COSPAR 2018

42nd Assembly | 60th Anniversary

Systematic analysis of machine learning techniques for Kp prediction in the framework of the H2020 project 'SWAMI'

Irina Zhelavskaya[1,2], Yuri Shprits[1,2,3], Ruggero Vasile[1], Claudia Stolle[1,4], Jürgen Matzka[1,4]

[1]GFZ Potsdam, Germany, [2]University of Potsdam, Germany, [3]Department of Earth, Planetary, and Space Sciences, UCLA, [4]Institute of Earth and Environmental Science, University of Potsdam, Germany

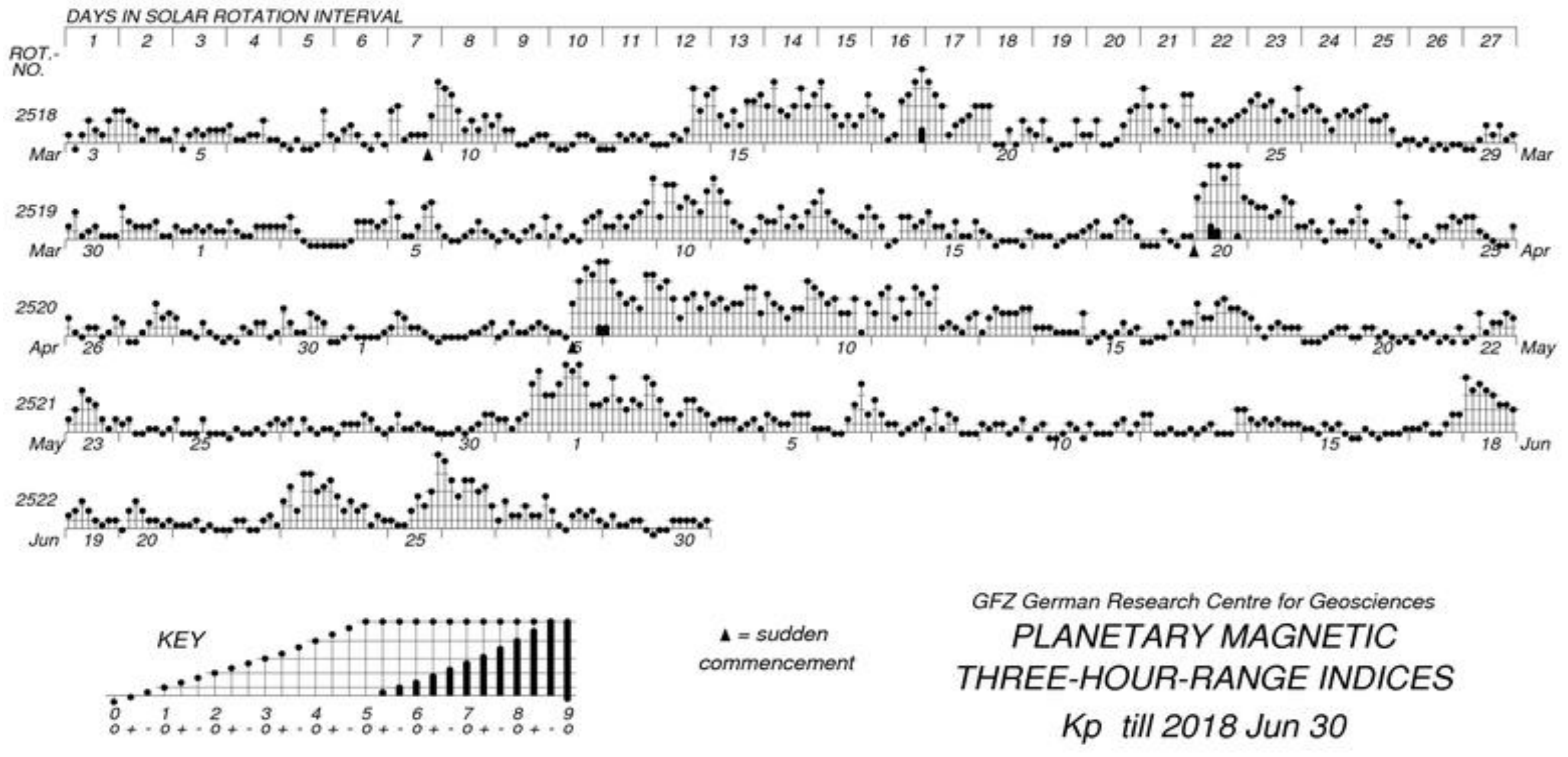
17 July 2018

OUTLINE

1. Motivation
2. Machine learning methods:
   a) Model development
   b) Feature selection
3. Comparison of the methods
4. Resulting models
5. Conclusions

KP INDEX

- Global measure of geomagnetic disturbance due to solar particle radiation.
- Used worldwide as a standard to measure geomagnetic activity.
- Average value of disturbance levels in the horizontal components of the magnetic field at 13 selected stations around the globe.
- Input parameter for important near-Earth space models, like air-drag, radiation belt, diffusion coefficients, plasmapause, thermospheric etc. models.



https://www.gfz-potsdam.de/en/kp-index/

CURRENT METHODOLOGIES

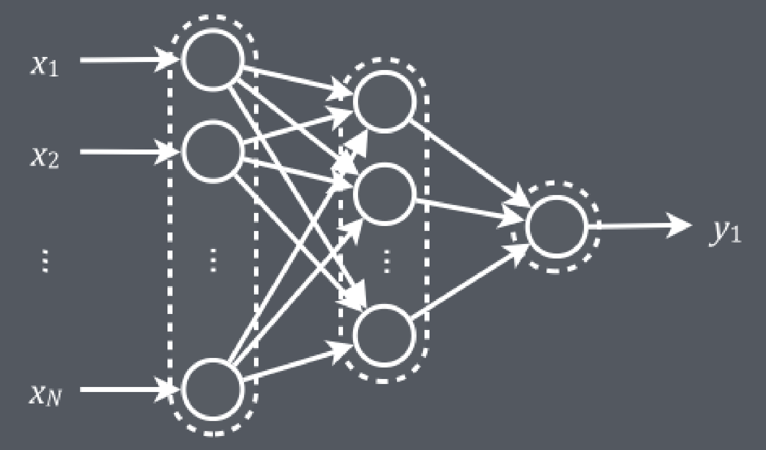
Input

Time history of
(1) solar wind speed
(2) proton density
(3) IMF B
(4) IMF Bz …

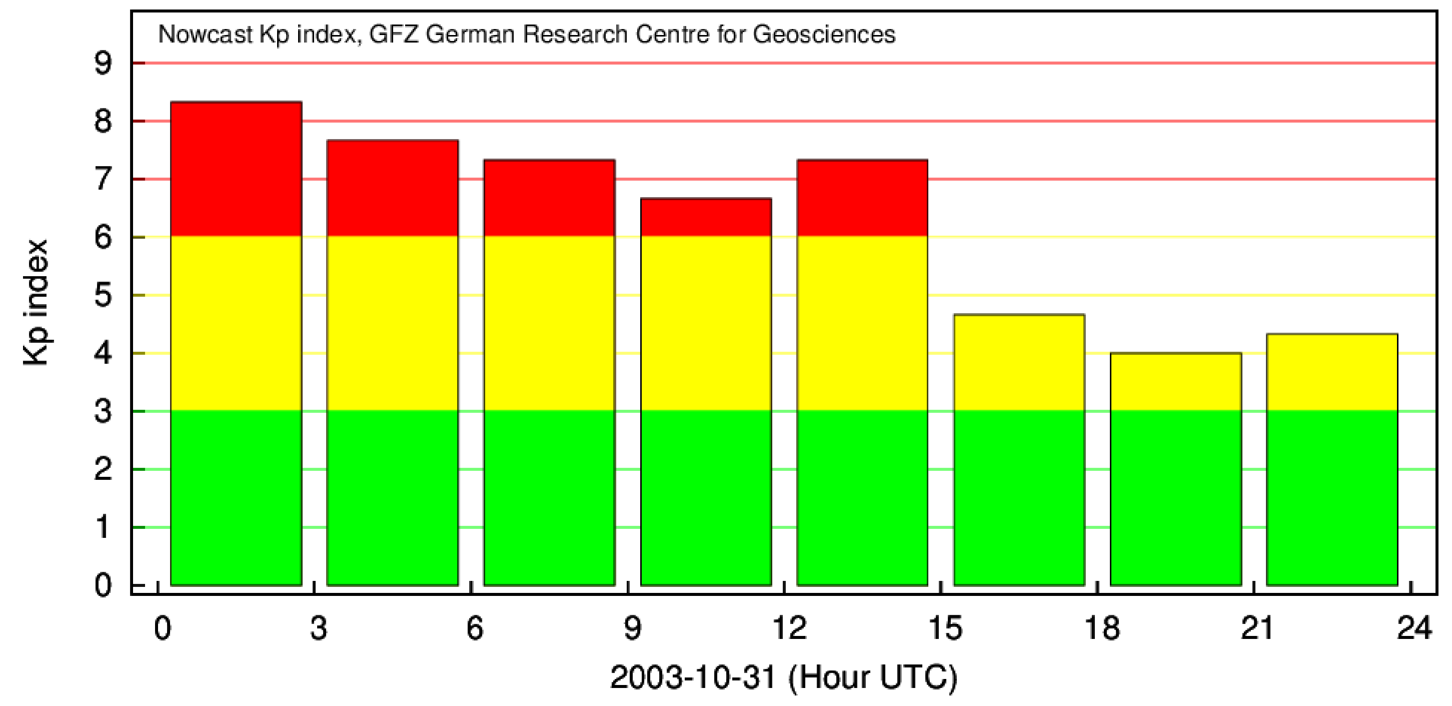
Output

Kp index

Neural networks



Kp prediction



CURRENT METHODOLOGIES

Input

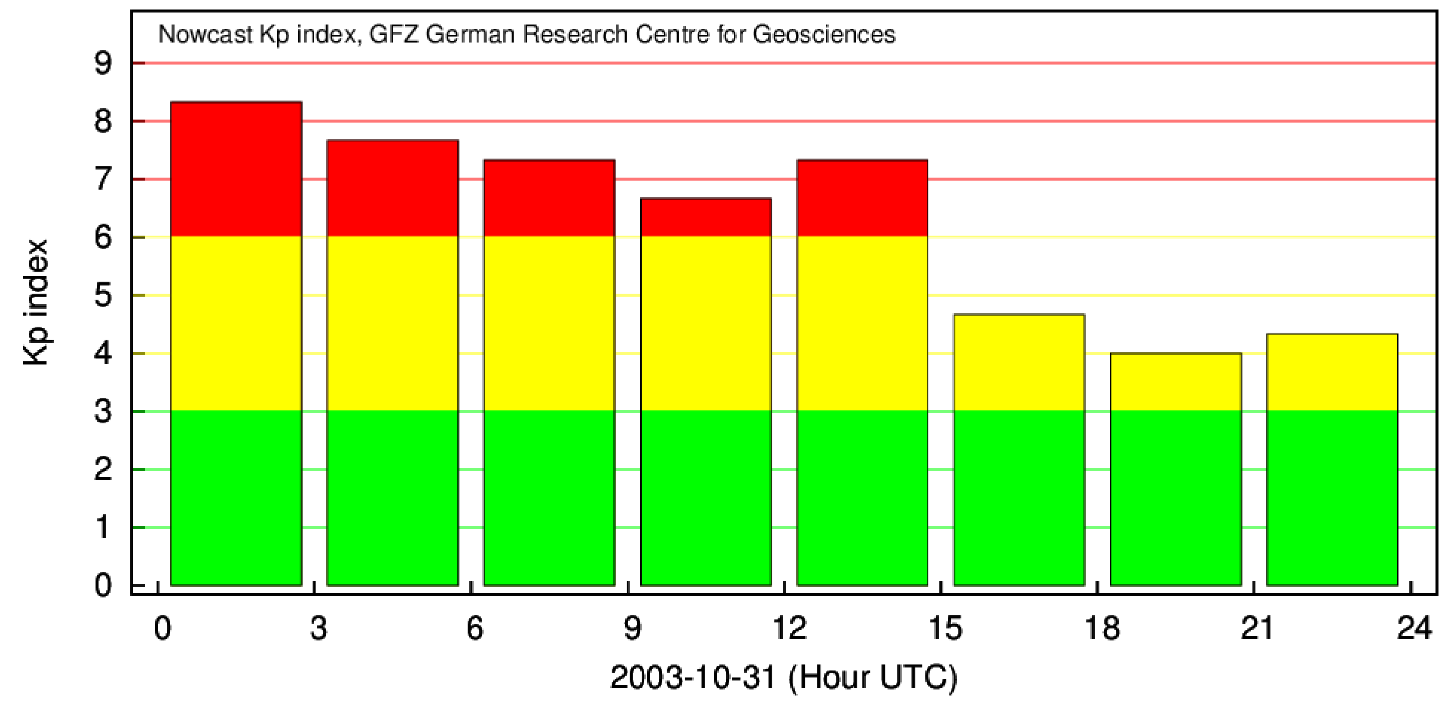
Time history of
(1) solar wind speed
(2) proton density
(3) IMF B
(4) IMF Bz …

Output

Kp index

→ ? →

Kp prediction



MACHINE LEARNING METHODS

- **Model development methods:**
  - Feedforward Neural Networks (NN)
  - Gradient Boosting (GB)
  - Linear Regression (LR)

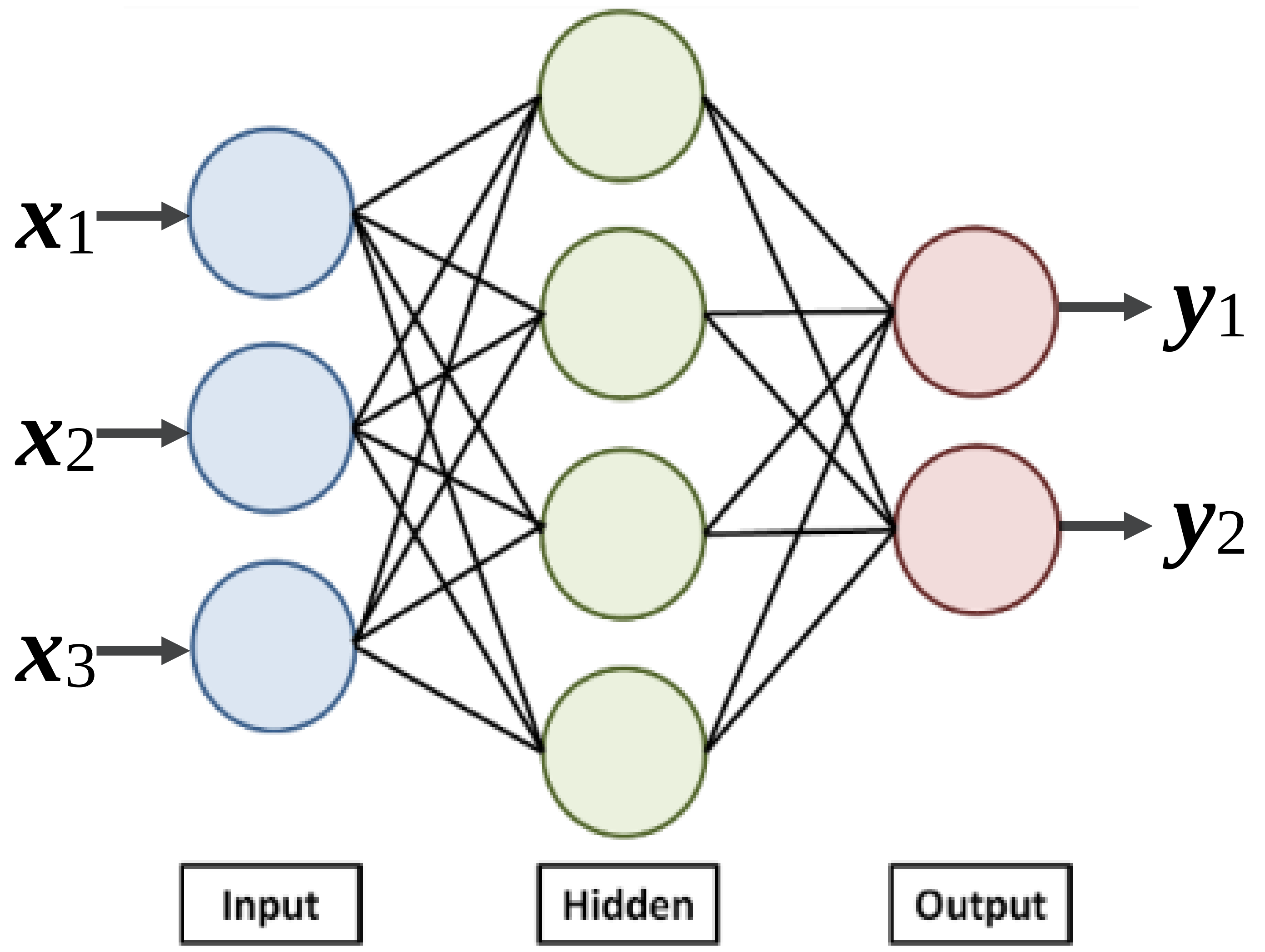
MACHINE LEARNING METHODS

- **Model development methods:**
  - Feedforward Neural Networks (NN)
  - Gradient Boosting (GB)
  - Linear Regression (LR)

