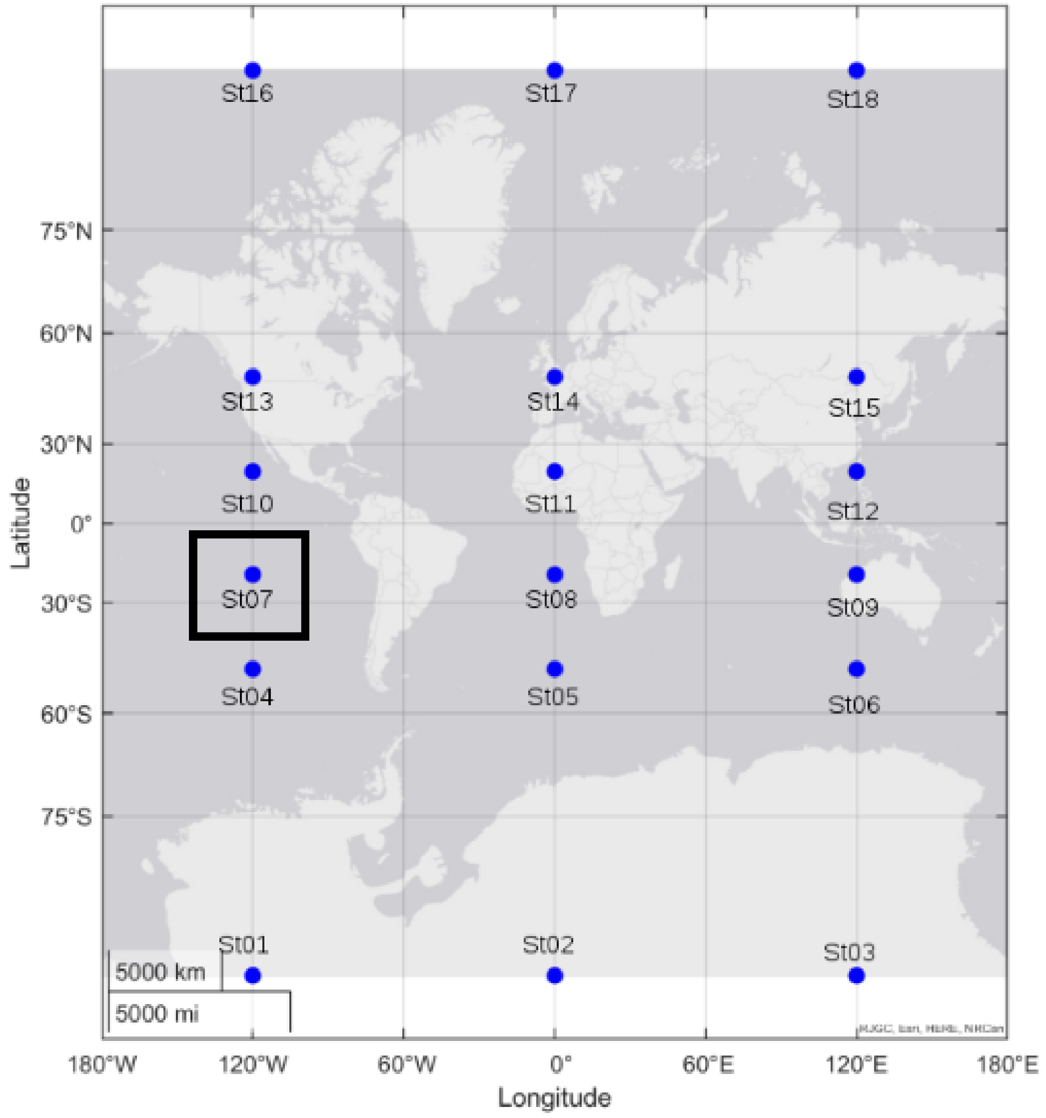
Objectives:

- Global TEC forecast 24 hs ahead using DL
- Propose a semi-operative prototype

- 2 stages: **a) single station forecasting (ML)**; b) extended forecasting
- 3 meridional sectors covering low, mid & high latitude
- Covering land & oceanic regions
- Input: TEC from GIMs + External input (Kp)







