Forecasting AE indices using neural networks

Magnus Wik, Peter Wintoft and Juri Katkalov

Swedish Institute of Space Physics
Lund, Sweden

This work has been supported by the European Union's Horizon 2020 grant agreement No 637302 (PROGRESS)

Summary

- New forecast models for *AE*, *AL* and *AU* have been developed, with temporal resolution of 5 and 30 minutes, based on ACE level 2 data.

- The models are analysed considering, e.g. input parameters, network topology, lead time, and time delays.

- We use the “flat delay” propagation method to propagate ACE data to the magnetopause.

- Models have been developed using Python and the add-on libraries Scipy and Keras, where Keras is a minimalist Python library for deep learning.

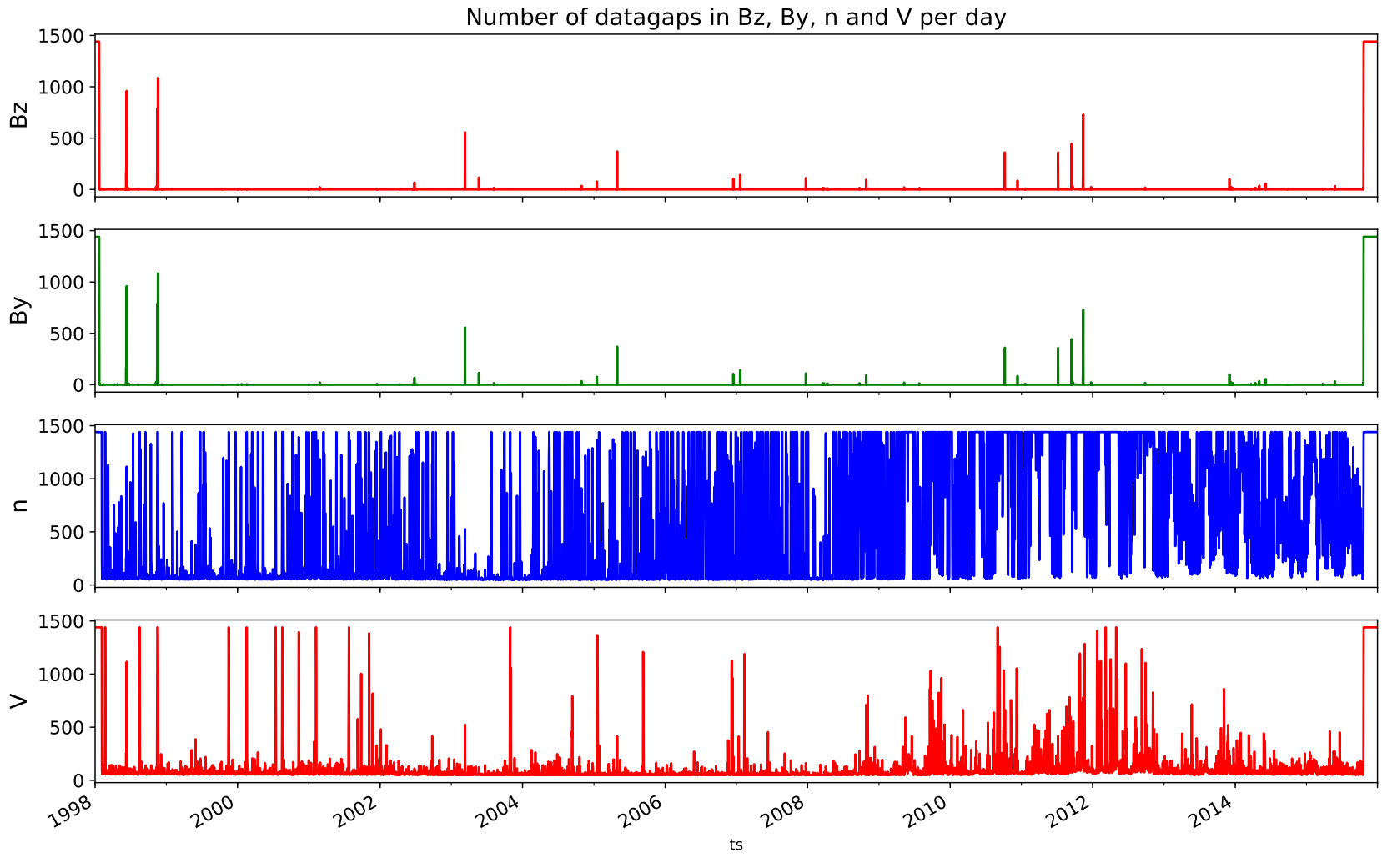
- Models use the feed-forward neural network algorithm with time delays up to 120 min.