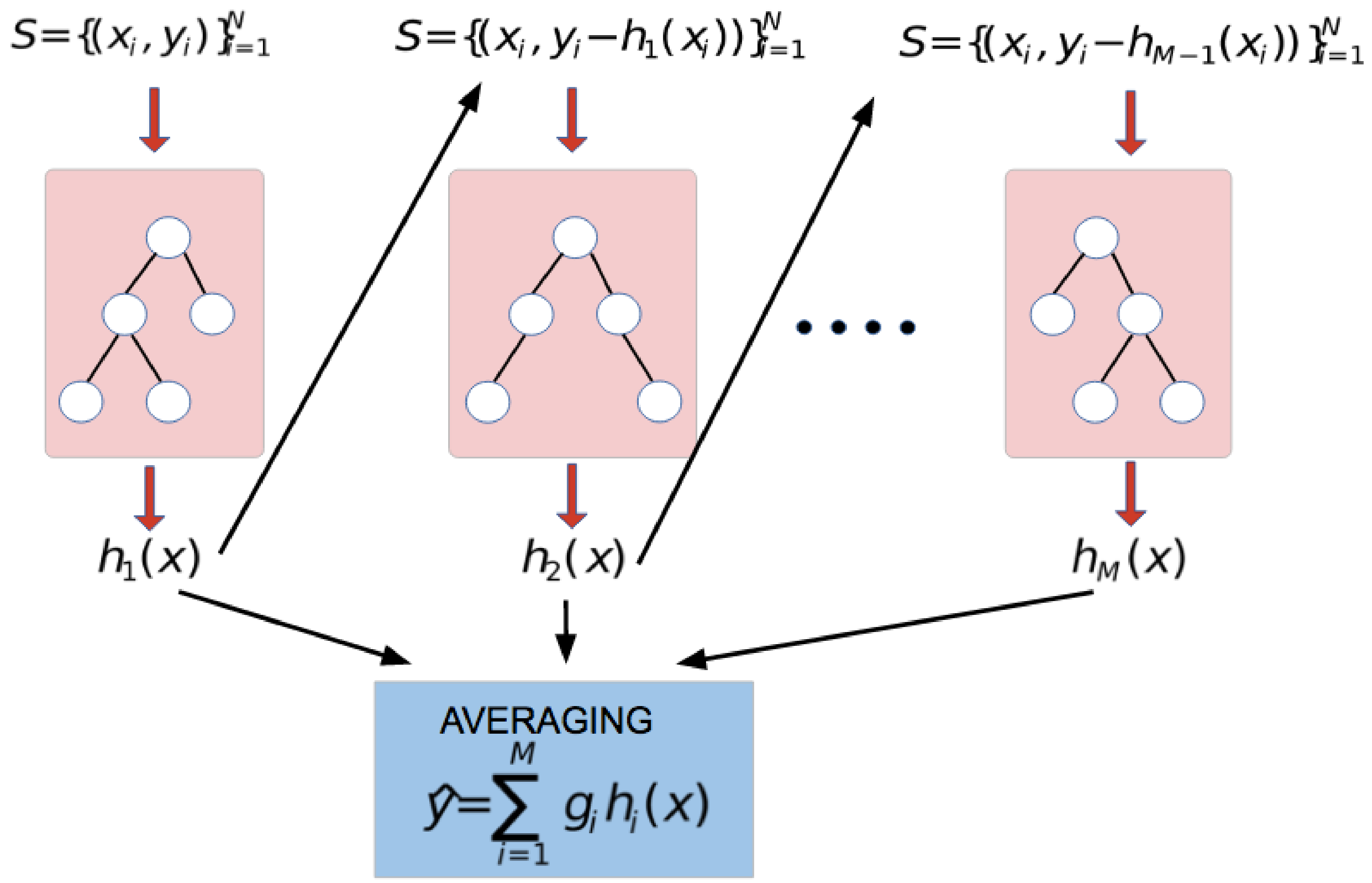
- **Input selection methods:**
  - Random Forest (RF)
  - Mutual Information (MI)

MODEL DEVELOPMENT METHODS

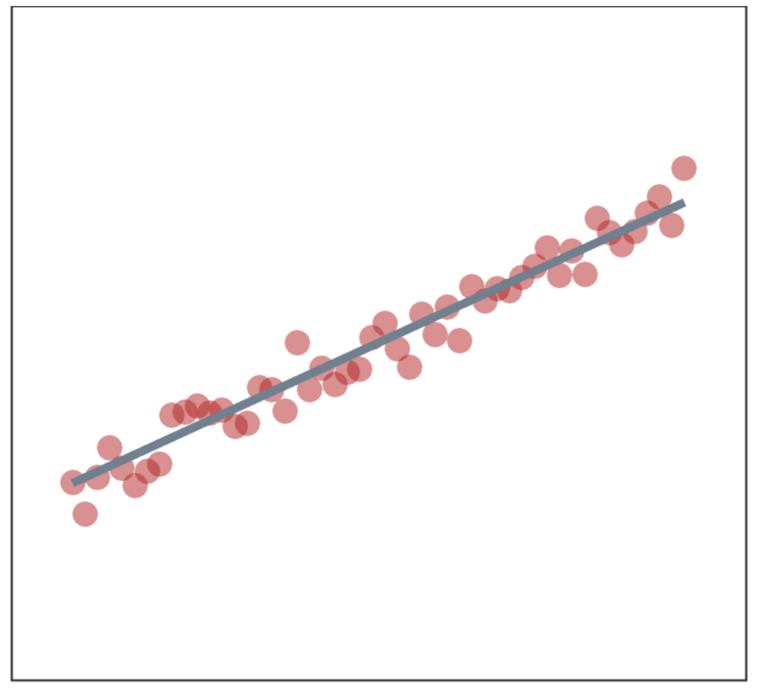
Feedforward Neural Networks (FNNs)



Gradient Boosting (GB)



Linear Regression (LR)

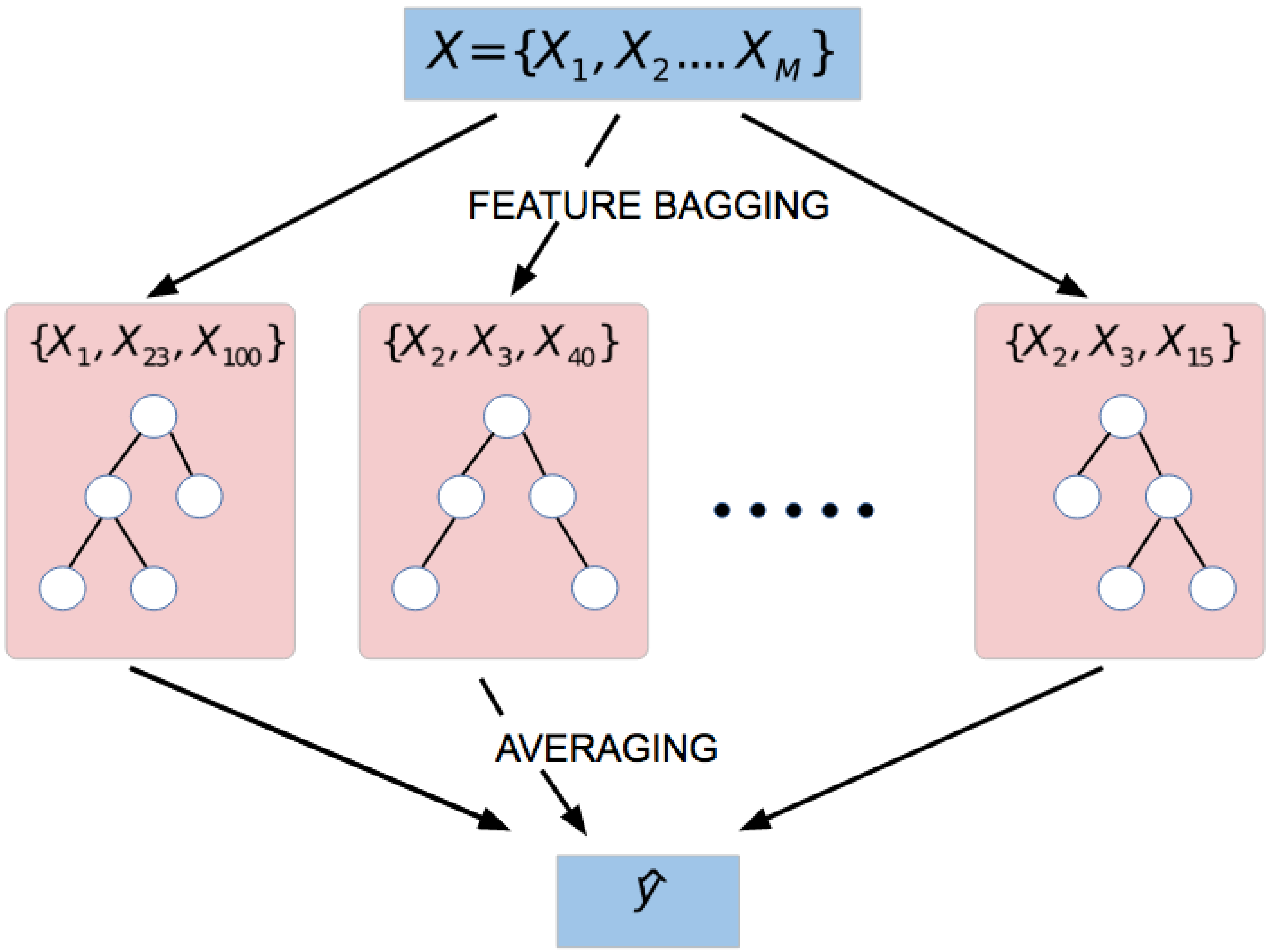


$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

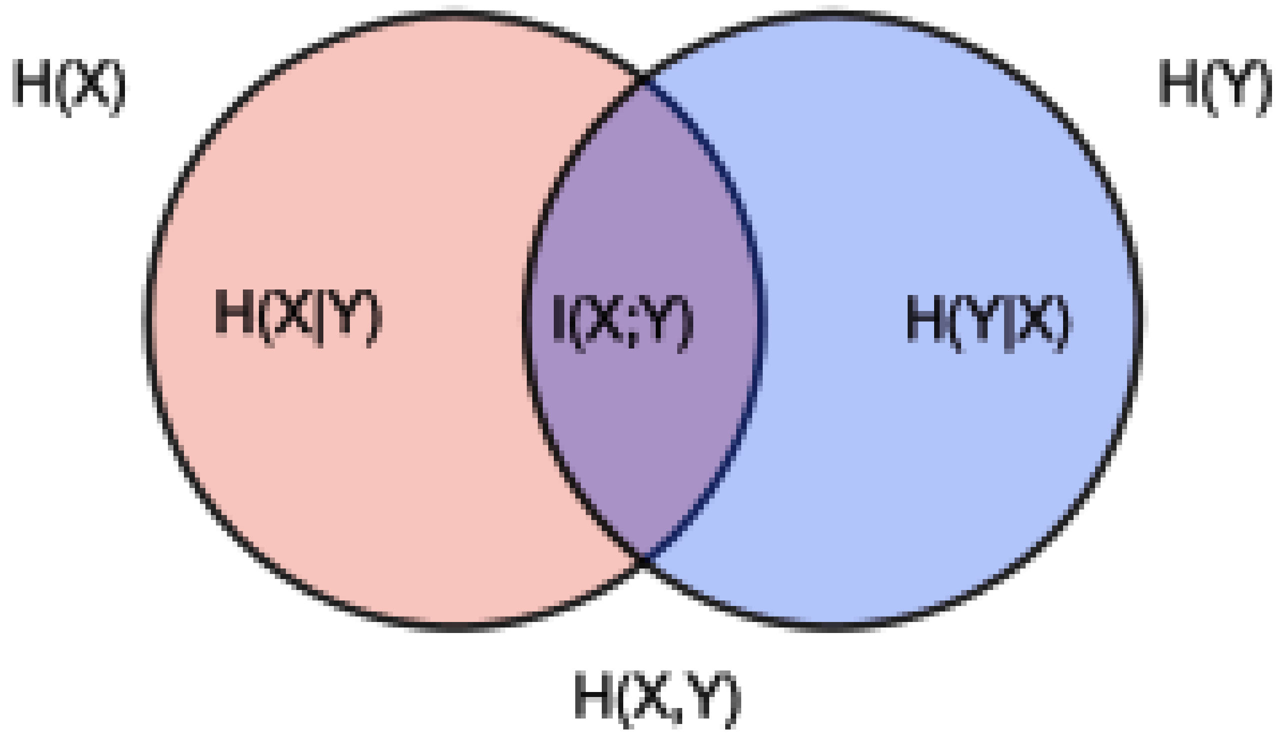
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

INPUT SELECTION METHODS

Random Forest (RF)



Mutual Information (MI)