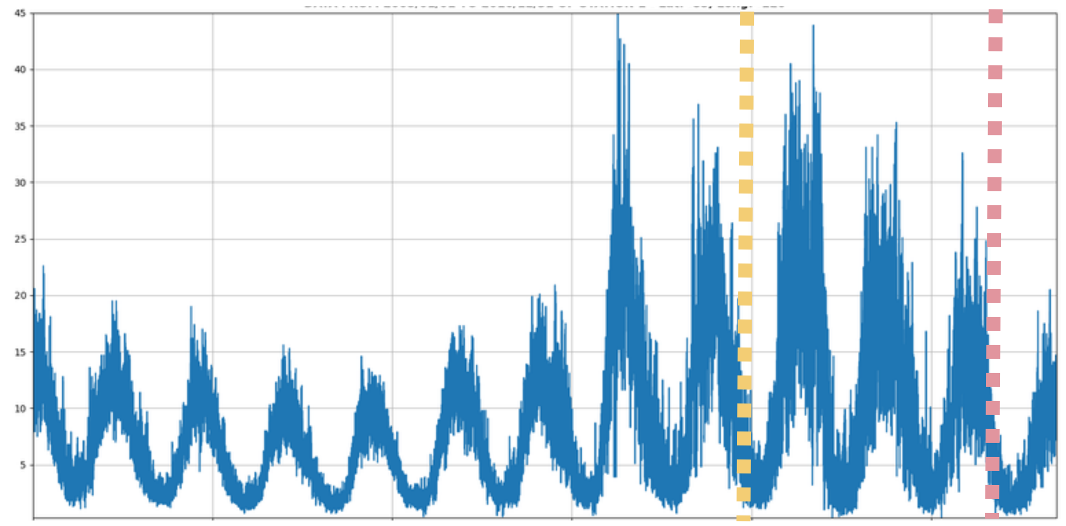
Cesaroni +, 2020

Data preparation & Feature selection

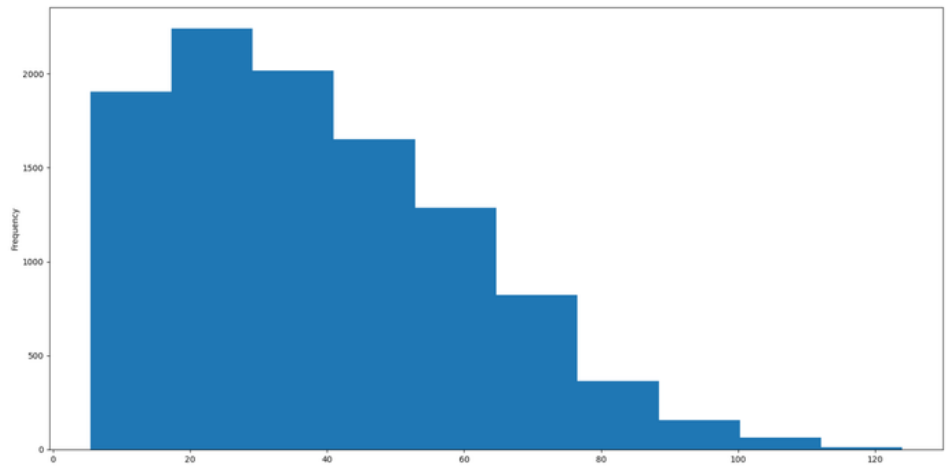
Loosely physics-informed approach

- Dataset:
  - 2005 - 2016
  - splitting strategy: 99% (99 train/1val) - 1% test (~43 days)
  - + cases study: geomagnetic storms in 2017
- Resolution (re-sampling):
  - TEC from GIMs - 2 hs resolution
  - Kp - 3hs resolution > K Nearest-neighbor interpolation
- Smart weight initialization (kernel initialization): GlorotNormal distribution + proper activation function (e.g. tanh).

St 01 TEC - dataset (2005 - 2016)

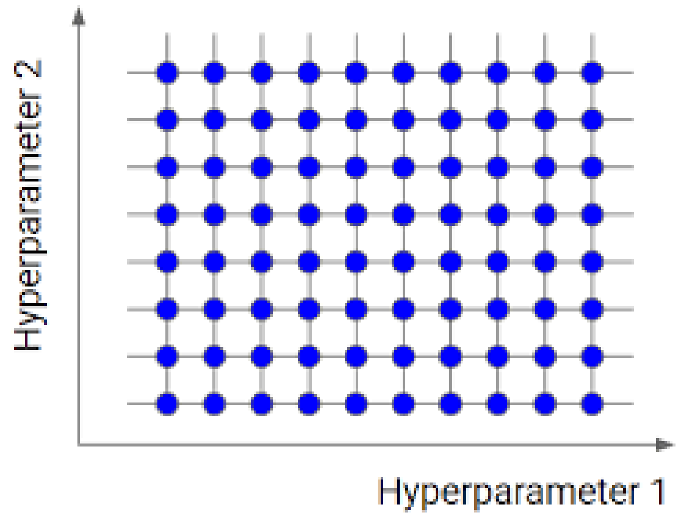


TEC - single ST Histogram

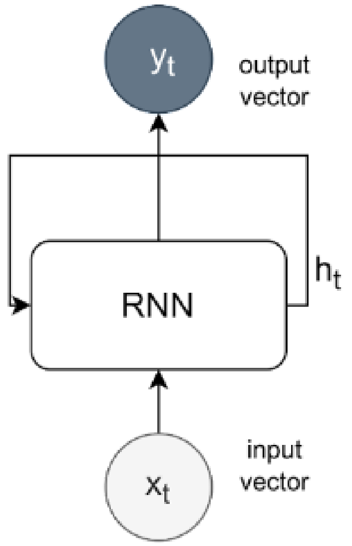


ML modelling

- 3 ML techniques:
  - 2 RNNs (LSTM & GRU)
  - CNN (1D)
- Time series
- Hyperparameter tuning:
  grid search



- # hidden layers (5,10,15,20,50,100 cells)
- batch size (16,32,64,128)
- #epochs (iterations) (5,10,15,20,30,40,50,100,200,500,1000)