- Inputs are B_z, B_y, B, speed, V, and density, n. Additional inputs are sine and cosine of UT hour and day of year (DOY).

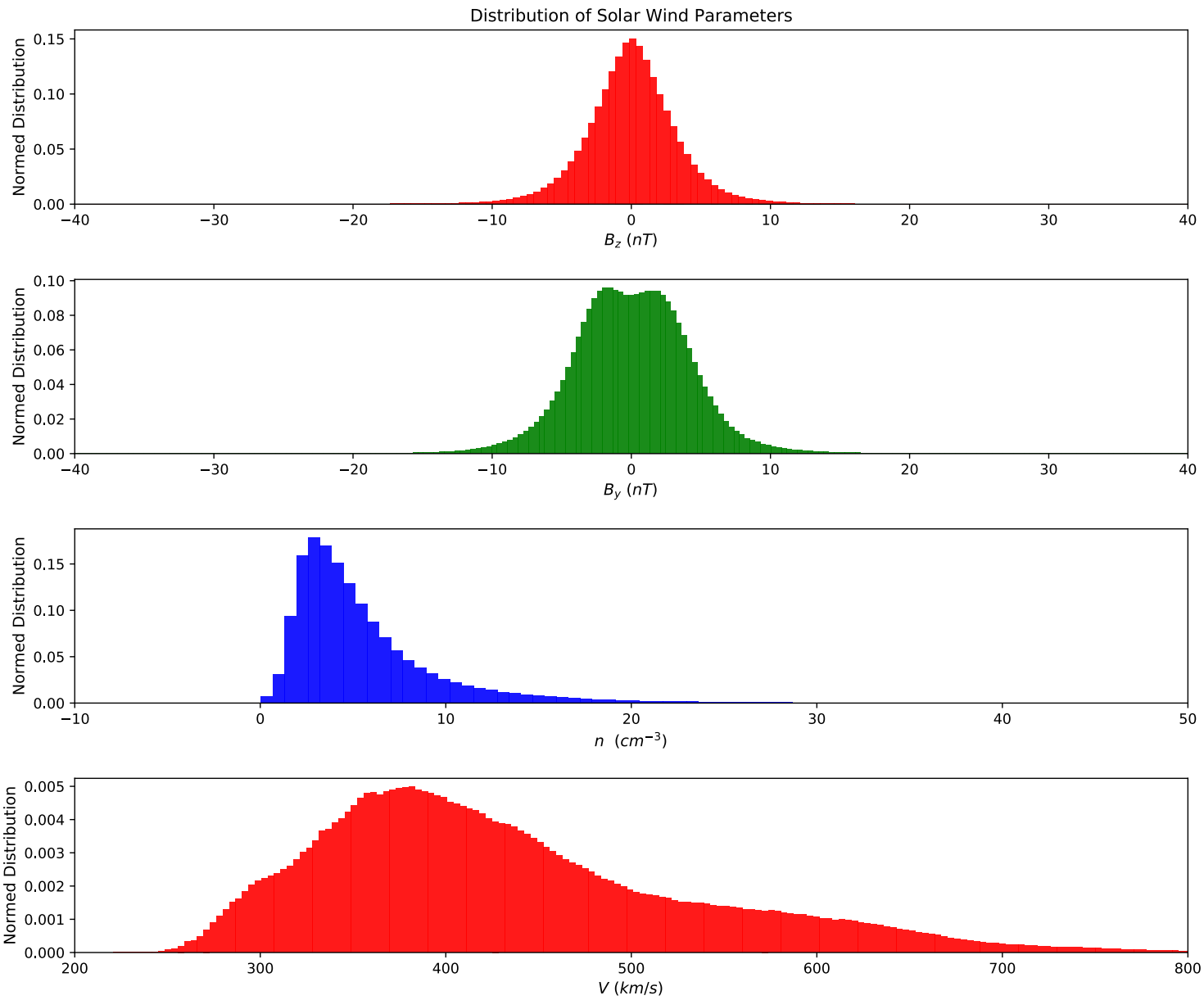
- We have used data from 1998 to 2015, in total 18 years. Training, validation and test sets compromise 10, 4 and 4 years.