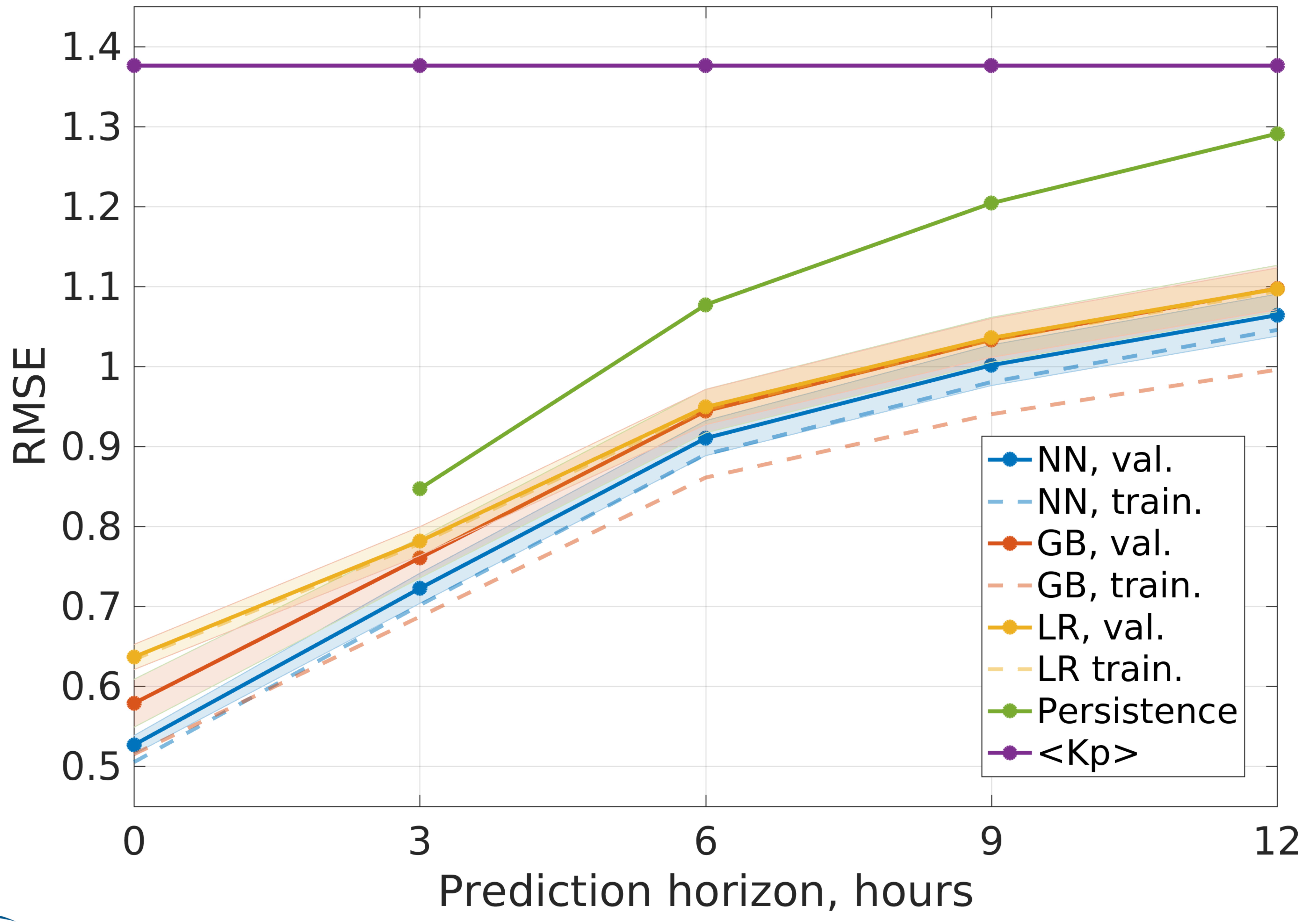


$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$$

$$I(X;Y) \geq 0, I(X;Y) = I(Y;X)$$

$$H(X) = \sum_{i=1}^{n} P(x_i) I(x_i) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

1. MODEL DEVELOPMENT METHODS PERFORMANCE



Input data

Averages, min, and max of:

- solar wind speed,
- proton density,
- IMF B,
- IMF Bz over 0-3, 3-6, 6-9 hours.

over 1998-2017 from OMNIWeb.

Training setup

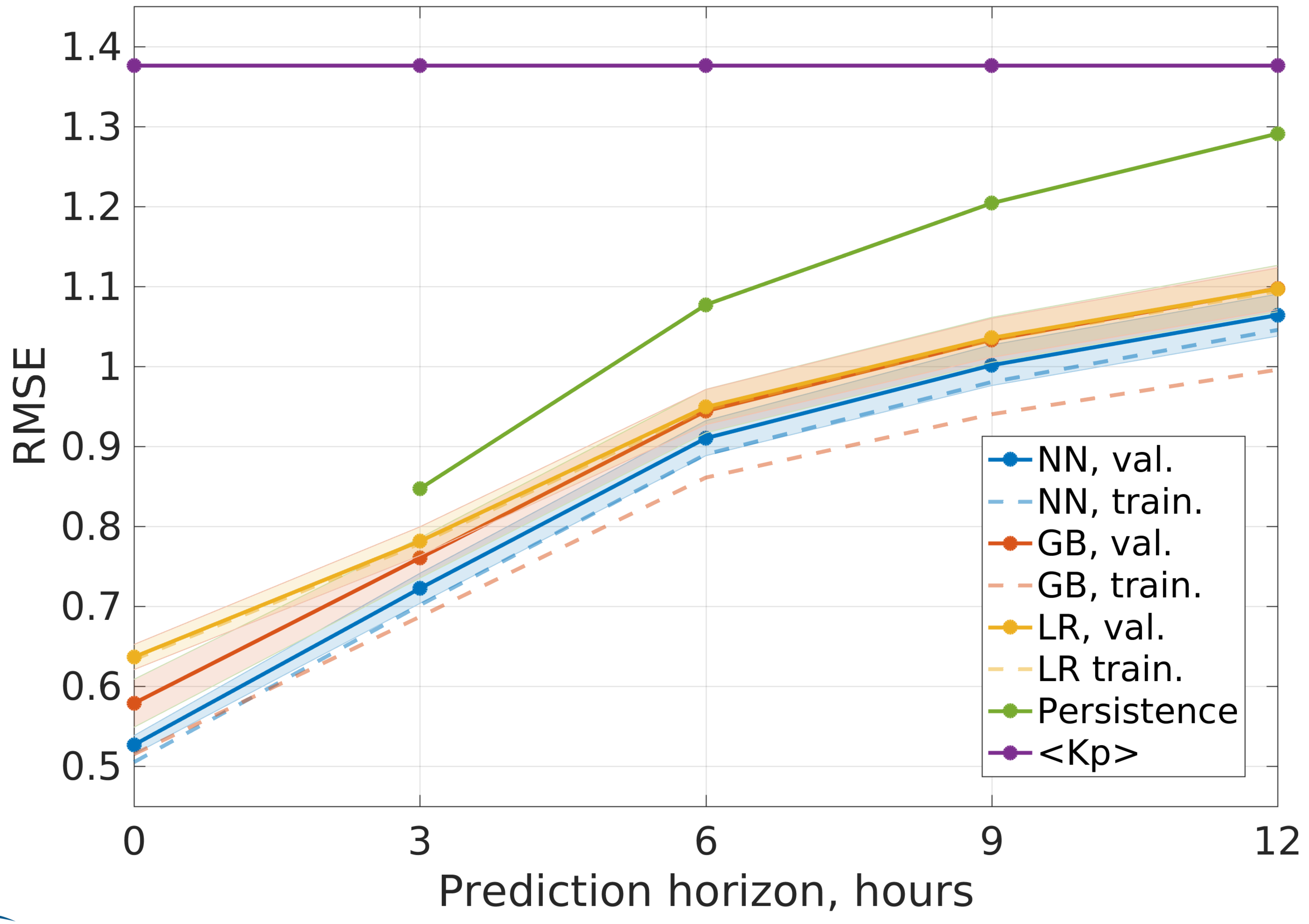
- 5-fold cross-validation (CV) with 10 repeats.
- Data are first split into 35-day chunks sequential in time.
- Separately from that, test set is left aside comprising 10%.

GFZ Helmholtz Centre Potsdam

HELMHOLTZ

1. MODEL DEVELOPMENT METHODS PERFORMANCE

Neural Networks outperform other methods



Input data

Averages, min, and max of:

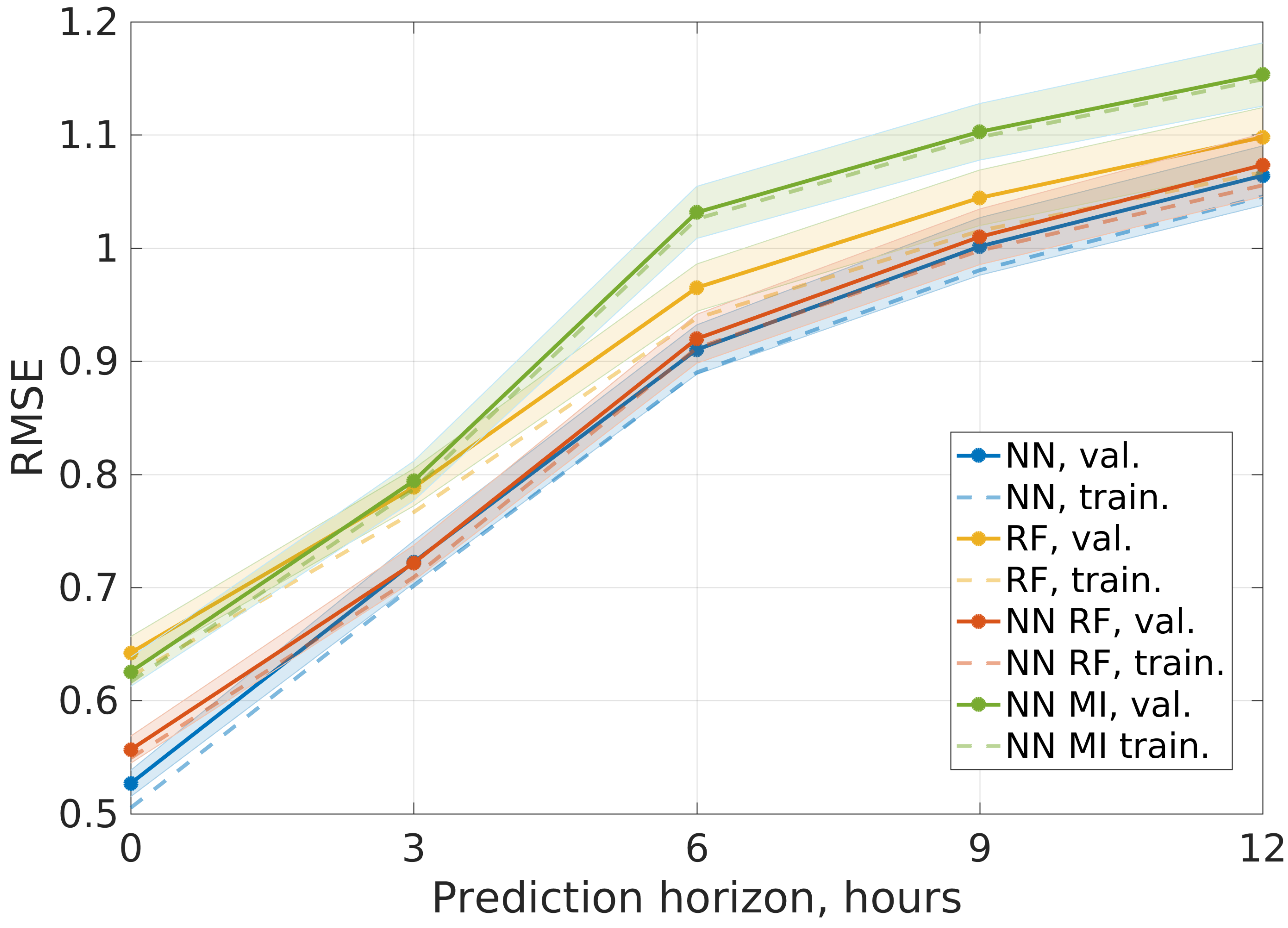
- solar wind speed,
- proton density,
- IMF B,
- IMF Bz over 0-3, 3-6, 6-9 hours.

over 1998-2017 from OMNIWeb.

Training setup

- 5-fold cross-validation (CV) with 10 repeats.
- Data are first split into 35-day chunks sequential in time.
- Separately from that, test set is left aside comprising 10%.

2. INPUT SELECTION METHODS PERFORMANCE

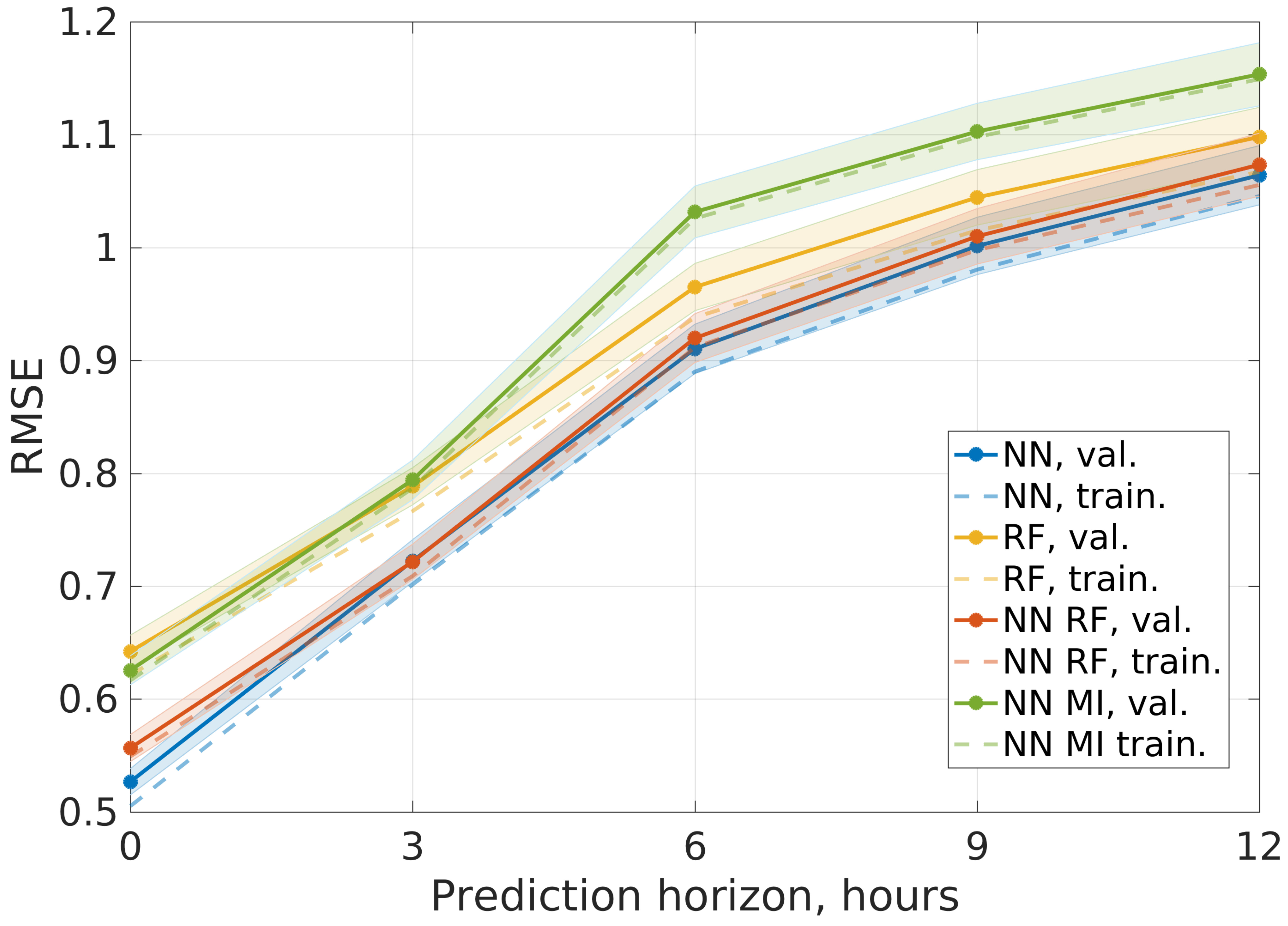


Example of selected features for horizon = 3 hours

RF: h = 3	**MI: h = 3**
$\mathbf{Bz_{min,0\text{-}3}}$	$\mathbf{Bz_{min,0\text{-}3}}$
$\mathbf{Vsw_{max,0\text{-}3}}$	$\mathbf{B_{max,0\text{-}3}}$
$Vsw_{avg,0\text{-}3}$	$B_{max,3\text{-}6}$
$\mathbf{B_{max,0\text{-}3}}$	$Bz_{min,3\text{-}6}$
$Vsw_{min,0\text{-}3}$	$B_{max,6\text{-}9}$
$Bz_{avg,0\text{-}3}$	$B_{max,9\text{-}12}$
$nProt_{max,0\text{-}3}$	$Bz_{min,6\text{-}9}$
$Vsw_{min,3\text{-}6}$	$B_{max,12\text{-}15}$
$nProt_{avg,0\text{-}3}$	$Bz_{min,9\text{-}12}$
$Vsw_{avg,3\text{-}6}$	$B_{max,15\text{-}18}$
$Vsw_{min,6\text{-}9}$	$\mathbf{Vsw_{max,0\text{-}3}}$
$B_{avg,0\text{-}3}$	$Bz_{min,12\text{-}15}$
$Vsw_{avg,6\text{-}9}$	$B_{max,18\text{-}21}$