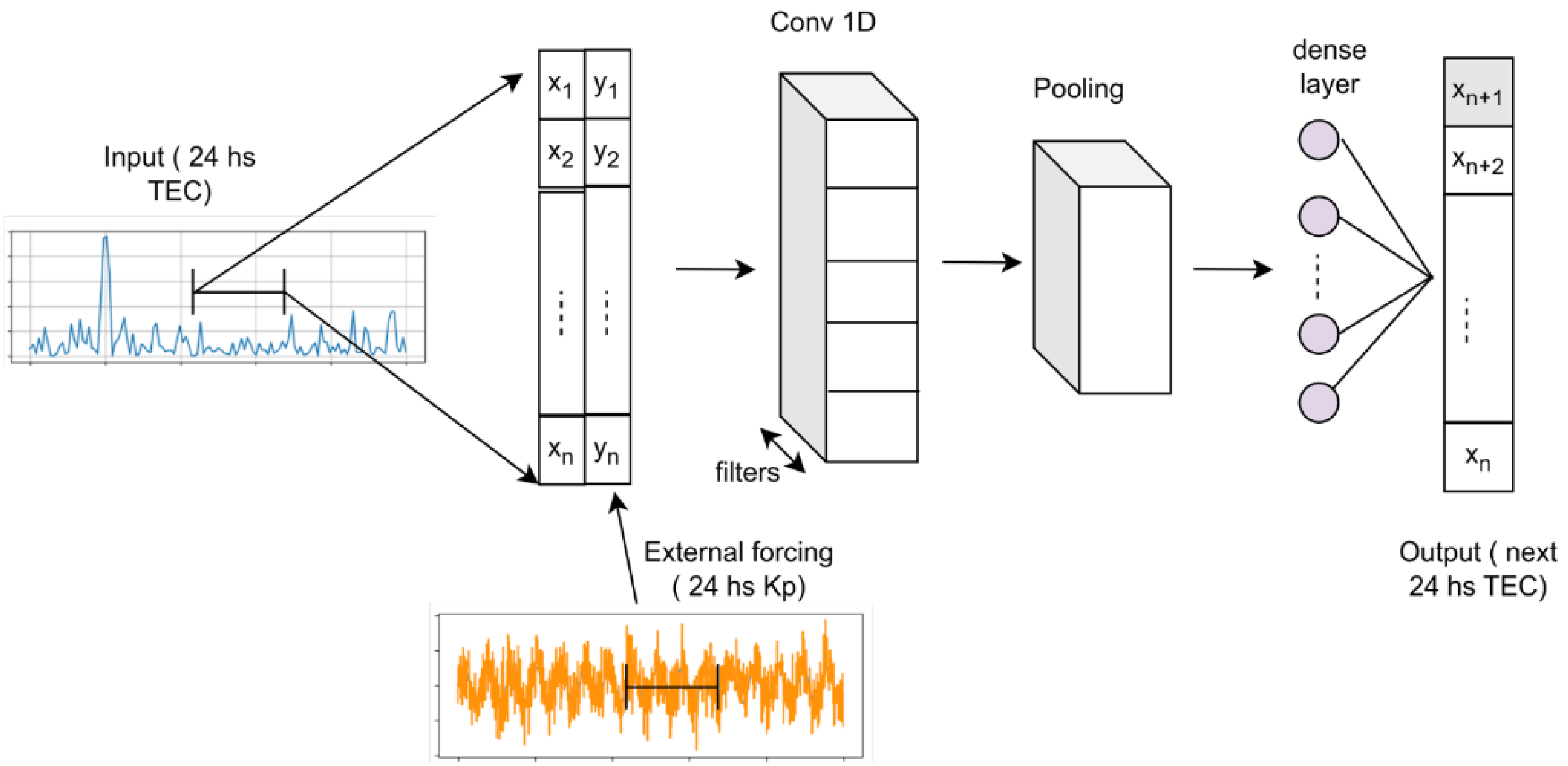


RNNs:

- Maintain order
- Memory (ht)
- Backpropagation through time
- Prone to overfitting, vanishing gradient problem
- LSTM & GRU -> gated cells -> long-term but not that long

CNN:

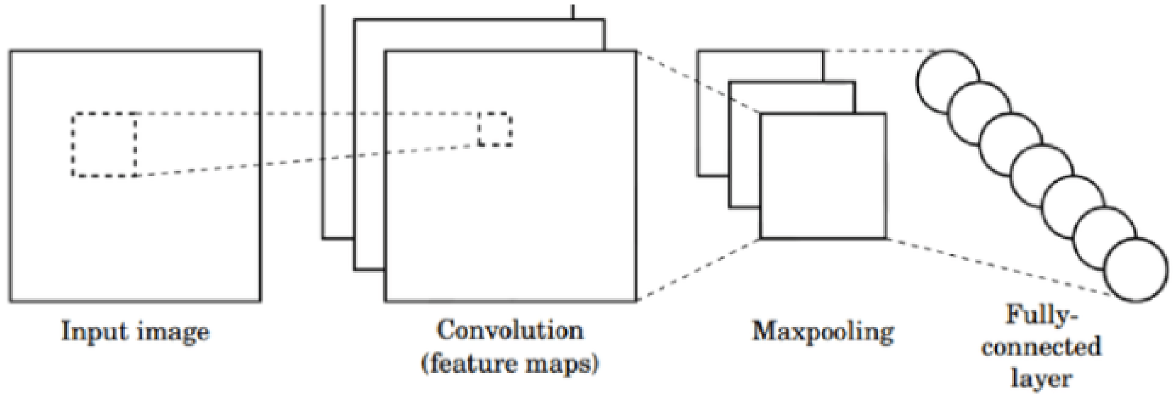
- kernel size = 2

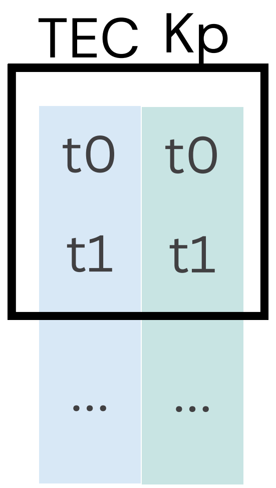


ML modelling

- Forecasting 24 hs ahead (quiet day)
- RMSE < 3 TECu
- CNN best at any station (– St16,17,18 -> TECu<=1 -> quiet day)
- Low lat + oceanic stations -> + challenging