ACE data

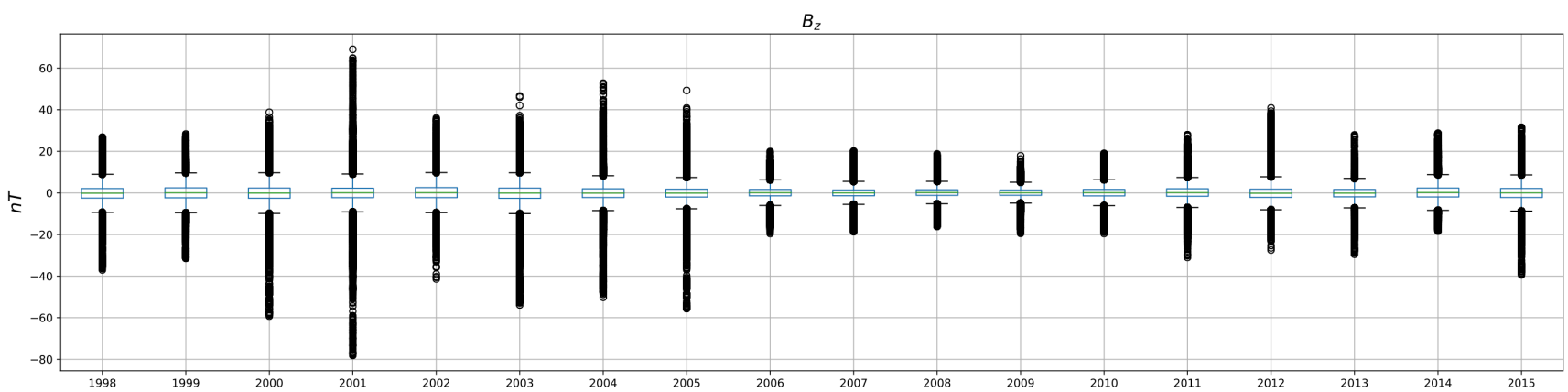
- The solar wind density, n, speed, V , IMF vectors B_z , B_y and the magnitude B, using the ACE level 2 data, for the period 1998 to 2015 were used for training the models.
- The solar wind density, has a data coverage of only 61%. This is partly due to plasma instrument outage during proton events. The speed and magnetic field vectors have coverage of 91% and 98%.
- The distribution for any parameter, also vary between each year. Our approach is to cover data for the whole period 1998 - 2015 into different data sets, to try and capture as much as possible of the variance in the data.
- To capture daily and seasonal variation, we also use the sine and cosine of UT and day of year.