2. INPUT SELECTION METHODS PERFORMANCE

Random Forest helps find optimal inputs and reduce input space



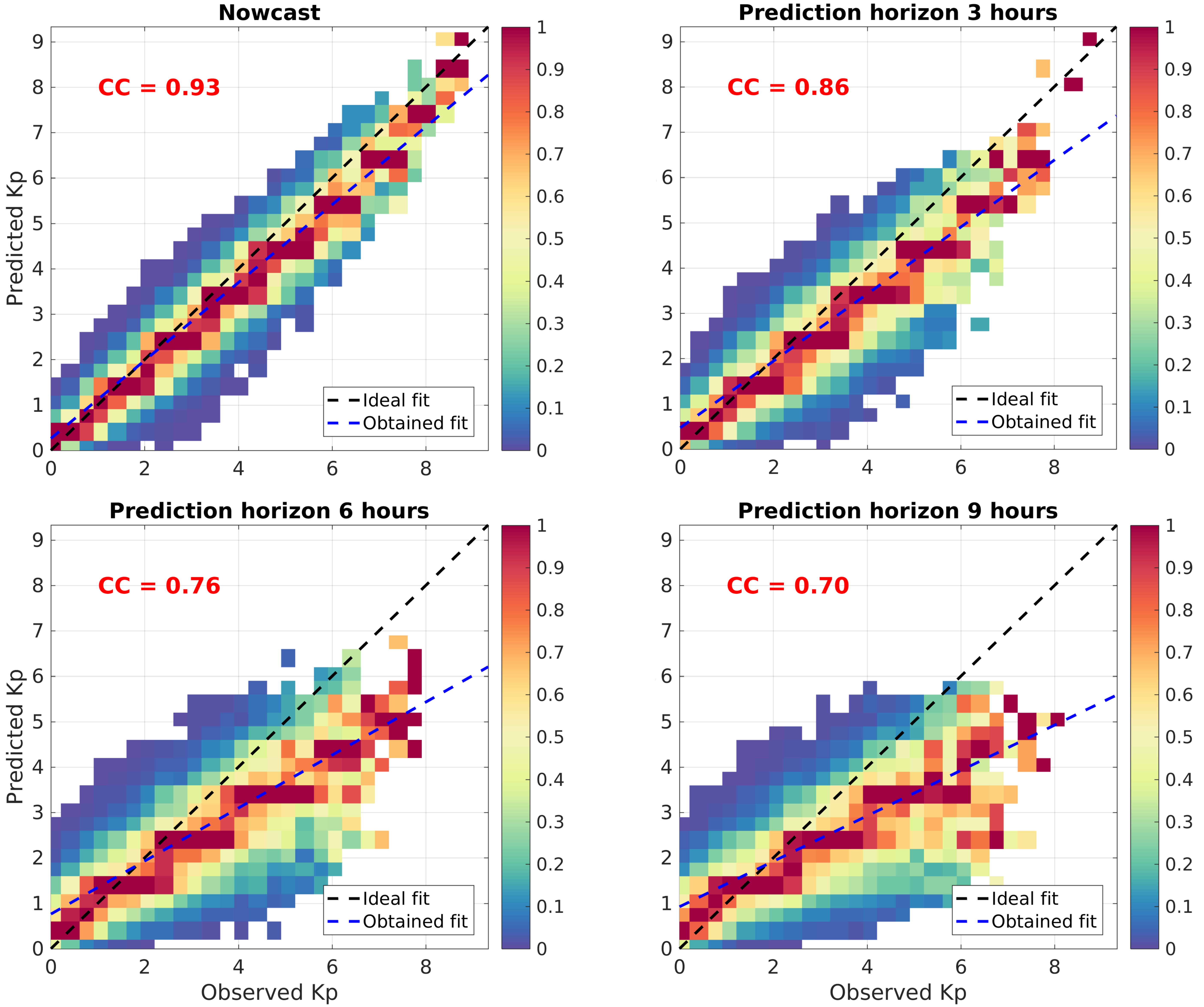
Example of selected features for horizon = 3 hours

RF: h = 3	MI: h = 3
$Bz_{min,0\text{-}3}$	**$Bz_{min,0\text{-}3}$**
$Vsw_{max,0\text{-}3}$	**$B_{max,0\text{-}3}$**
$Vsw_{avg,0\text{-}3}$	$B_{max,3\text{-}6}$
$B_{max,0\text{-}3}$	$Bz_{min,3\text{-}6}$
$Vsw_{min,0\text{-}3}$	$B_{max,6\text{-}9}$
$Bz_{avg,0\text{-}3}$	$B_{max,9\text{-}12}$
$nProt_{max,0\text{-}3}$	$Bz_{min,6\text{-}9}$
$Vsw_{min,3\text{-}6}$	$B_{max,12\text{-}15}$
$nProt_{avg,0\text{-}3}$	$Bz_{min,9\text{-}12}$
$Vsw_{avg,3\text{-}6}$	$B_{max,15\text{-}18}$
$Vsw_{min,6\text{-}9}$	**$Vsw_{max,0\text{-}3}$**
$B_{avg,0\text{-}3}$	$Bz_{min,12\text{-}15}$
$Vsw_{avg,6\text{-}9}$	$B_{max,18\text{-}21}$

GFZ Helmholtz Centre POTSDAM

HELMHOLTZ

RESULTING MODELS



CONCLUSIONS

- We have explored how three different algorithms (Neural Networks, Gradient Boosting, Linear Regression) perform on the task of predicting the Kp index for 5 different prediction horizons (up to 12 hours), and assessed the performance of the two feature selection methods based on Mutual Information and Random Forests.

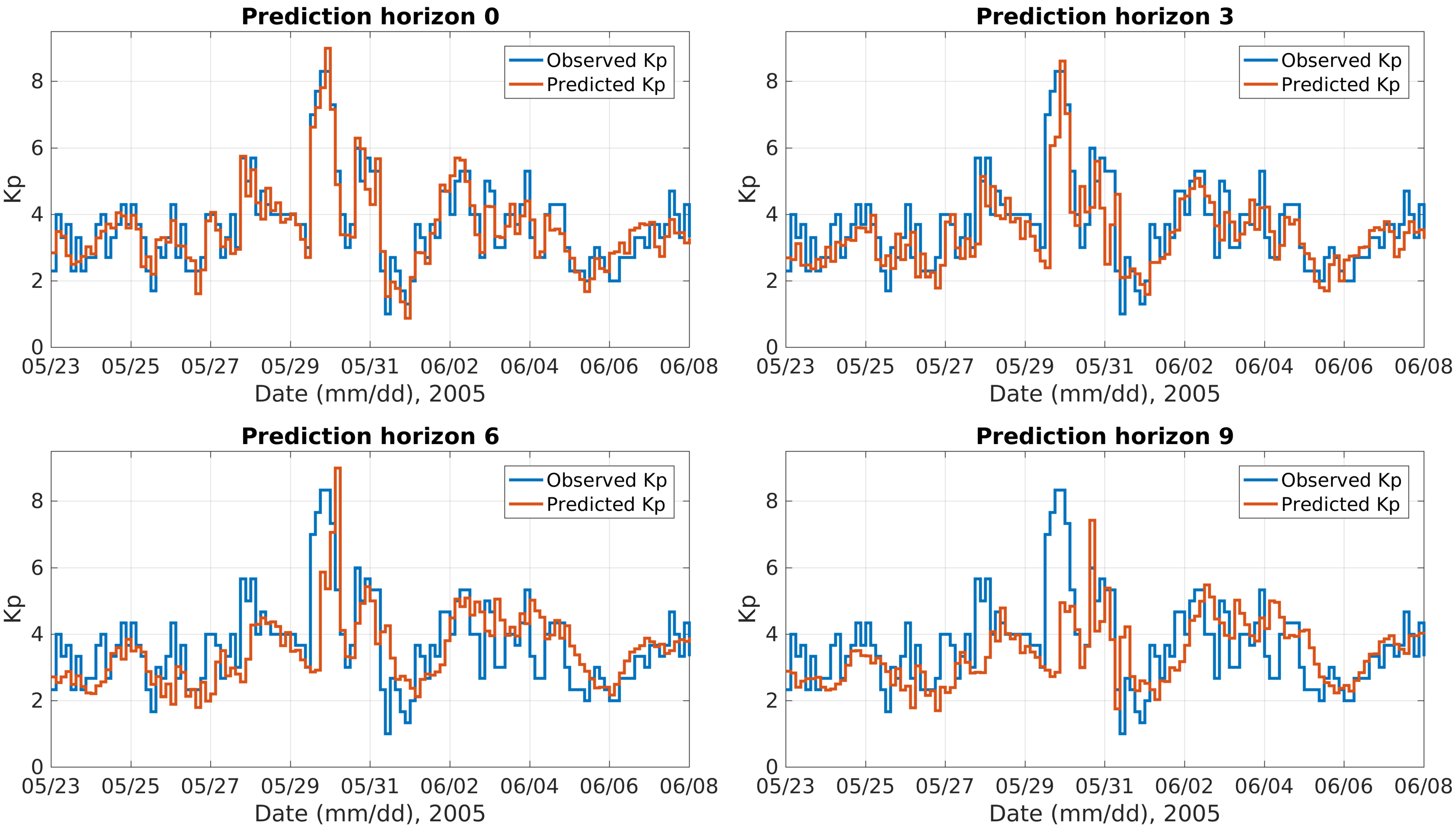
- Neural networks outperformed other models. Models based on the inputs selected by Random Forest perform similarly to the models based on the inputs selected using the domain knowledge, while the input space is significantly reduced using the RF input selection (models can be trained faster).

- More information about validation of Kp predictive models in **C0.1-005-18 on Thursday, 19 July, 15:15-15:30**.

THANK YOU!