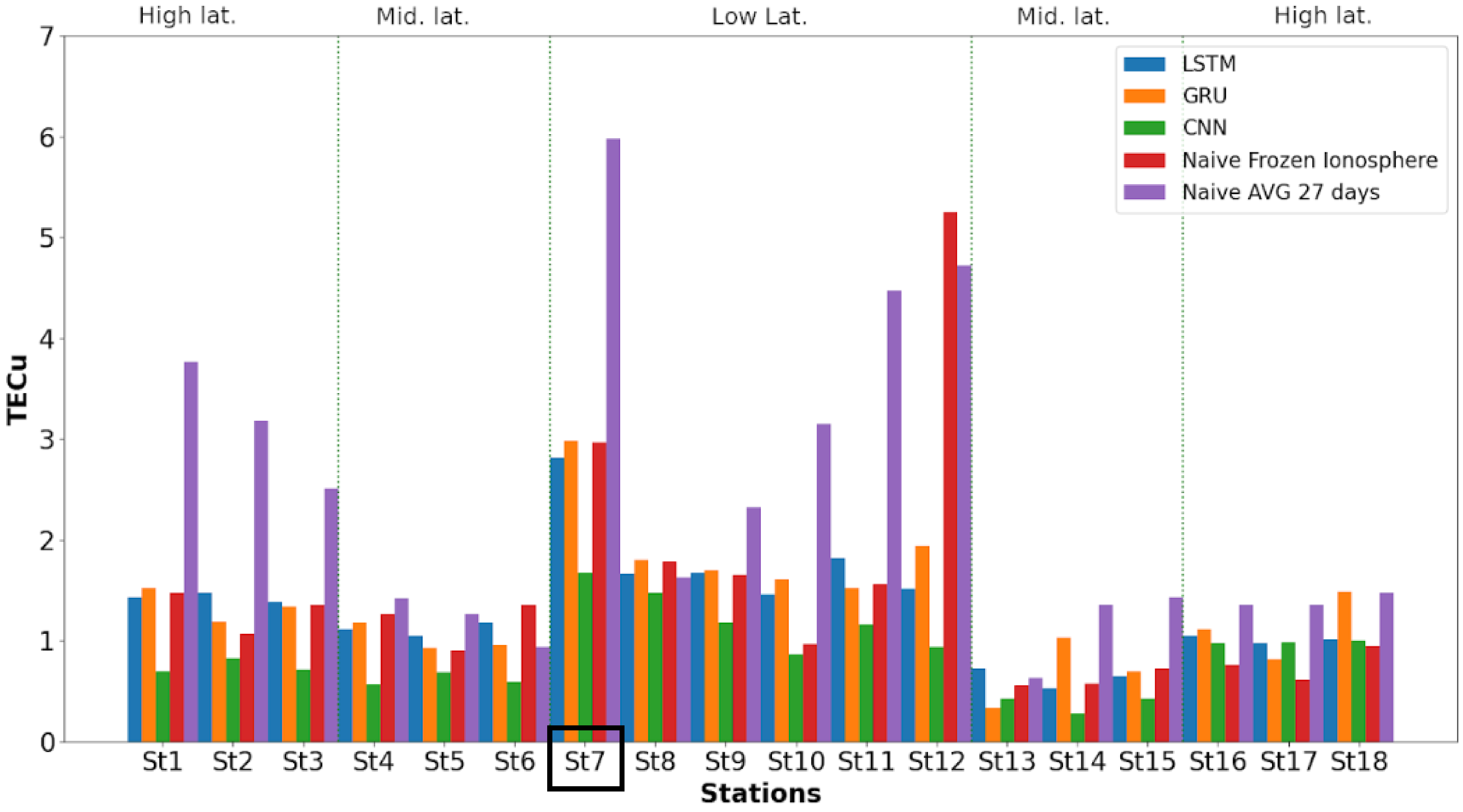
- Why these results?
    - LSTM & GRU -> difficult to catch fast changes and peaks
    - CNN (1D) -> spatial relationship = short term relationships