ACE data



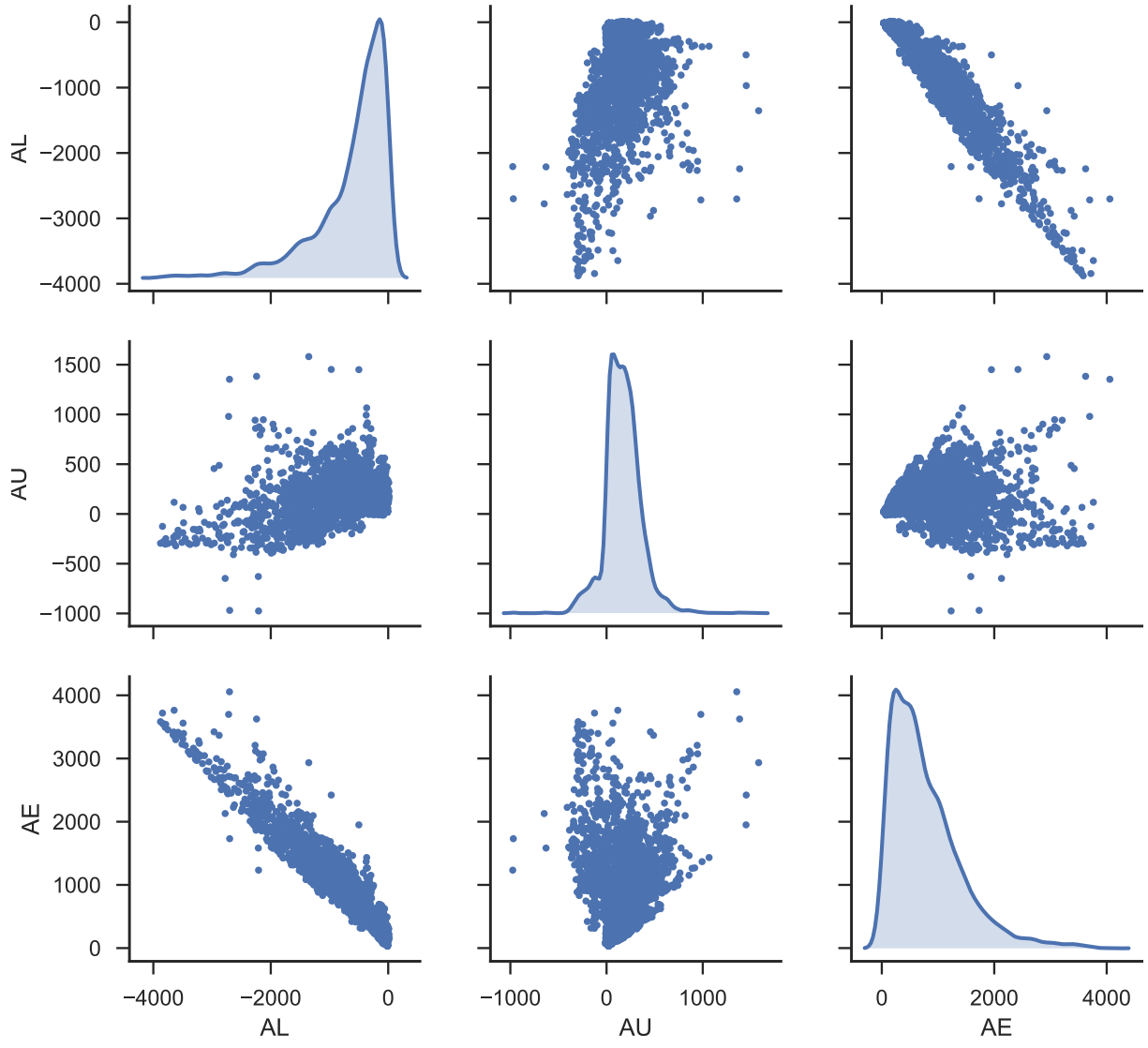
The distribution, for the years 1998 to 2015, of the solar wind parameters B_Z , B_Y , n and speed, V.



Box and Whisker plots of the solar wind parameter B_Z. The boxes, represent the middle 50% of the data, and the Whiskers represent the percentage of data outside the middle 50%. Data points beyond whiskers on the two sides are possible outliers, or extreme values.