BACKUP SLIDES

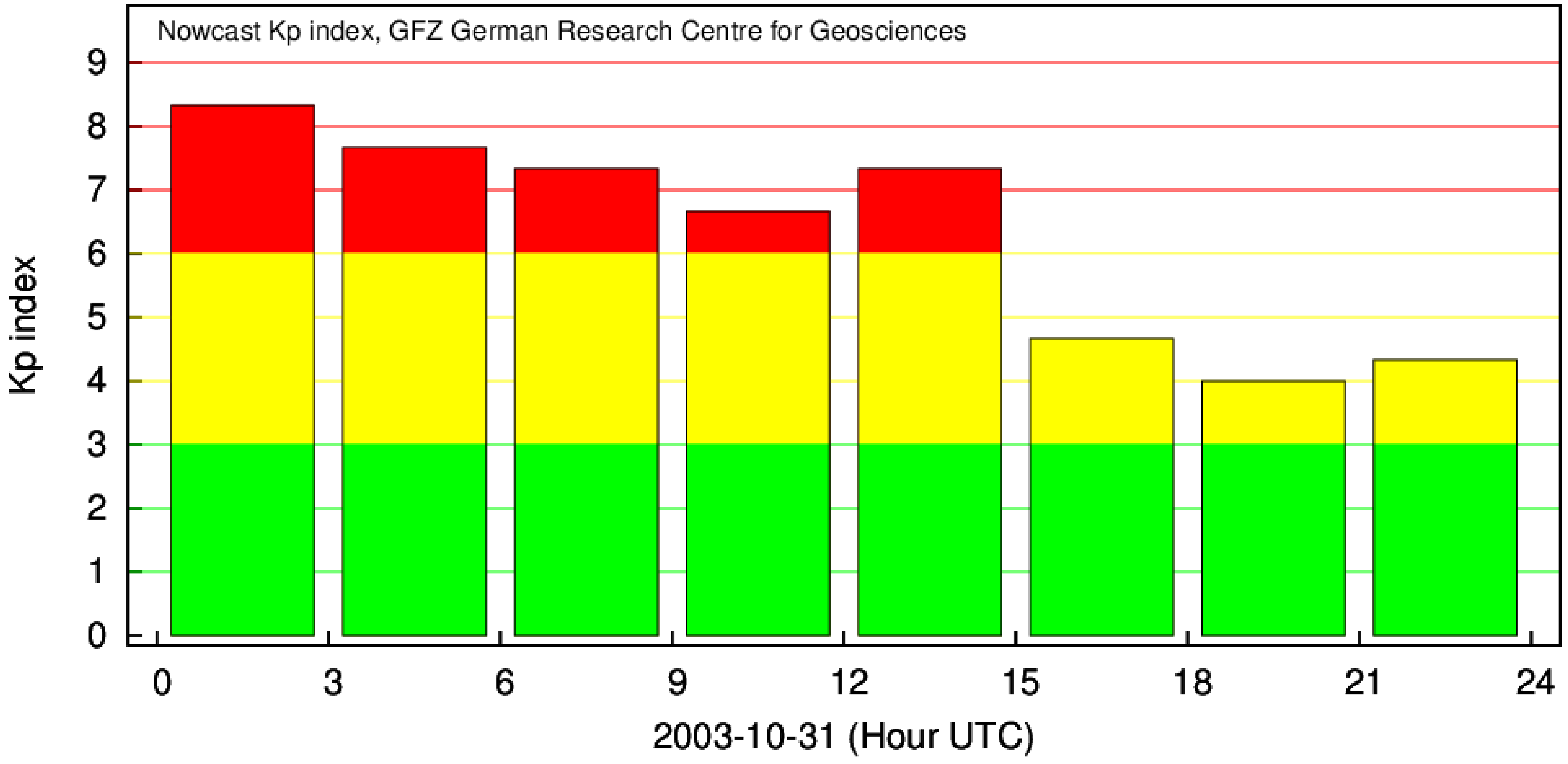
EXAMPLES OF PREDICTION



GFZ
Helmholtz Centre
POTSDAM

HELMHOLTZ

KP INDEX



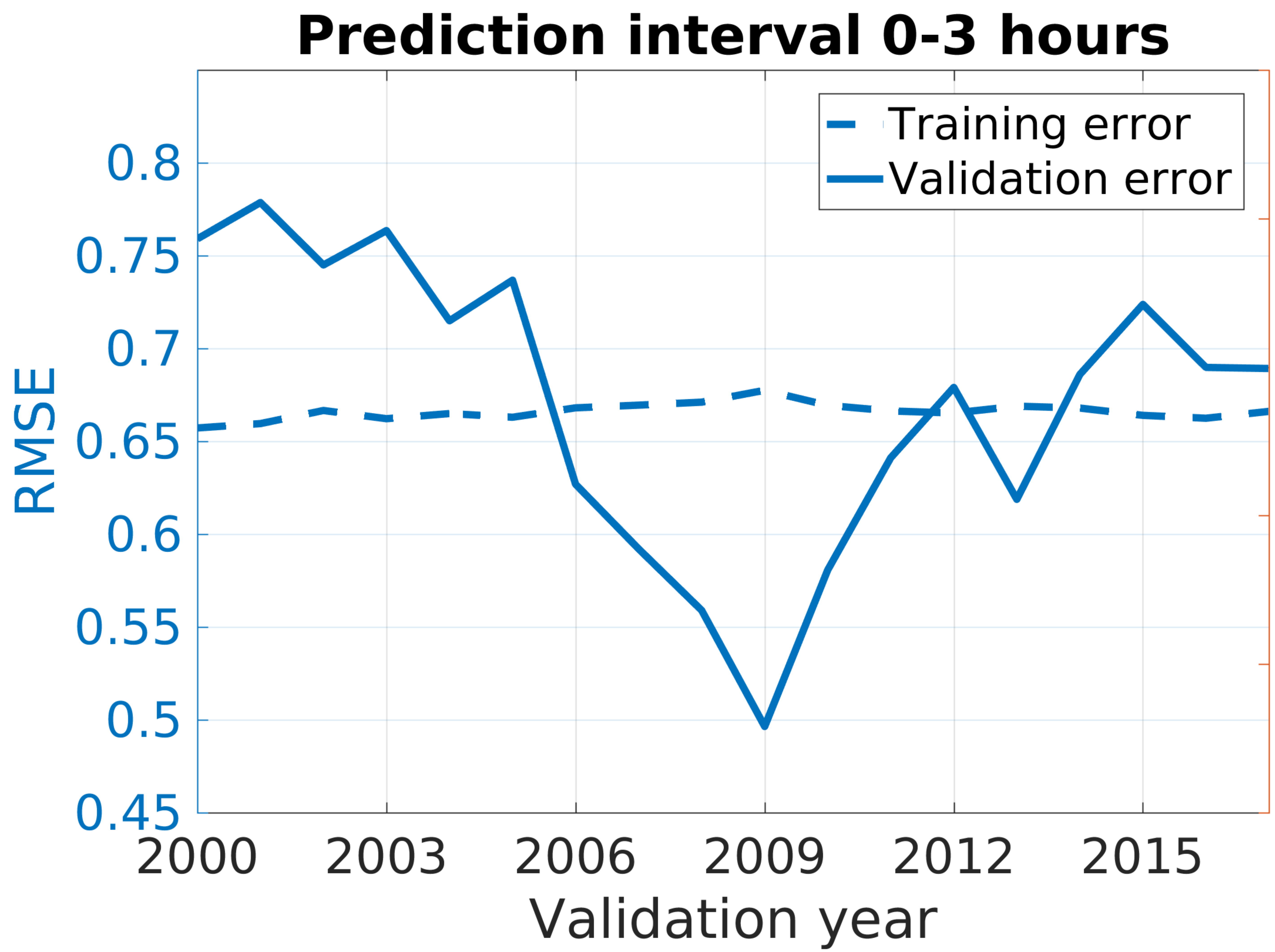
0o 0+ 1- 1o 1+ 2- 2o 2+ 3- 3o 3+ 4- 4o 4+

5- 5o 5+ 6- 6o 6+ 7- 7o 7+ 8- 8o 8+ 9- 9o

- 3-hour interval index
- Values from 0 (quiet), 0+, 1-,... to 9-, 9 (very active/disturbed)
- Input parameter for important near-Earth space models, like air-drag, radiation belt, diffusion coefficients, plasmapause, thermospheric etc. models

https://www.gfz-potsdam.de/en/kp-index/

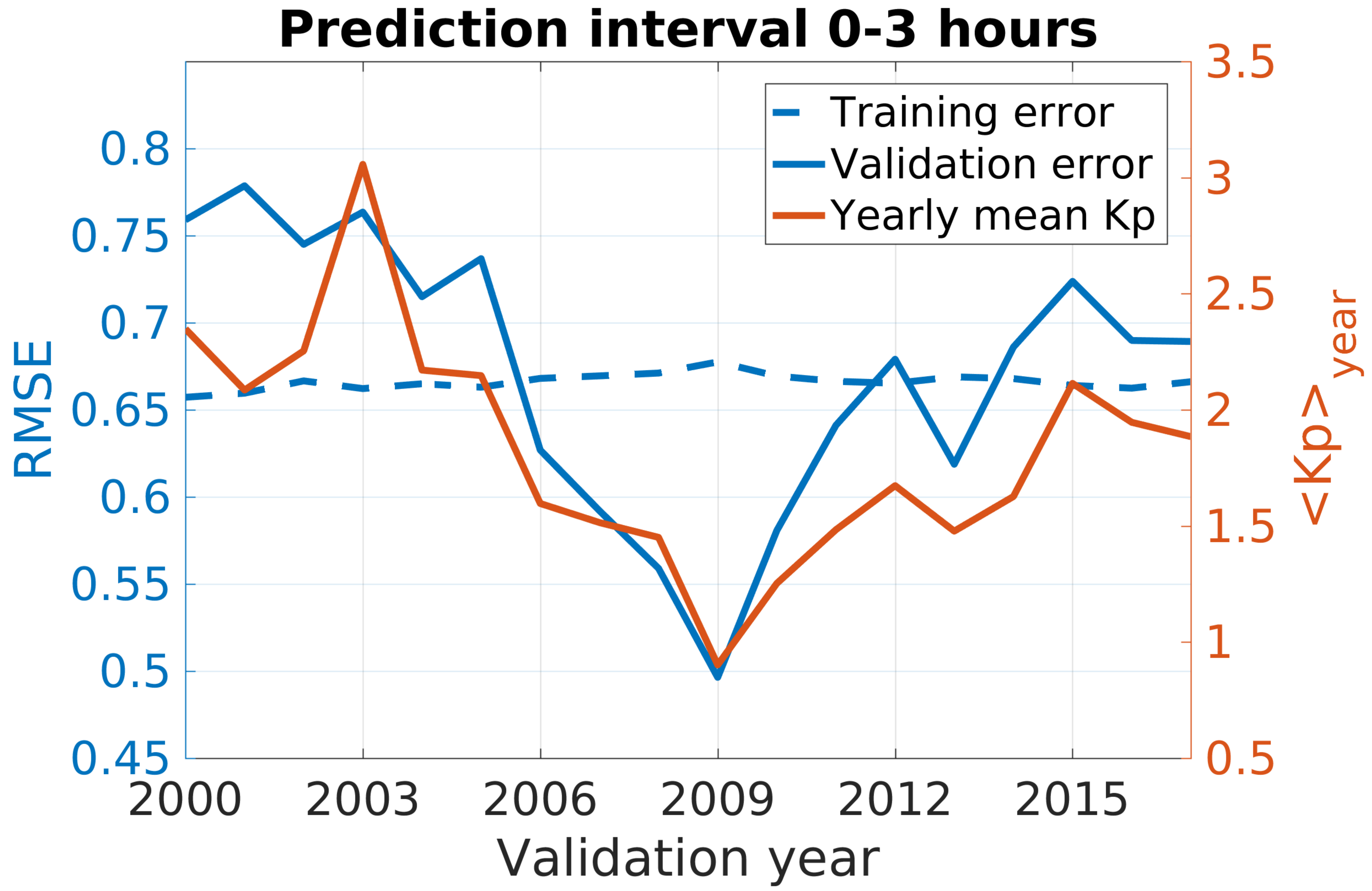
DEPENDENCE OF MODEL ERROR ON SOLAR CYCLE&KP



GFZ
Helmholtz Centre
POTSDAM

HELMHOLTZ

DEPENDENCE OF MODEL ERROR ON SOLAR CYCLE&KP



GFZ Helmholtz Centre POTSDAM

HELMHOLTZ

EXISTING KP PREDICTION MODELS

Model	Prediction horizon	Accuracy	Train/validation/test data sets splitting
Costello et al., 1997	1 hour	CC = 0.757	Training: 1970,1976,1978,1980,1981,1982,1989 Test: 1969,1979,1986,1990
Boberg et al., 2000	3 hours	CC = 0.768; RMSE = 0.985	Training and validation sets of equal lengths selected from 1976-1985 Test: 1986-1996
Wing et al., 2005	1 and 4 hours	1h: CC= 0.92 4h: CC = 0.79	Training and test sets of equal lengths, randomly selected from1975-2001
Bala and Reiff, 2012	1, 3, and 6 hours	1h: CC = 0.88; RMSE = 0.615 3h: CC = 0.86; RMSE = 0.659 6h: CC = 0.76; RMSE = 0.851	Training & test: 1998/01-2001/03, 2001/05-2005/12, 2008/01-2009/12 Validation: 2001/04, 2006/01-2007/12
Wintoft et al., 2017	ACE lead time (20-90 minutes)	CC = 0.92; RMSE= 0.55	Training: 1998-2015 except for: Validation: 2000, 2006, 2012 Test: 2001, 2011
Tan et al., 2018	3 hour	CC = 0.8147; RMSE = 0.6382; MAE = 0.4765	Training: 2000/08-2011/10 Validation: 2012/11-2013/11 Test: 2013/12-2014/9

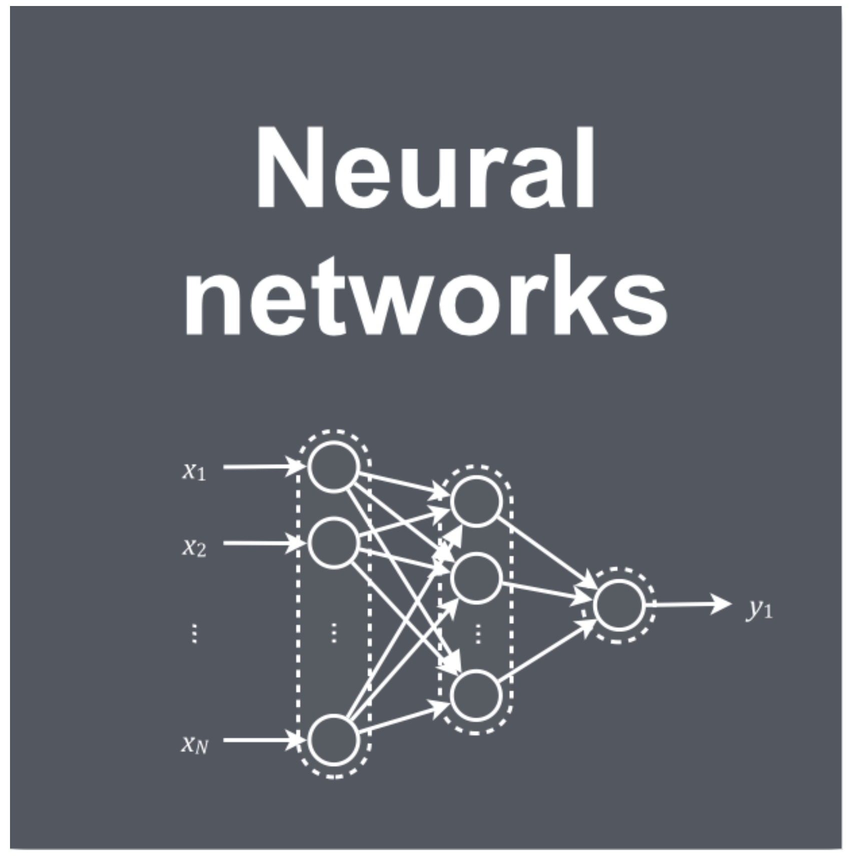
METHODOLOGY

Inputs

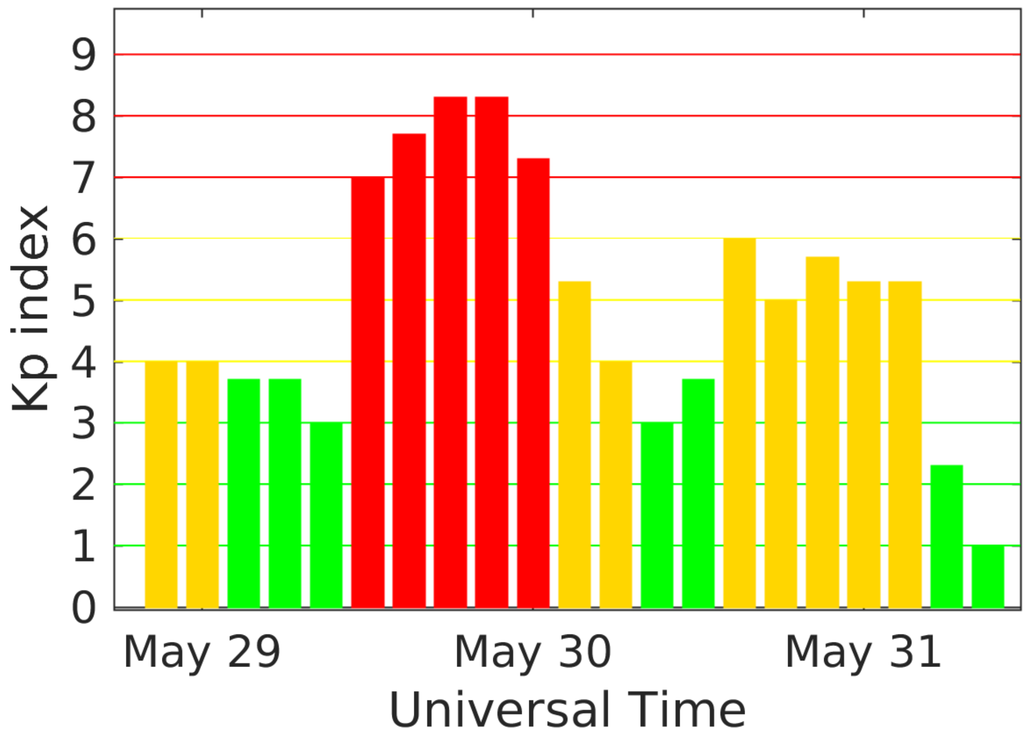
Time history of
(1) solar wind speed,
(2) proton density,
(3) IMF B,
(4) IMF Bz.