Kernel = 2

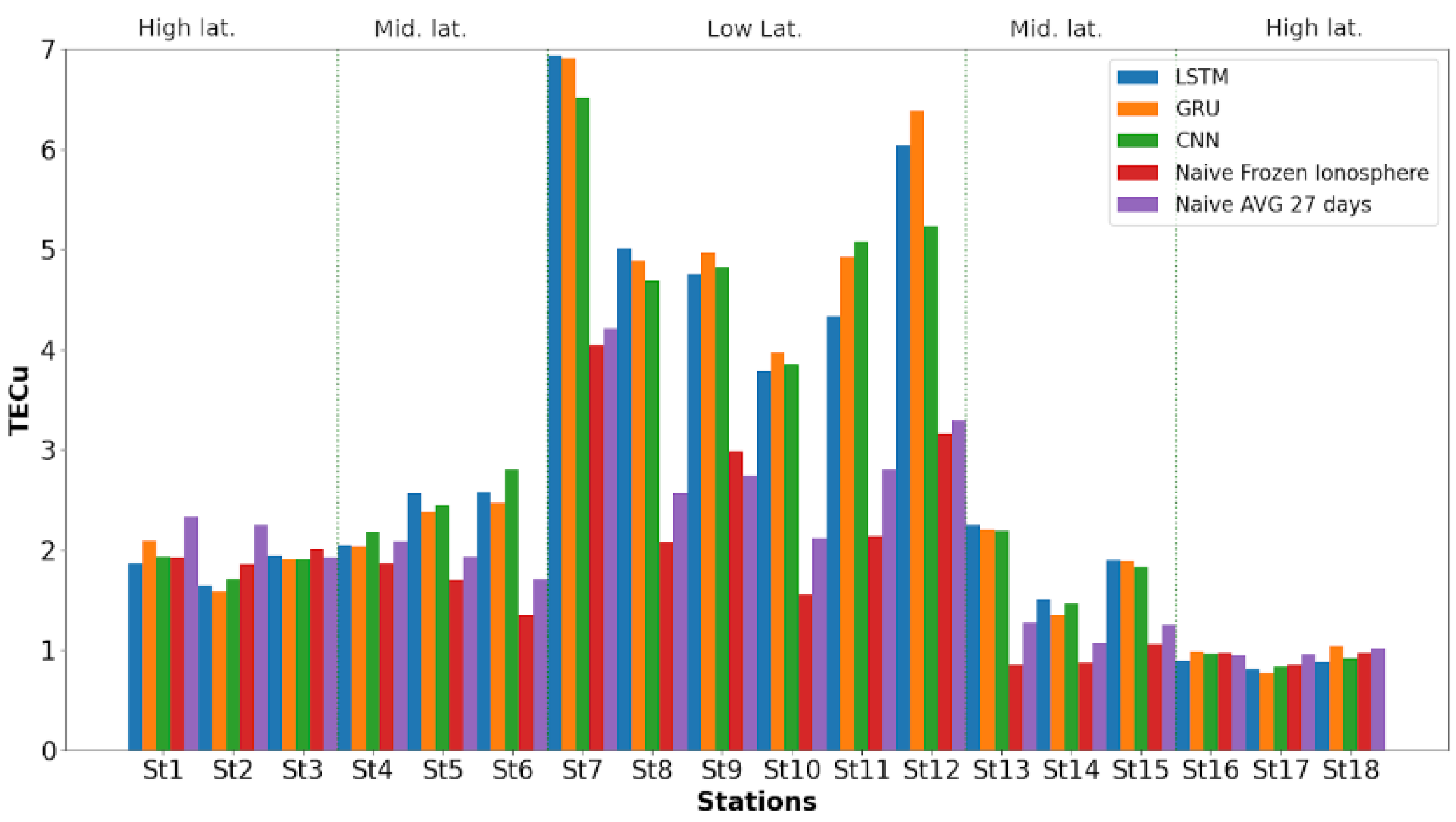
RMSE



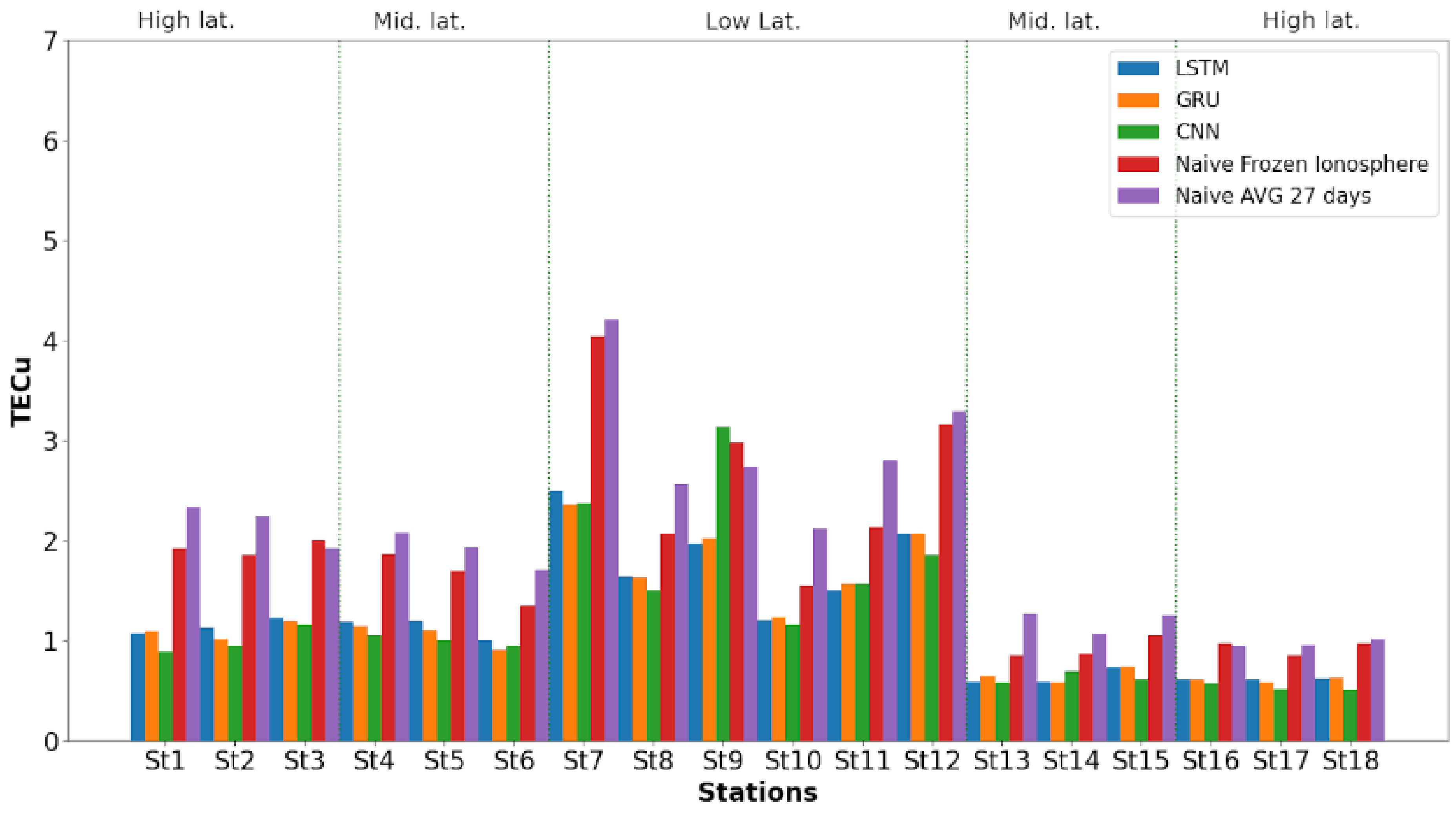
ML modelling

- In general: in SWx, few extreme cases (unbalanced datasets) -> forecasting may fail when new data arrives (generalization is a problem)-> Incremental learning

RMSE



Test set -> 43 days wiht the basic models

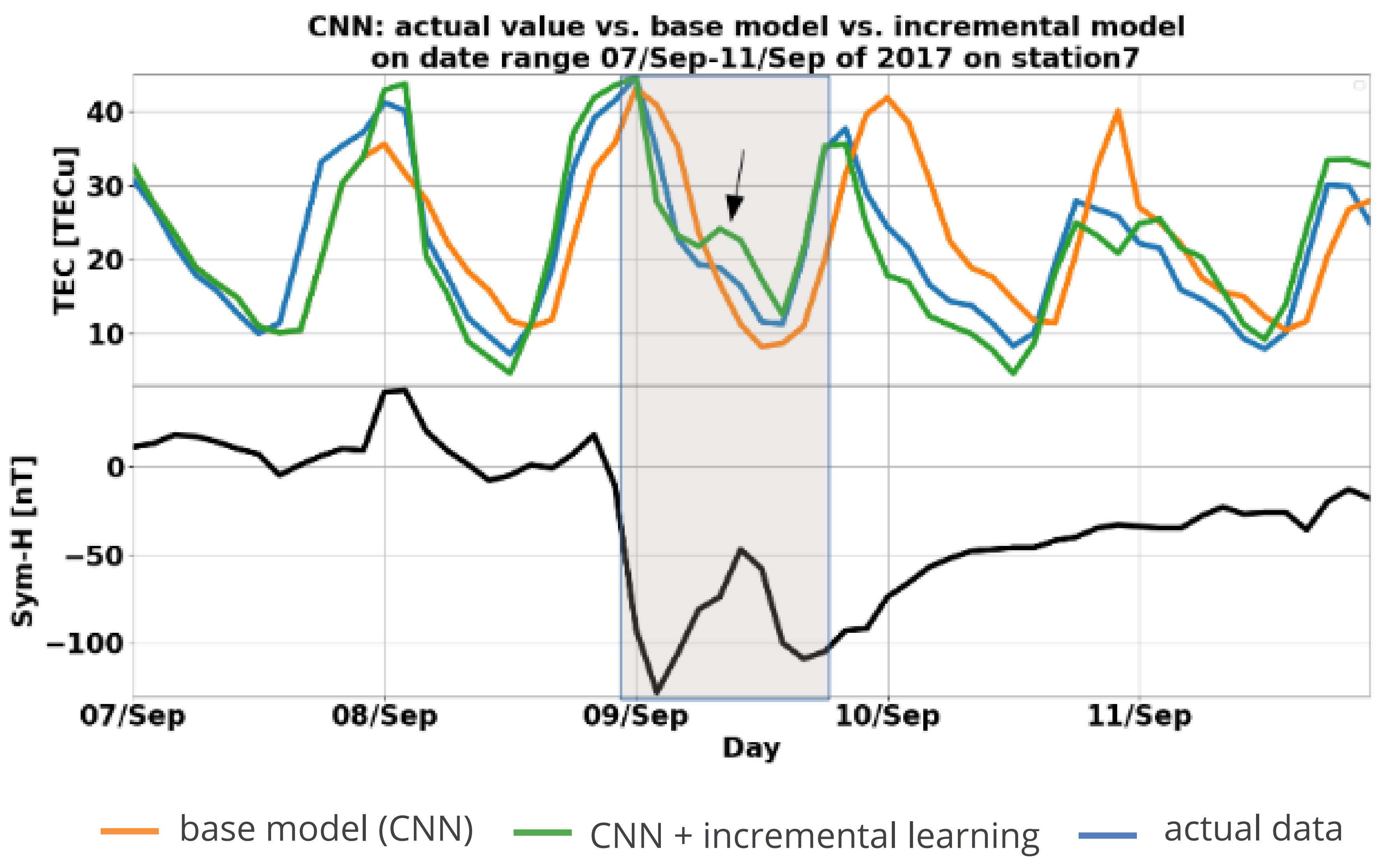


Test set -> 43 days with the models + incremental learning (updating each 24 hs)

