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]Earth, Planetary and Space Sciences, UCLA, [4]Institute of Earth and Environmental Science,University of Potsdam, Potsdam, Germany

irina.zhelavskaya@gfz-potsdam.de

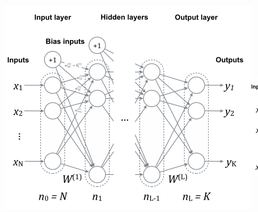
Abstract

The **Kp index** is a global measure of geomagnetic activity and it represents short-term magnetic variations driven by space weather. The Kp index is used as an input to various thermosphere and radiation belt models, and it is therefore important to predict it accurately. In this study, we systematically test how different machine learning techniques perform on the task of **nowcasting** and forecasting Kp for **3, 6, and 9 hours prediction** horizons. Additionally, we investigate two feature selection schemes based on **Mutual Information** and **Random Forests**. Finally, we evaluate and report the optimal combinations of input parameters and the best performing machine learning model.

Methods description

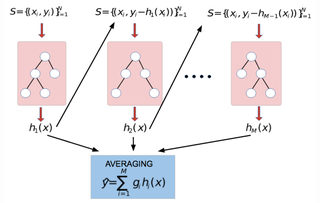
Feedforward Neural Network

A computational model that is very powerful for finding multivariate nonlinear relationships between input and output data. It has an input, output and a number or hidden layers. Each node in the layer is a Neuron, which can be thought of as the basic processing unit of a Neural Network. The information in the FNN moves only forward, from input to output (there are no feedback connections or loops).



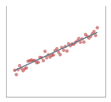
Gradient Boosting

Ensemble machine learning algorithm for classification and regression problems, which produces a hierarchy of weak prediction models iteratively. In a regression problem, every new model is trained to fit the residual between the actual target variable and the prediction value given by the previous model. It requires a differentiable loss function.



Linear Regression

A linear approach for modelling the relationship between a scalar dependent variable and one or more explanatory variables (or independent variables).

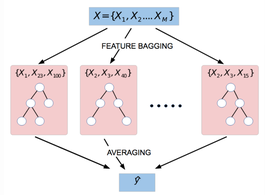


$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

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

Random Forest

Ensemble machine learning algorithm for classification and regression problems. It builds an ensemble of decision tree models, training each of them using a subset of input features obtained through feature bagging out of the training set. In a regression problem, the final model is the average of the ensemble models.



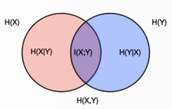
Mutual Information

The mutual information (MI) of two random variables is a measure of the mutual dependence between them. More specifically, it quantifies the "amount of information" (in units of bits) obtained about one random variable, through the other random variable. The concept of mutual information is intricately linked to that of entropy of a random variable defining the "amount of information" held in a random variable.

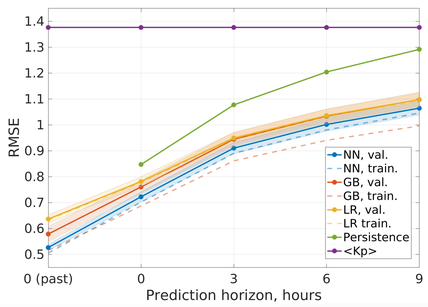
$$I(X_1;X_2) = H(X_1) + H(X_2) - H(X_1,X_2)$$

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

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



Performance of different ML methods



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%.

Existing models performance

Model (h=0, past)	RMSE	CC
Wintoft et al., 2017	0.55	0.92
Wing et al., 2005	-	0.92

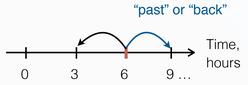


Table 1. Optimal inputs to the models derived from the CV procedure.

h = 0 (past), h = 0		h = 3	h = 6	h = 9
$Bz_{avg,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$Vsw_{avg,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$Bz_{avg,3\text{-}6,\,6\text{-}9,\,9\text{-}12}$	$Bz_{avg,6\text{-}9,\,9\text{-}12,\,12\text{-}15}$	$Bz_{avg,\,9\text{-}12,\,12\text{-}15,\,15\text{-}18}$
$Bz_{min,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$Vsw_{min,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$B_{avg,3\text{-}6,\,6\text{-}9,\,9\text{-}12}$	$B_{avg,6\text{-}9,\,9\text{-}12,\,12\text{-}15}$	$B_{avg,\,9\text{-}12,\,12\text{-}15,\,15\text{-}18}$
$B_{avg,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$nProt_{avg,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$By_{avg,3\text{-}6,\,6\text{-}9,\,9\text{-}12}$	$Vsw_{avg,6\text{-}9,\,9\text{-}12,\,12\text{-}15}$	$Vsw_{avg,\,9\text{-}12,\,12\text{-}15,\,15\text{-}18}$
$B_{min,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$nProt_{min,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$Vsw_{avg,3\text{-}6,\,6\text{-}9,\,9\text{-}12}$	$nProt_{avg,6\text{-}9,\,9\text{-}12,\,12\text{-}15}$	$nProt_{avg,\,9\text{-}12,\,12\text{-}15,\,15\text{-}18}$
$B_{max,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$nProt_{max,0\text{-}3,\,3\text{-}6,\,6\text{-}9}$	$nProt_{avg,3\text{-}6,\,6\text{-}9,\,9\text{-}12}$	sin(T), cos(T), sin(D), cos(D)	sin(T), cos(T), sin(D), cos(D)
sin(T), cos(T)	sin(D), cos(D)	sin(T), cos(T), sin(D), cos(D)		