Output

Kp index



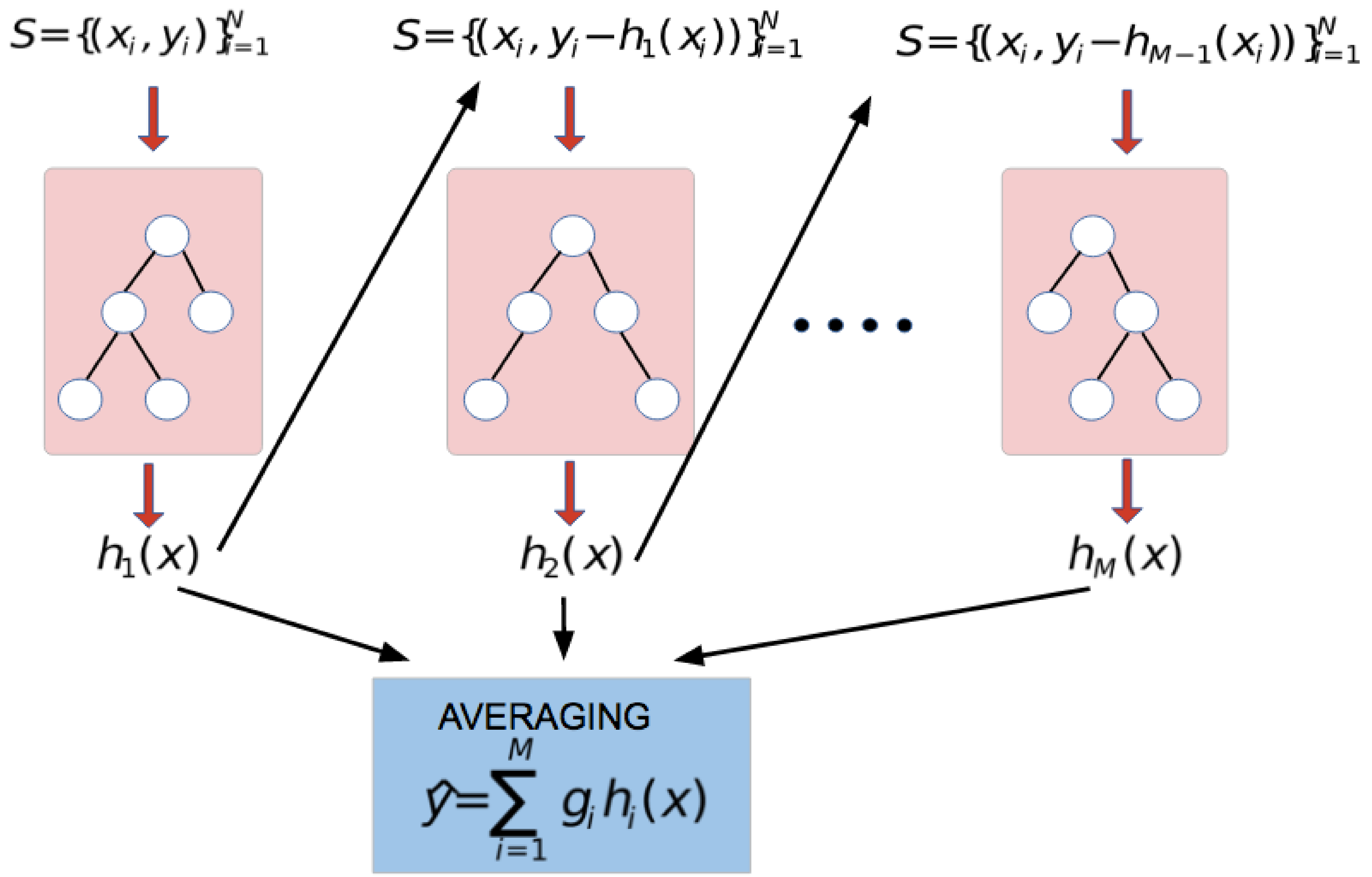
Kp prediction



- Methodology: neural network based empirical modelling.
- Data: solar wind and IMF data from ACE (available at OMNIWeb), Kp index from GFZ Potsdam, 1993 - 2017.

ML METHODS DESCRIPTION

- **Model development:**
  - Feedforward Neural Networks (NN)
  - **Gradient Boosting (GB)**
  - Linear Regression
- Feature selection:
  - Random Forest
  - Mutual Information

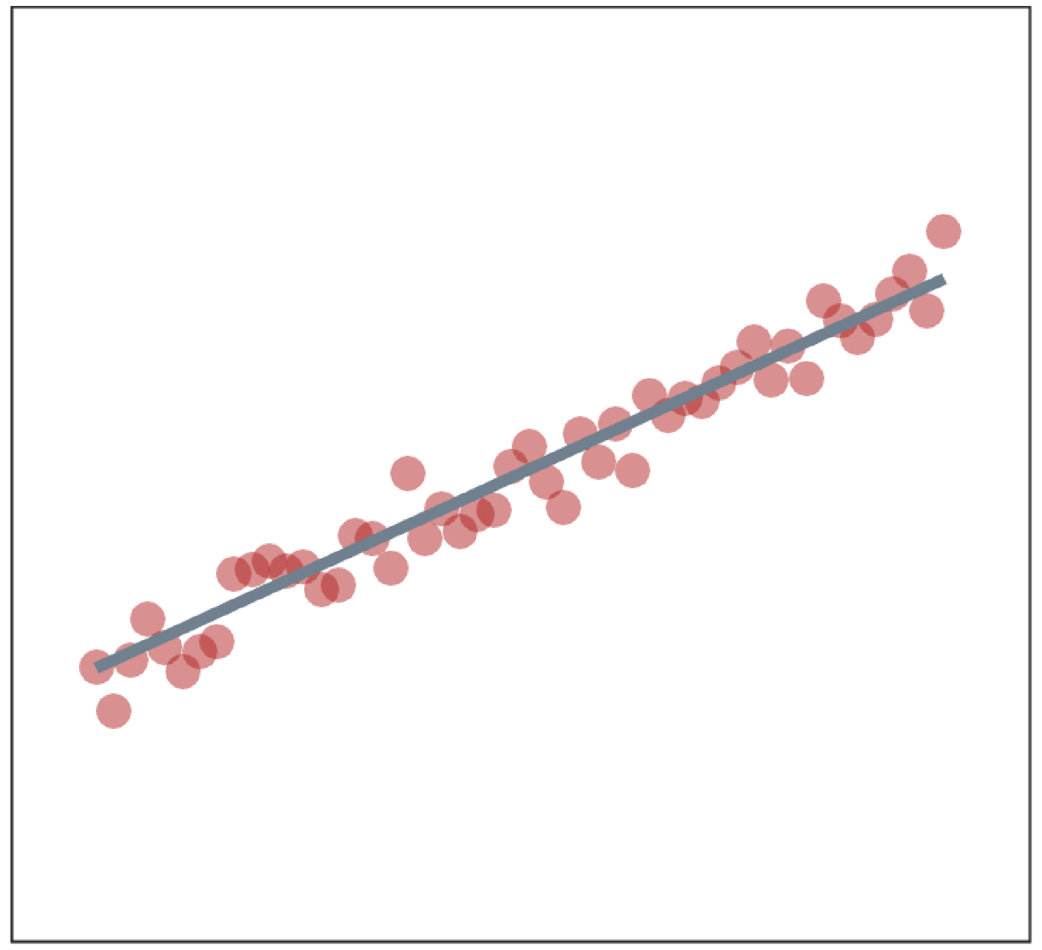


GFZ Helmholtz Centre Potsdam

HELMHOLTZ

ML METHODS DESCRIPTION

- **Model development:**
  - Feedforward Neural Networks (NN)
  - Gradient Boosting (GB)
  - **Linear Regression**
- Feature selection:
  - Random Forest
  - Mutual Information

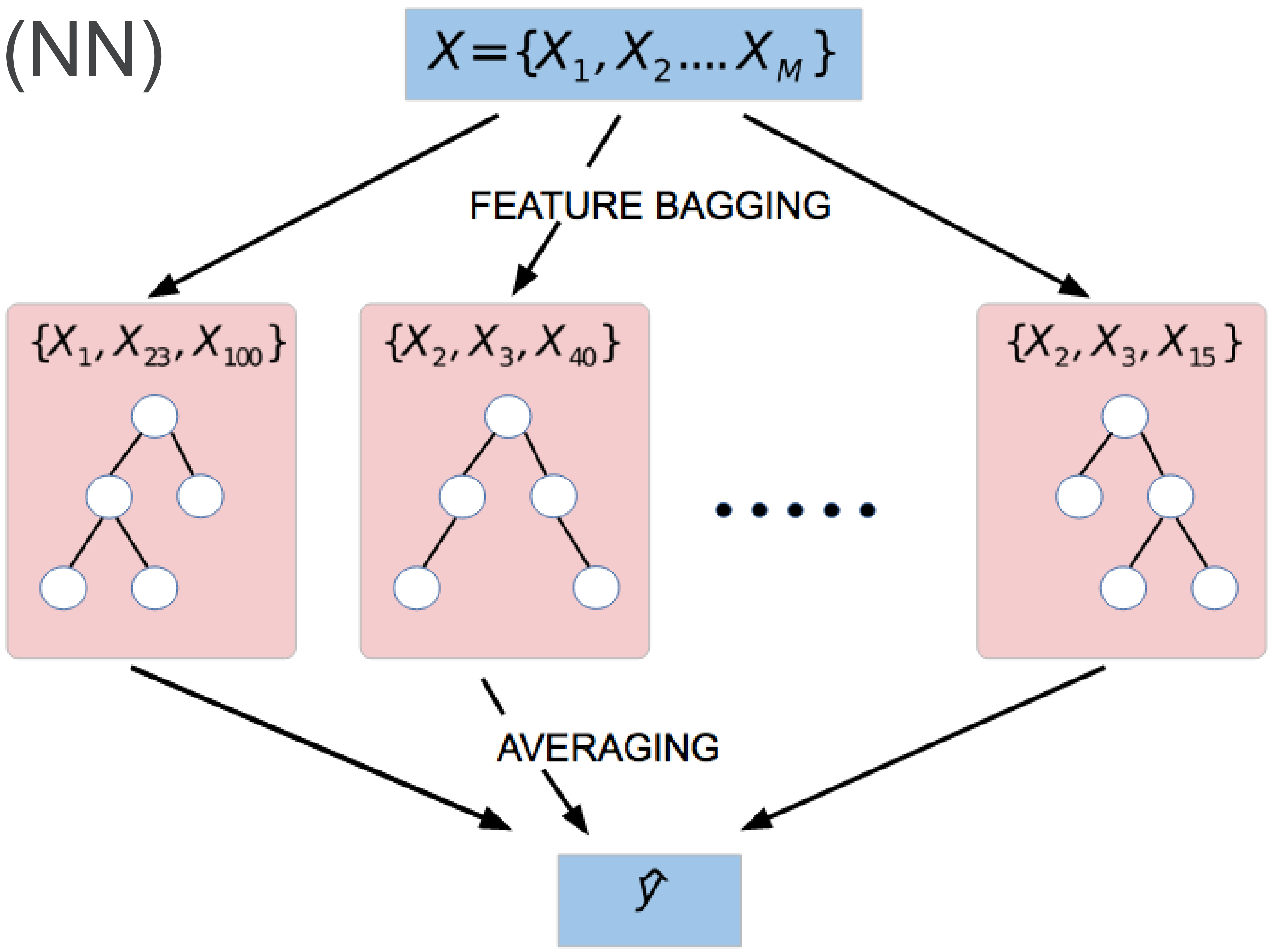


$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

ML METHODS DESCRIPTION

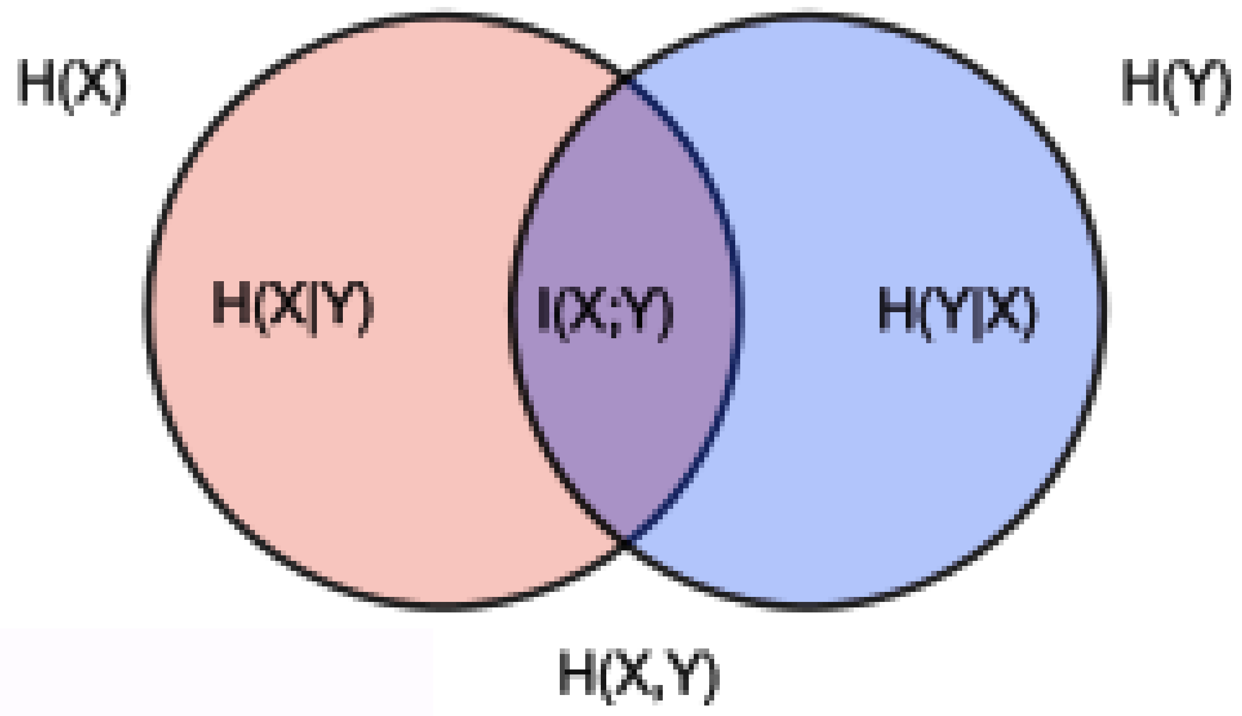
- Model development:
  - Feedforward Neural Networks (NN)
  - Gradient Boosting (GB)
  - Linear Regression
- **Feature selection:**
  - **Random Forest**
  - Mutual Information



ML METHODS DESCRIPTION

- Model development:
  - Feedforward Neural Networks (NN)
  - Gradient Boosting (GB)
  - Linear Regression
- **Feature selection:**
  - Random Forest
  - **Mutual Information**

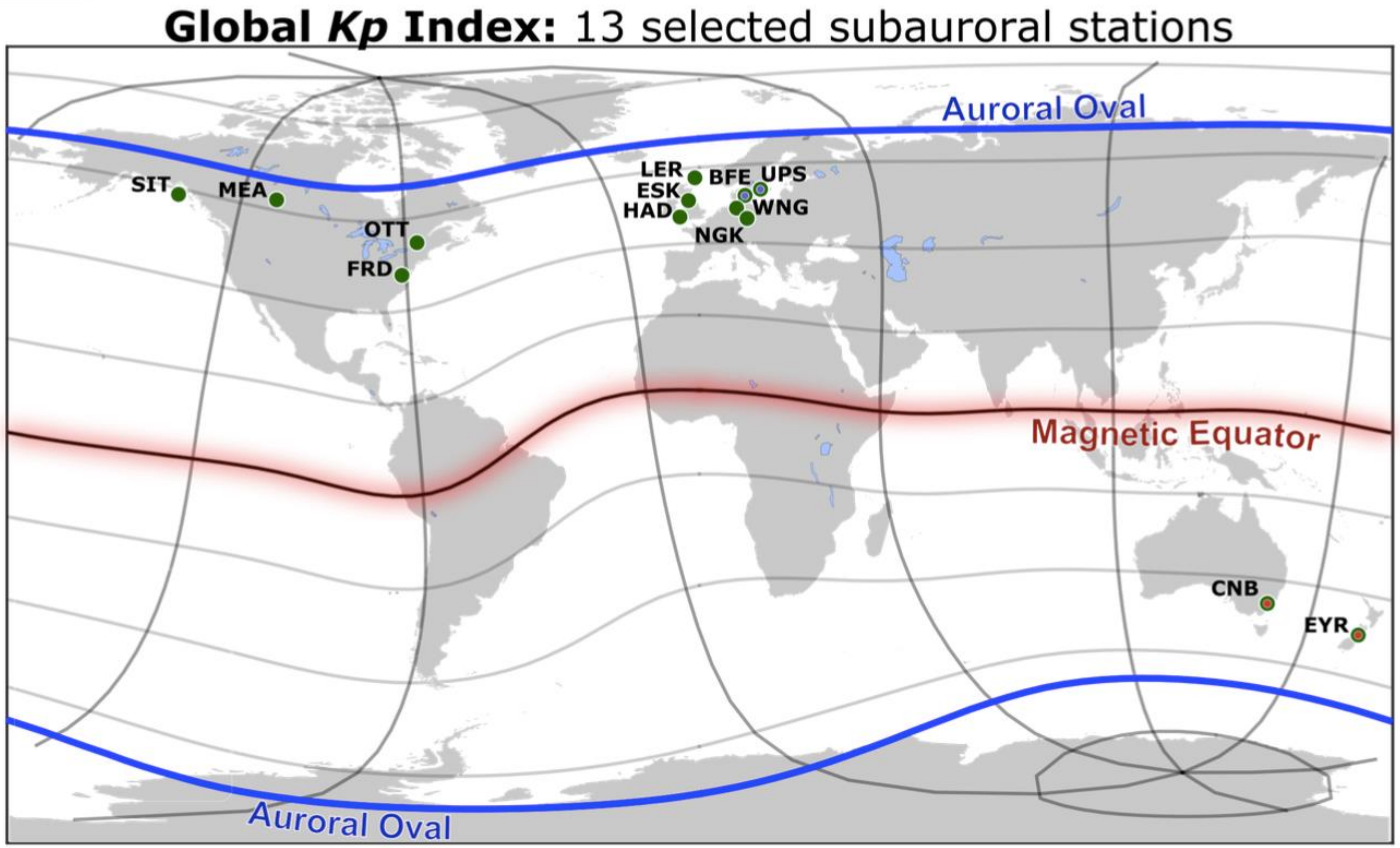
$$I(X_1;X_2) = H(X_1) + H(X_2) - H(X_1,X_2)$$