AE indices

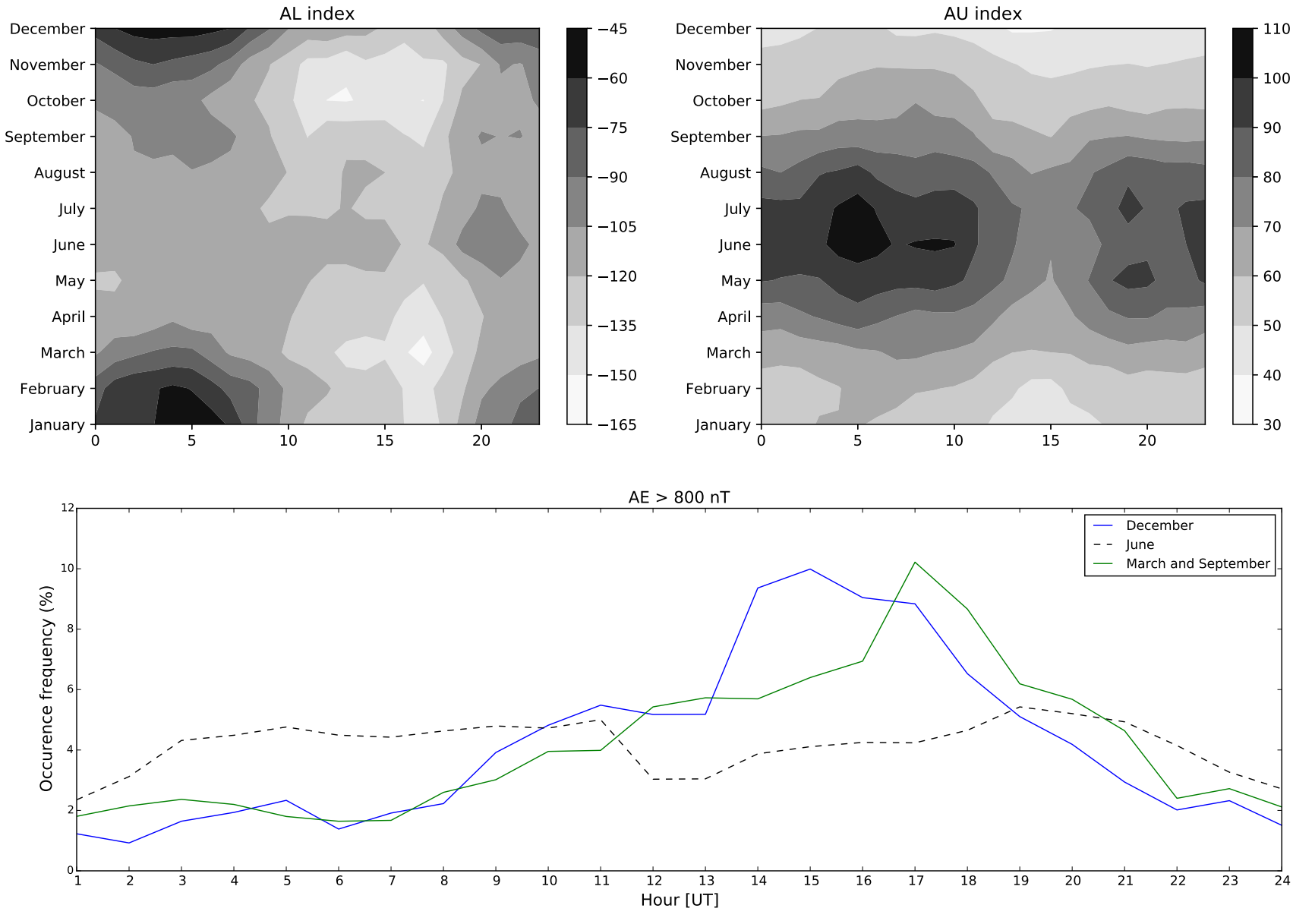
- We use the three geomagnetic indices *AE*, *AL* and *AU* as target data for training the models, where the *AE* index is the difference between *AU* and *AL*.
- As is shown, in the scatter plot matrix, for 28-30 October, 2003, the indices are clearly skewed. For all 18 years, the correlation between the *AE* index and the *AL* index is -0.96, and between *AE* and *AU* 0,83.