ML modelling

$$Global\ \Delta TEC = \frac{1}{st}\sum^{st} \Delta TEC$$

- We considered cases study from 2017 under different geomagnetic conditions

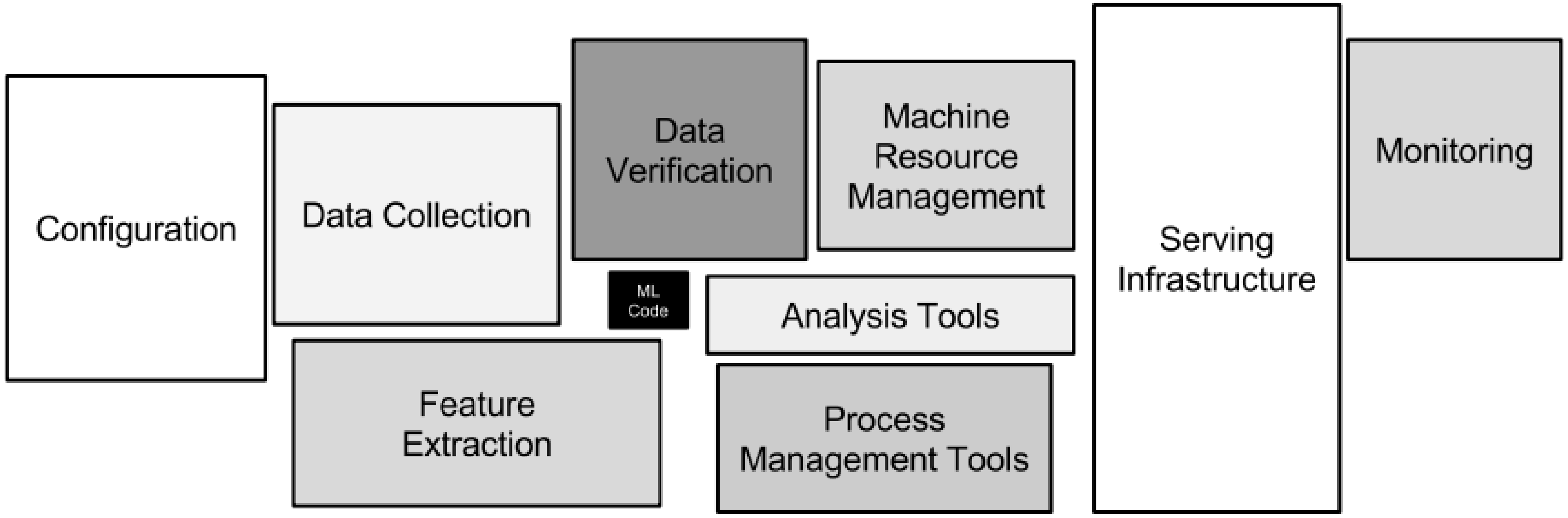


Molina +(submitted)

R2O

Considerations

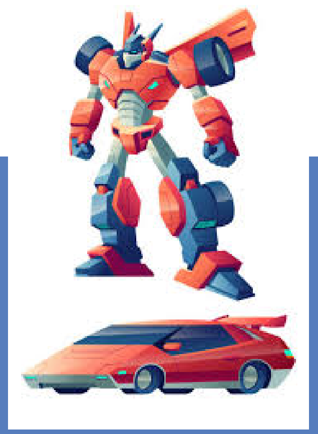
Hidden Technical Debt in Machine Learning Systems, D. Sculley et.al (2015)



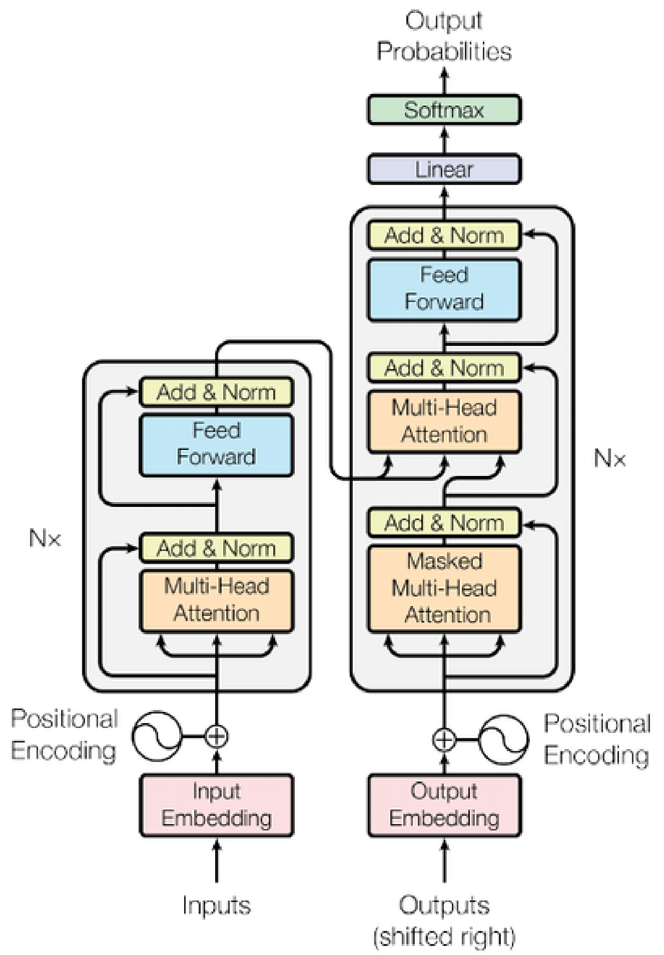
- The modelling is just a small part of an operational system