$$I(X;Y) \geq 0, I(X;Y) = I(Y;X)$$

$$H(X) = \sum_{i=1}^{n} P(x_i)I(x_i) = -\sum_{i=1}^{n} P(x_i)\log_b P(x_i)$$

KP INDEX



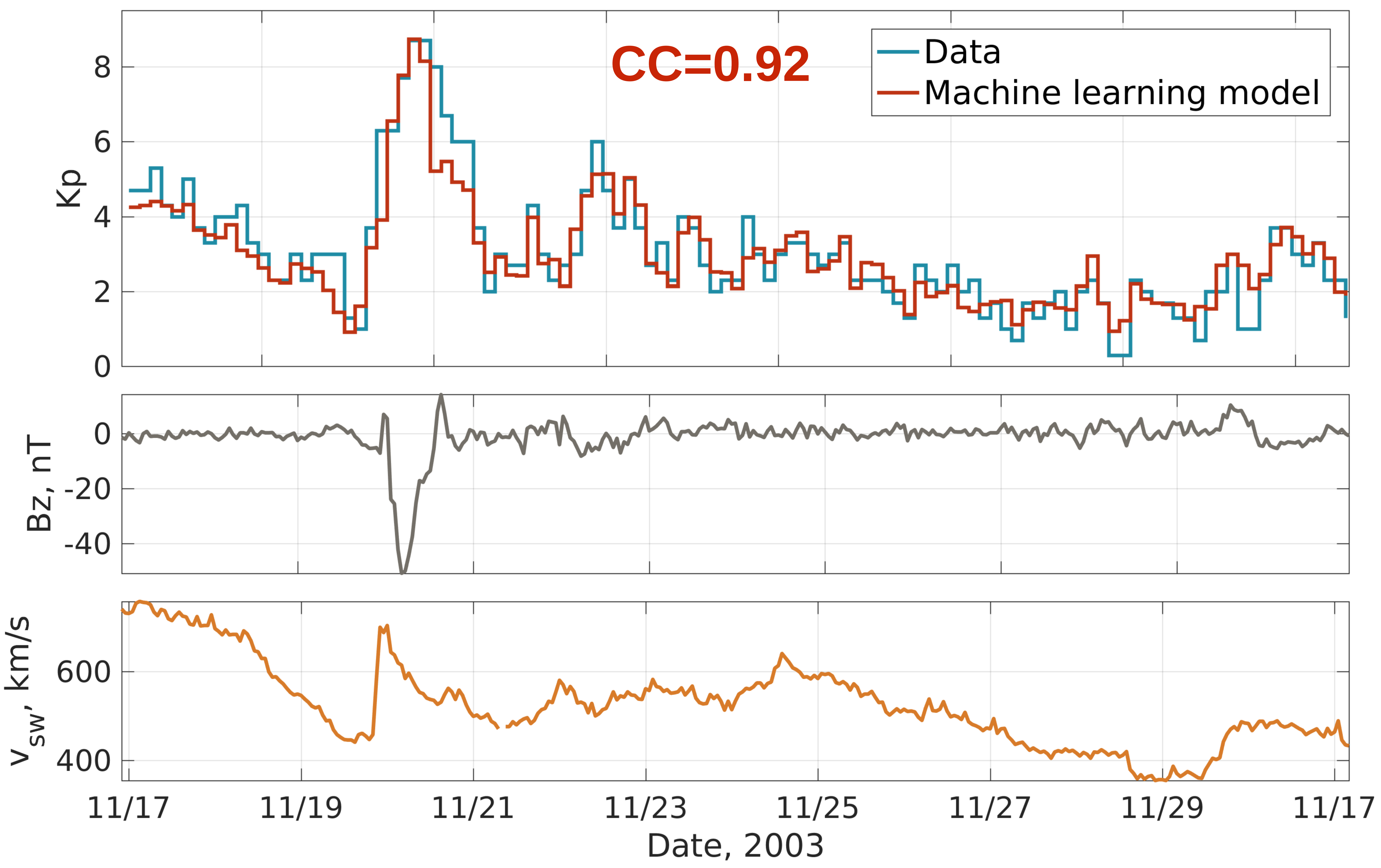
https://www.gfz-potsdam.de/en/kp-index/

- Global measure of geomagnetic disturbance due to solar particle radiation.
- Used worldwide as a standard to measure geomagnetic activity.
- Average value of disturbance levels in the horizontal components of the magnetic field at 13 selected stations around the globe.
- Input parameter for important near-Earth space models, like air-drag, radiation belt, diffusion coefficients, plasmapause, thermospheric etc. models.

GEOMAGNETIC ACTIVITY AND RISKS

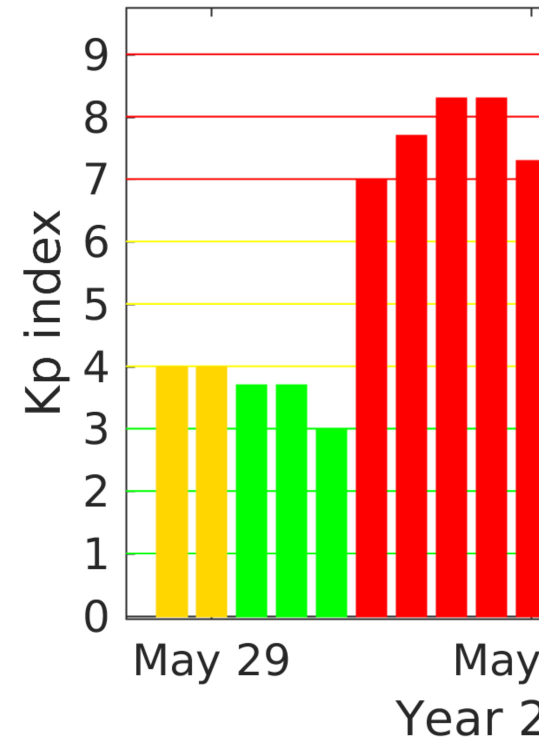
- Damages to satellites
  - Single event upset mechanism
  - Deep dielectric charging
  - Surface charging
- Disruption of power grids
  - Blackouts
  - Repair costs
- Radiation on humans
  - Long term space missions
  - Radiation exposure for frequent flyers

PRELIMINARY RESULTS



METHODOLOGY

- Methodology: neural network based empirical modelling.
- Data: solar wind and IMF data from ACE (available at OMNIWeb), Kp index from GFZ, 1993 - 2017.
- Model (input and internal neural network parameters) selection: nested cross validation.



Training data

Input: Averages, min, and max of (1) solar wind speed, (2) proton density, (3) IMF B, (4) IMF Bz over 0-3, 3-6, 6-9 hours.

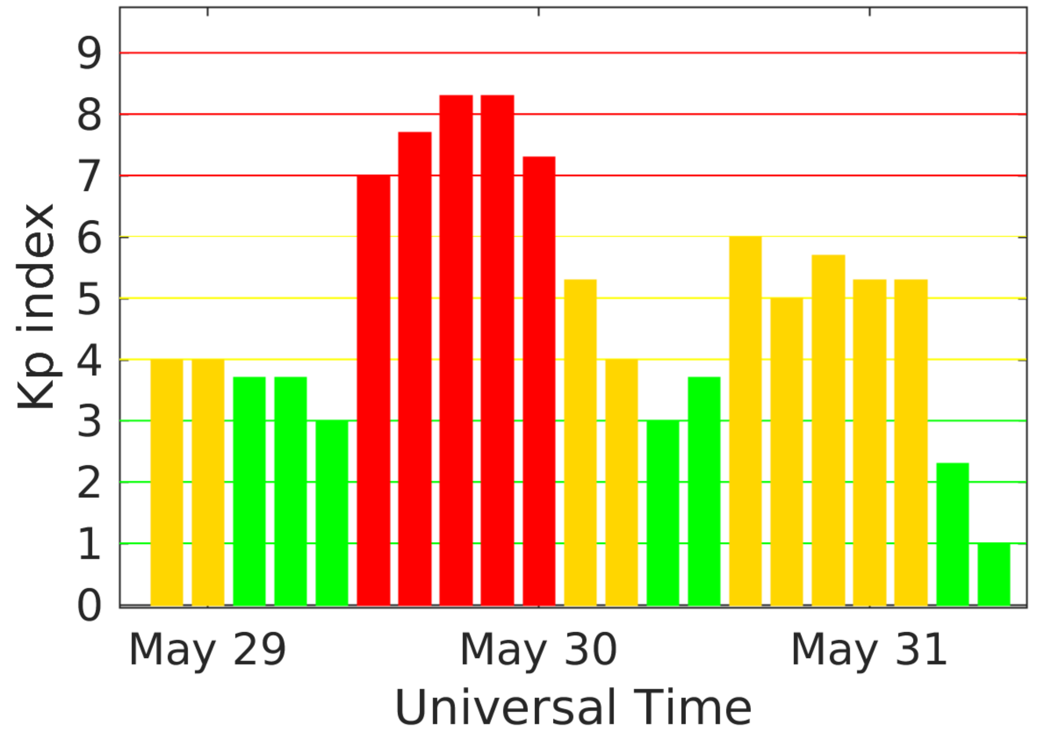
Output: Kp index.

→

Neural networks

→

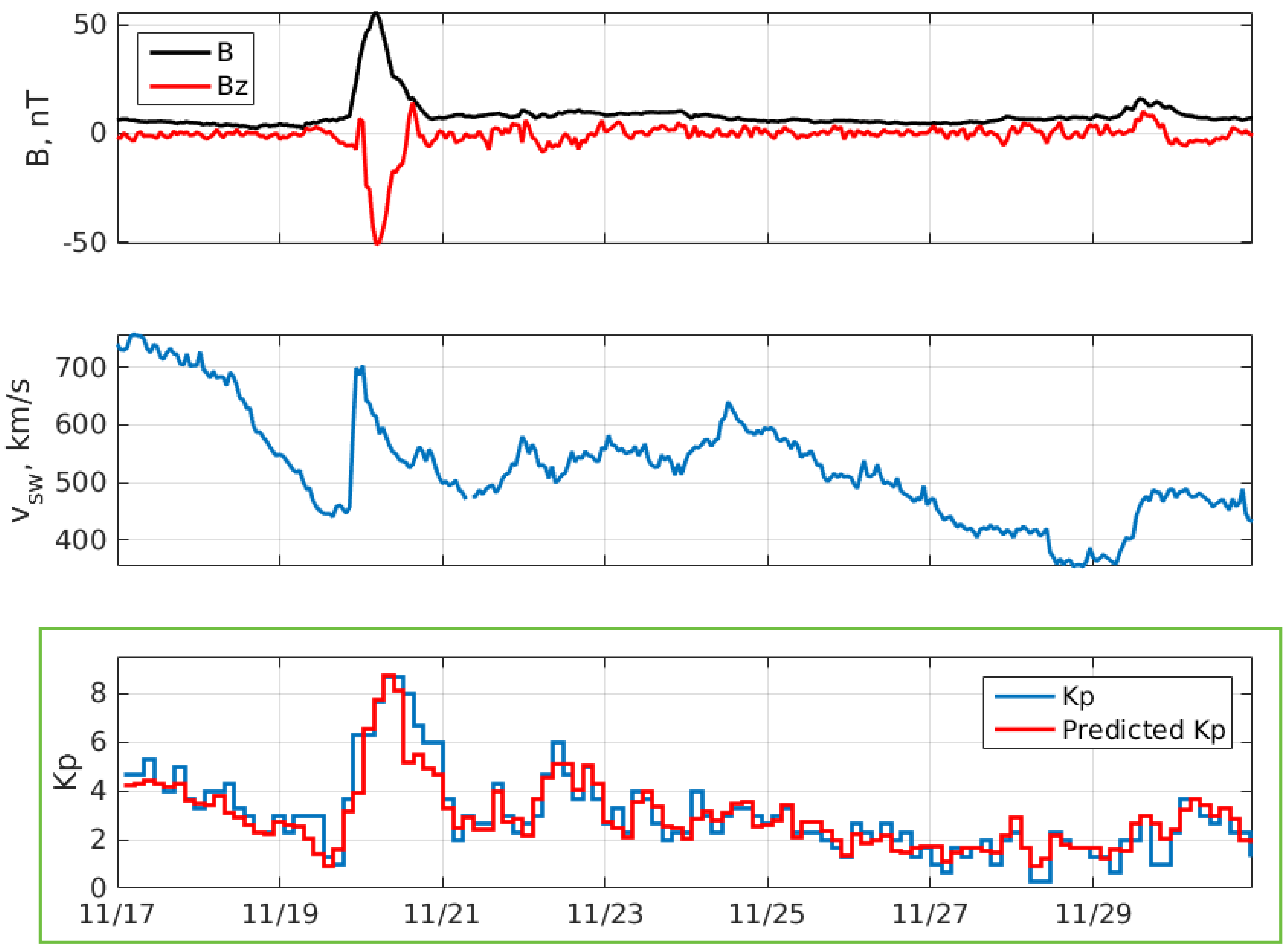
Kp prediction



MODEL PERFORMANCE

- Correlation coefficient between observed and predicted *Kp* on the training, validation, and test sets = 0.92.
- RMSE on the training, validation, and test sets = 0.55, 0.55 and 0.56.