Scatter Plot Matrix - AE indices

AE indices

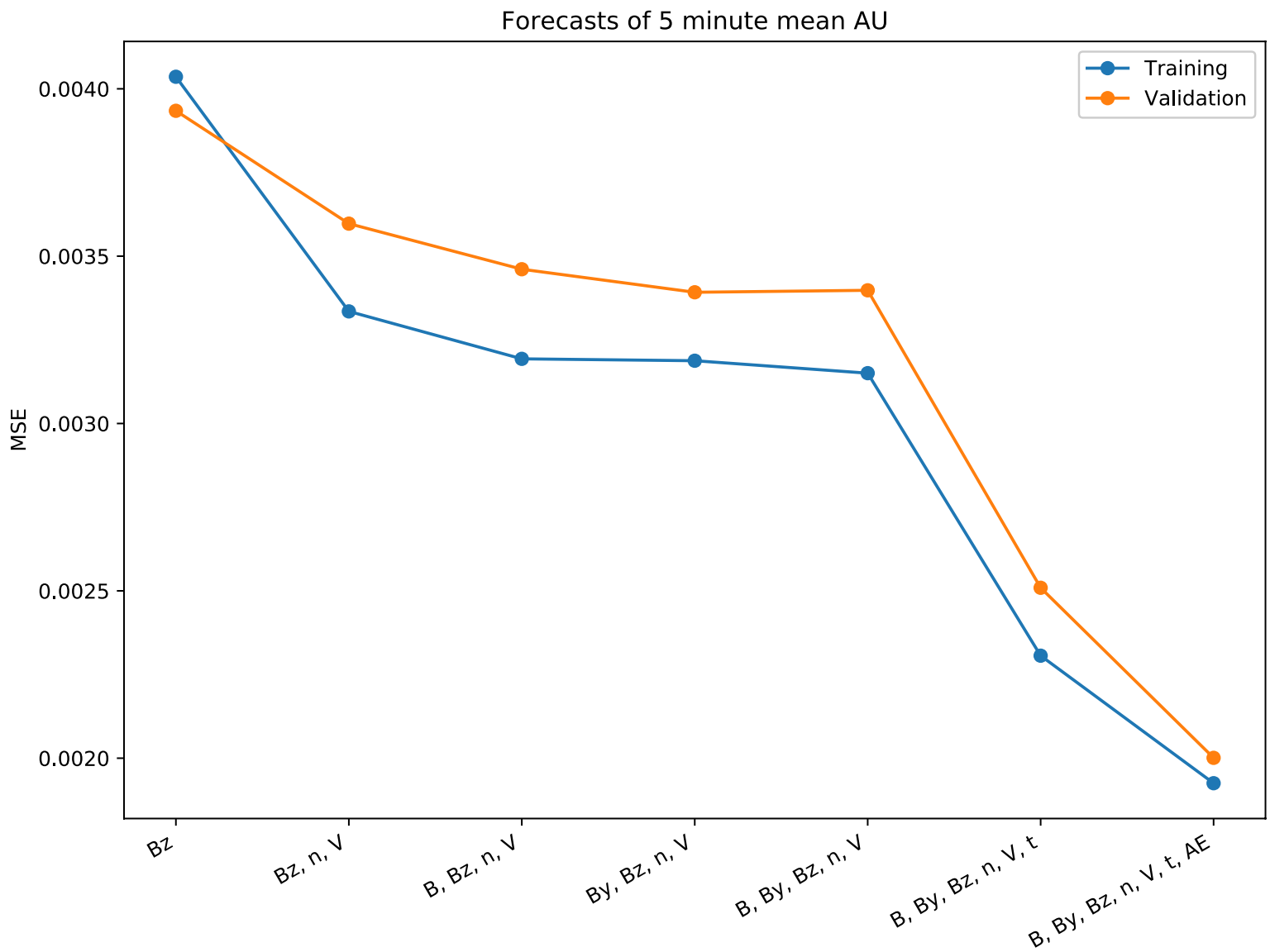
- A contour plot of the average *AL* and *AU* as function of UT and month is shown. For the *AL* index there is a minimum in the summer time, whereas for the *AU* index there is a maximum.
- Similarly we can also spot a UT variation, that changes during the year, for both *AL* and *AU* . A UT variation is also seen, for *AE* > 800 nT. During Winter, the occurrence frequency is highest around 15 UT, and at around 17 UT during the equinoxes. During summer these peaks have vanished.
- These results indicate that we need to add the seasonal and UT variation to the network inputs when training the models.