T = 2π * (UT hour) / 24

D = 2π * (UT DoY) / 365,

DoY = day of year

Comparison of Mutual Information and Random Forests for feature selection

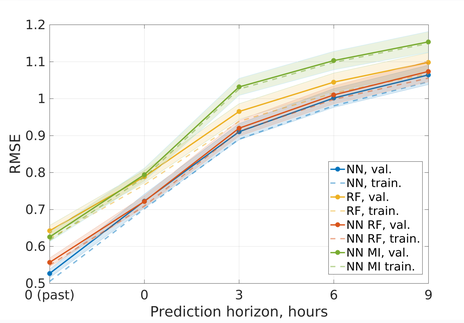


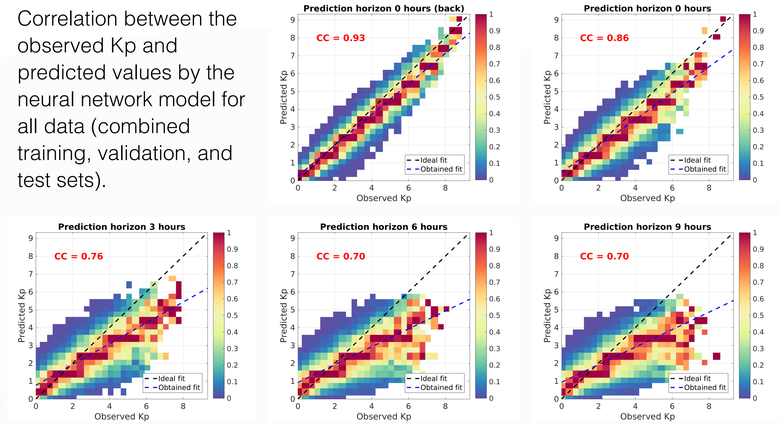
Table 2. Features selected by MI and RF (in the order of importance).

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

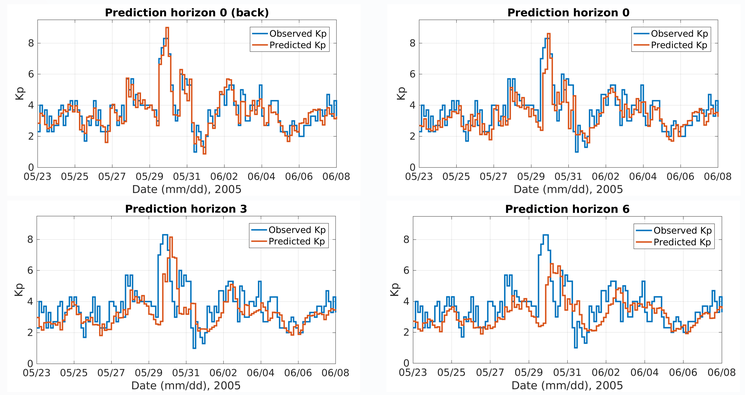
RF: h = 0	MI: h = 0	RF: h = 3	MI: h = 3	RF: h = 6	MI: h = 6	RF: h = 9	MI: h = 9
$Bz_{min,0\text{-}3}$	**$Bz_{min,0\text{-}3}$**			$B_{avg,6\text{-}9}$		$B_{avg,9\text{-}12}$	
$Vsw_{max,0\text{-}3}$	**$B_{max,0\text{-}3}$**			$Vsw_{avg,6\text{-}9}$		$Vsw_{avg,9\text{-}12}$	
$Vsw_{avg,0\text{-}3}$	$B_{max,3\text{-}6}$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

Resulting models

Correlation between the observed Kp and predicted values by the neural network model for all data (combined training, validation, and test sets).



Examples of Kp prediction for different horizons.



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 9 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 features selected by Random Forest perform similarly to the models based on features selected using the domain knowledge, while the input space is significantly reduced using the RF feature selection (models can be trained faster).

References

Wing S, Johnson JR, Jen J, Meng CI, Sibeck DG, Bechtold K, Freeman J, Costello K, Balikhin M, Takashi K. 2005. Kp forecastmodels. J Geophys Res 110: A04203. DOI: 10.1029/2004JA010500.

Wintoft P, Wik M, Matzka J, Shprits Y. 2017. Forecasting Kp from solar wind data: input parameter study using 3-hour averages and 3-hour range values. J. Space Weather Space Clim. 7: A29

Acknowledgements

This work was supported by H2020 project SWAMI. I.Z was funded by Geo.X, the Research Network for Geosciences in Berlin and Potsdam. Solar wind data and geomagnetic indices were obtained from http://omniweb.gsfc.nasa.gov/form/dx1.html. Kp index of geomagnetic activity was obtained from the GSFC/SPDF OMNIWeb interface at http://omniweb.gsfc.nasa.gov and produced by GFZ, Potsdam.