Example of Kp prediction for 17 Nov - 01 Dec 2003
(out of training sample)



KP PARAMETERISED MODELS

- T89 Magnetic field model

  N.A. Tsyganenko, Planet. Space Sci., Vol. 37-1, 5-20, 1989

- Carpenter-Anderson plasmapause model

  D.L. Carpenter and R.R. Anderson, J. Geo. Res., Vol 97, 1097-1108, 1992

$$\boxed{L_{pp} = 5.6 - 0.46 \langle Kp \rangle_{24h}}$$

- Brautigam-Albert Radial radial diffusion model

  D.H. Brautigam and J.M. Albert, J. Geo. Res., Vol 105, 291-309, 2000

$$\boxed{D_{L\,L}(Kp\,,\,L) = 10^{0.506\,Kp - 9.325}\,L^{10}}$$

- DTM 2013 Thermospheric model

  S. Bruinsma, J. Space Weather Space Clim. , 5 A1, 2015

EXISTING KP PREDICTION MODELS

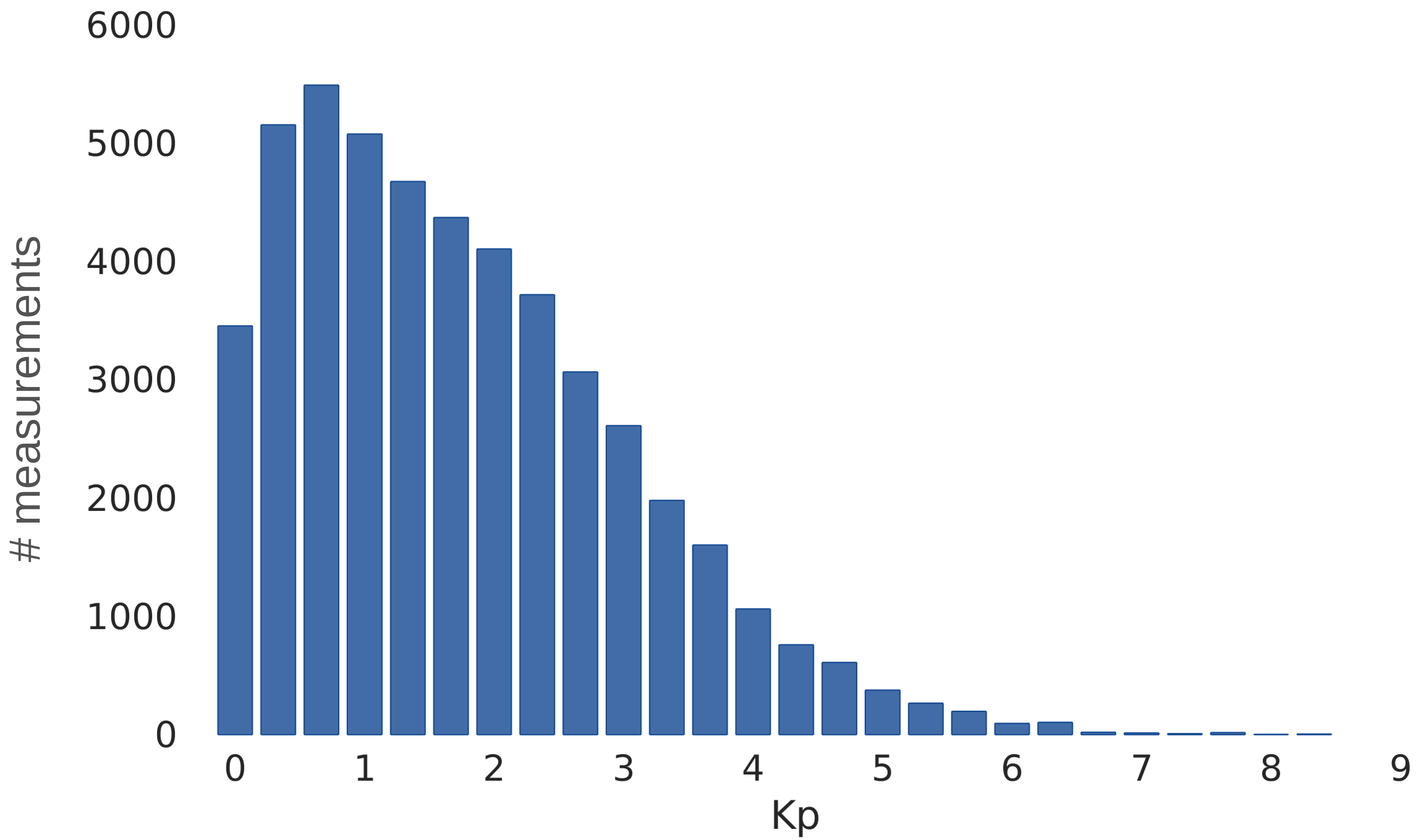
- *Costello et al*., 1997
- *Boberg et al*., 2000
- *Bala et al*., 2012
- *Wing et al*., 2015
- *Wintoft et al*., 2017
- *Tan et al*., 2018

ISSUES

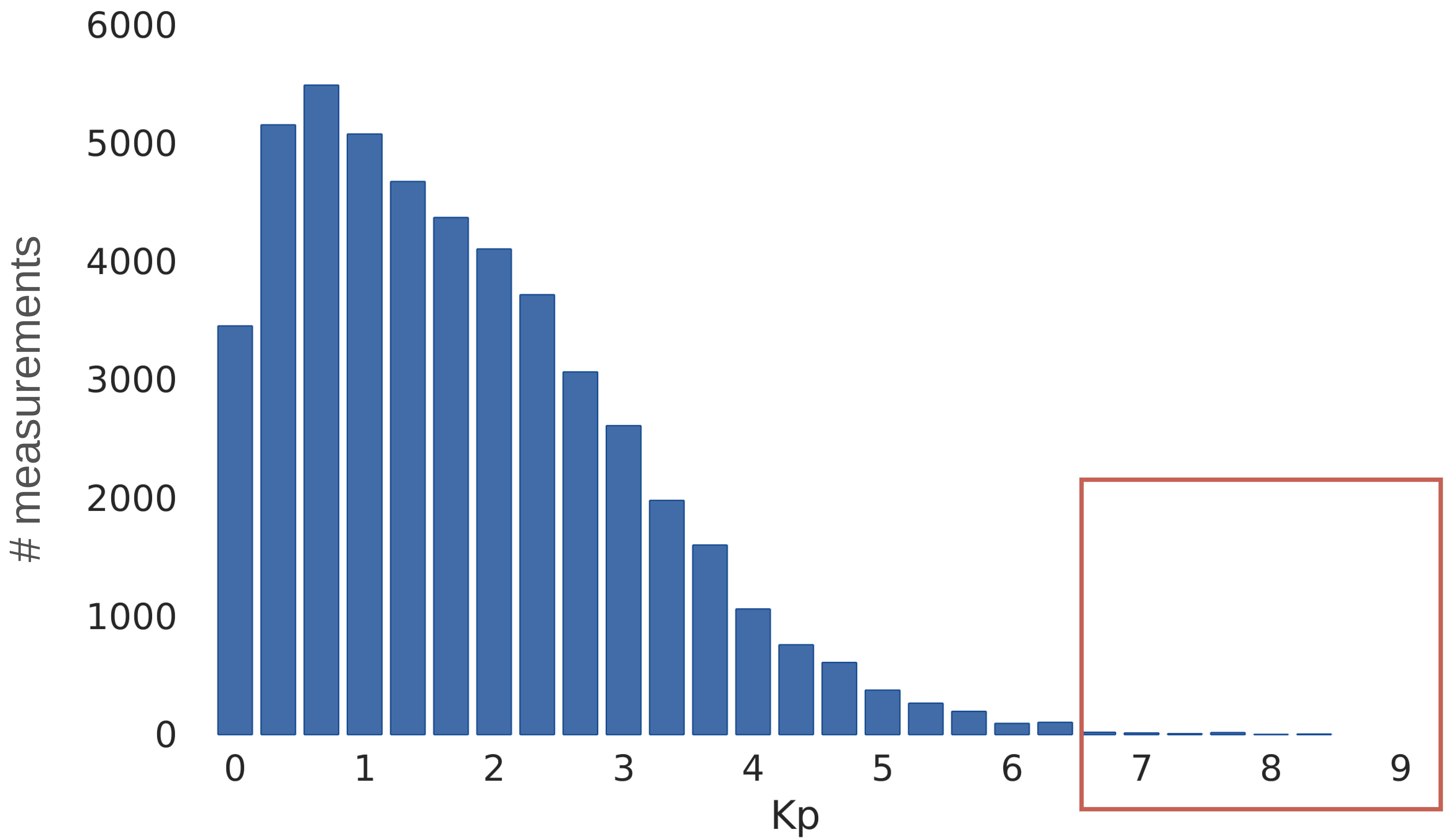
1. No common / Ambiguous names for different forecast horizons - no agreed upon convention for names.
2. Not always the same performance metrics are used for evaluation of models - hard to compare between models.
3. No common / agreed-upon way to split data into training, validation, and test sets.
4. Data augmentation techniques should not affect the test set.

NEURAL NETWORKS LEARN FROM DATA

NEURAL NETWORKS LEARN FROM DATA…



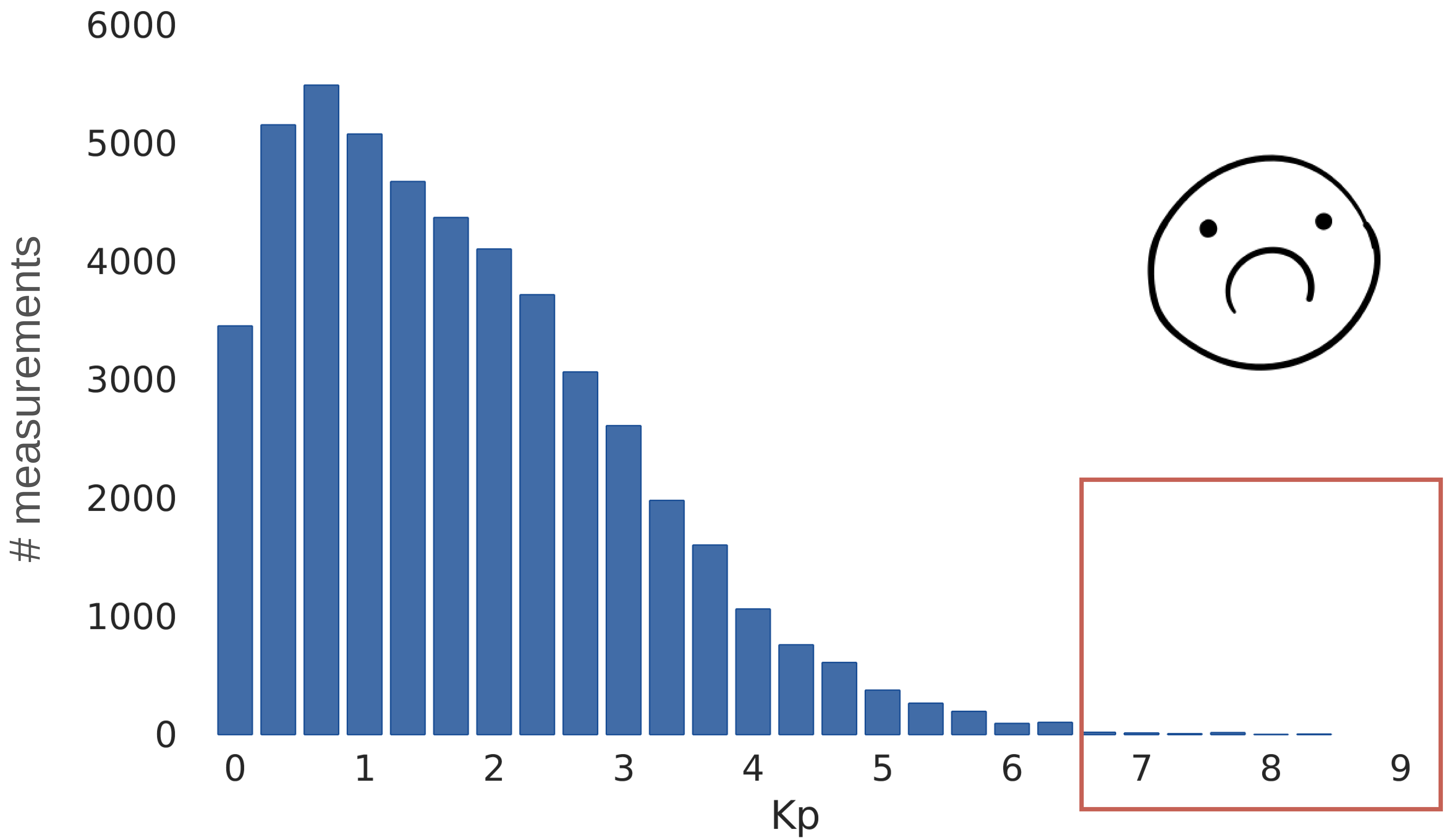
NEURAL NETWORKS LEARN FROM DATA…



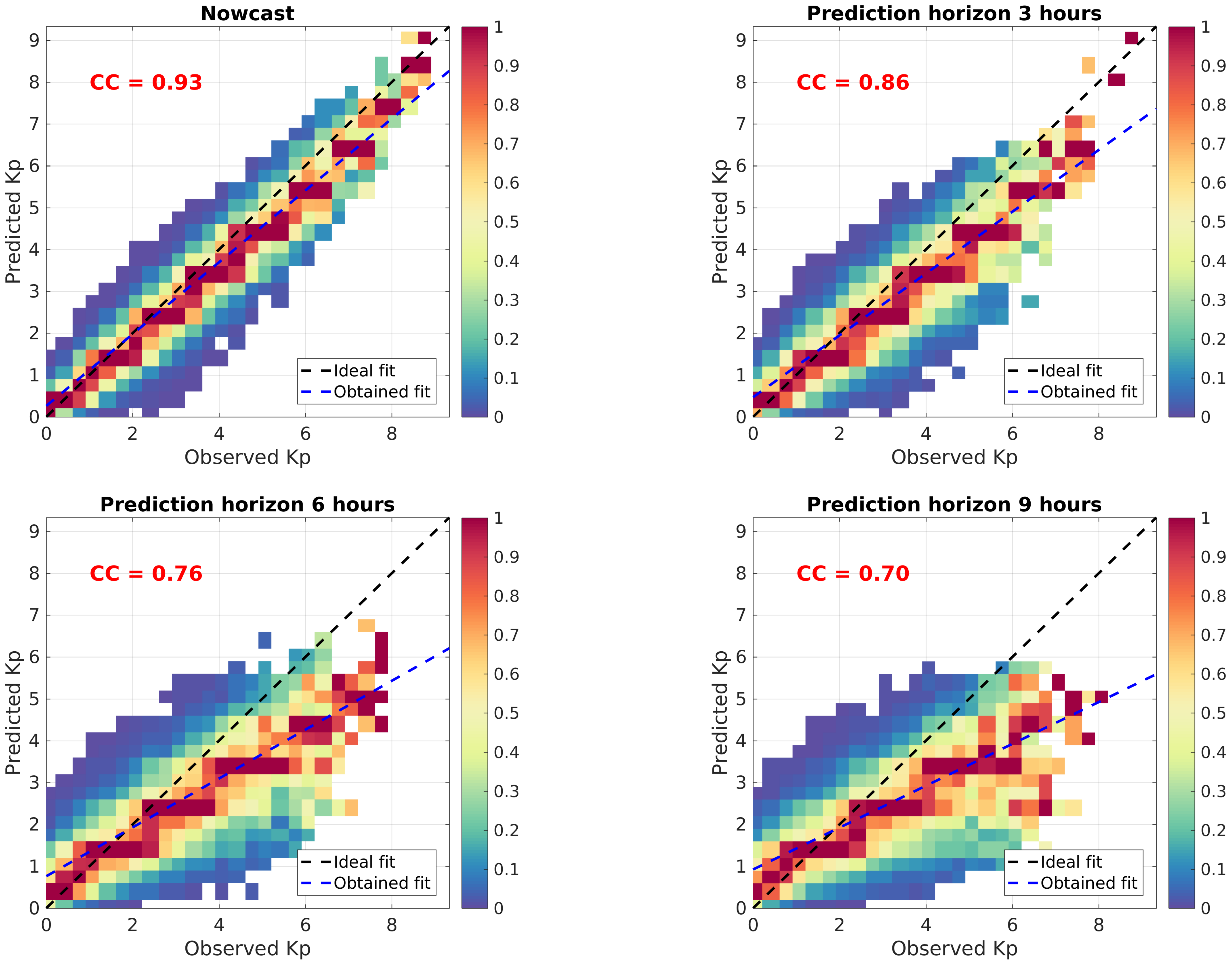
GFZ
Helmholtz Centre
POTSDAM

HELMHOLTZ

NEURAL NETWORKS LEARN FROM DATA...



RESULTING MODELS



EXAMPLES OF PREDICTION