- Software development -> in production
- **Trustworthiness** is key (e.g. uncertainty quantification)
- Better **data quality and real-time** data
- Better feature selection/engineering (e.g.. choose wisely the geomagnetic index, etc)
- Data enhancement/ surrogate data
- The most expensive and time-consuming stage is data preparation -> we need inter-operational data
- Continuous monitoring and validation

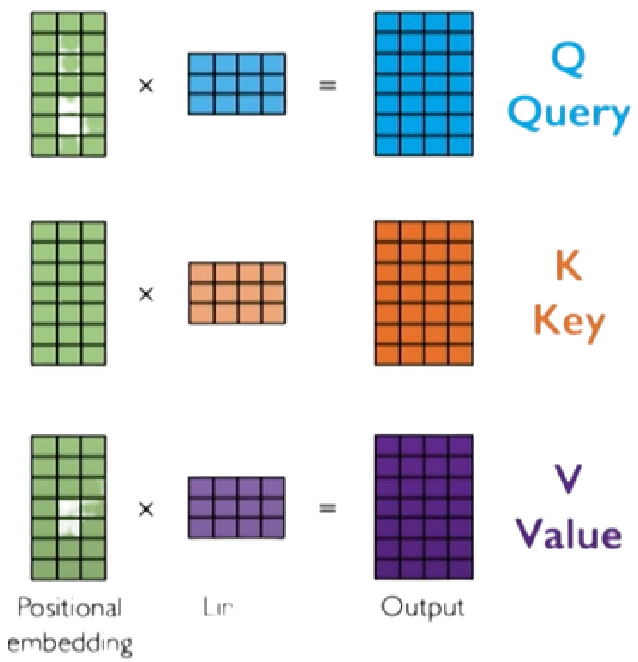
Next steps

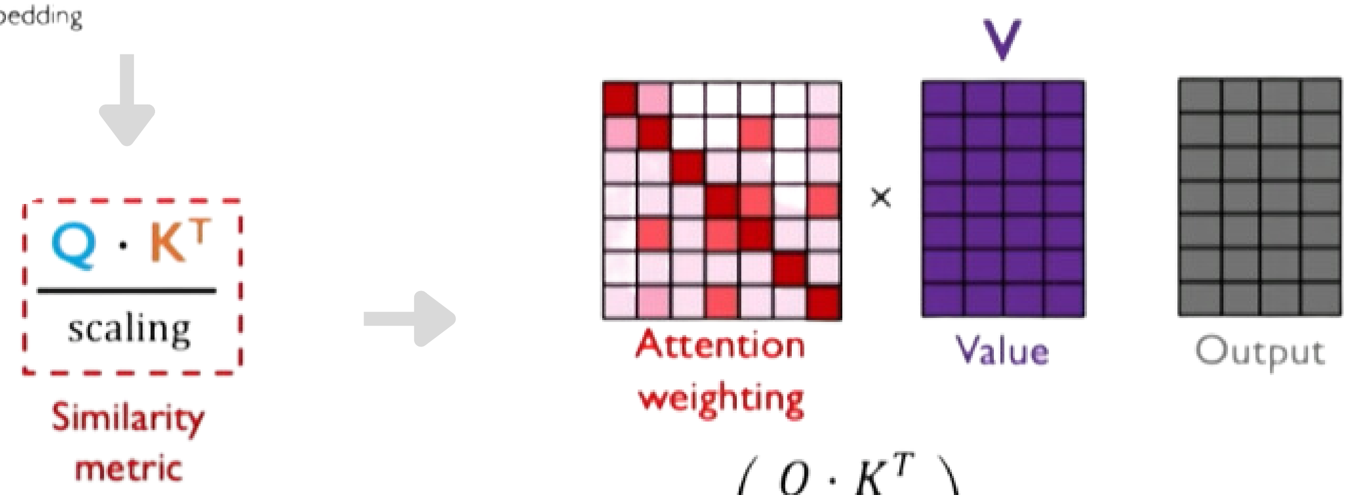


- Catch fast and far information (different scales)
- "Attend" to the more influential features within the data



Vaswani +, 2017





$$softmax\left(\frac{Q \cdot K^T}{scaling}\right) \cdot V = A(Q,K,V)$$

Self-attention-based models (transformers):

- More computationally efficient
- Eliminates recurrence -> positional encoding
- Multi-head -> different scales

Conclusion

- 3 techniques (LSTM, GRU and CNN): CNN obtain better performance and is able to catch fast changes within the time series even during geomagnetic storms.

- Considerations for operative implementation: Incremental learning

- Still, many things to consider: better data quality and real-time data, better hyperparameter tuning, better feature selection, etc.

- Further works:
  - change the architecture -> self-attention-based ML
  - better data, better features
  - Regional forecasting (different target parameters, e.g. foF2)

(*)Final Global TEC maps from IGS and developed by the Universitat Politècnica de Catalunya (UPC) are available at ftp://cddis.gsfc.nasa.gov. Kp data are available on NOAA Website. The SymH data was provided by the WDC for Geomagnetism, Kyoto (http://wdc.kugi.kyoto-u.ac.jp/wdc/Sec3.html).

María Graciela Molina
gmolina@herrera.unt.edu.ar