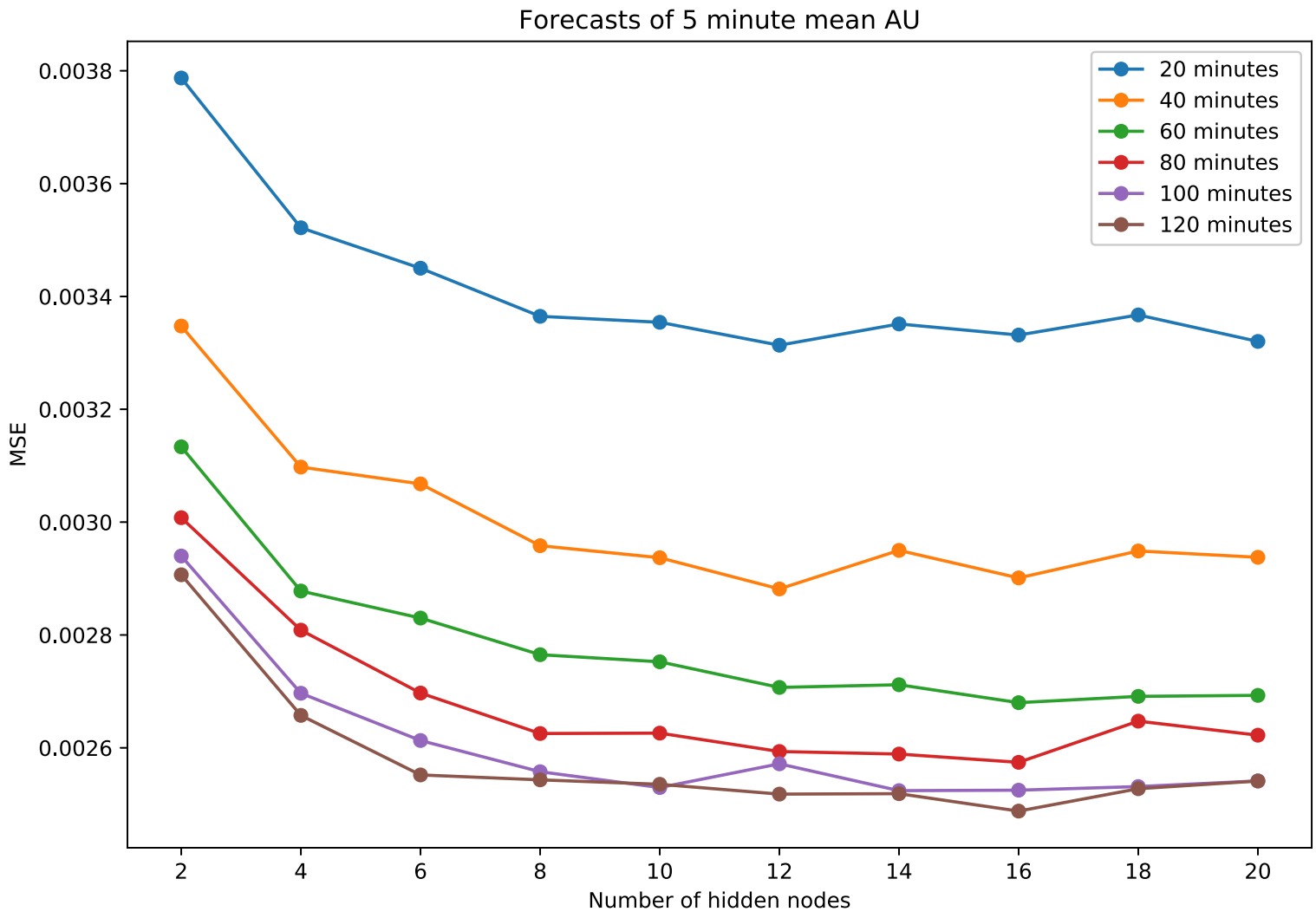


Model Training

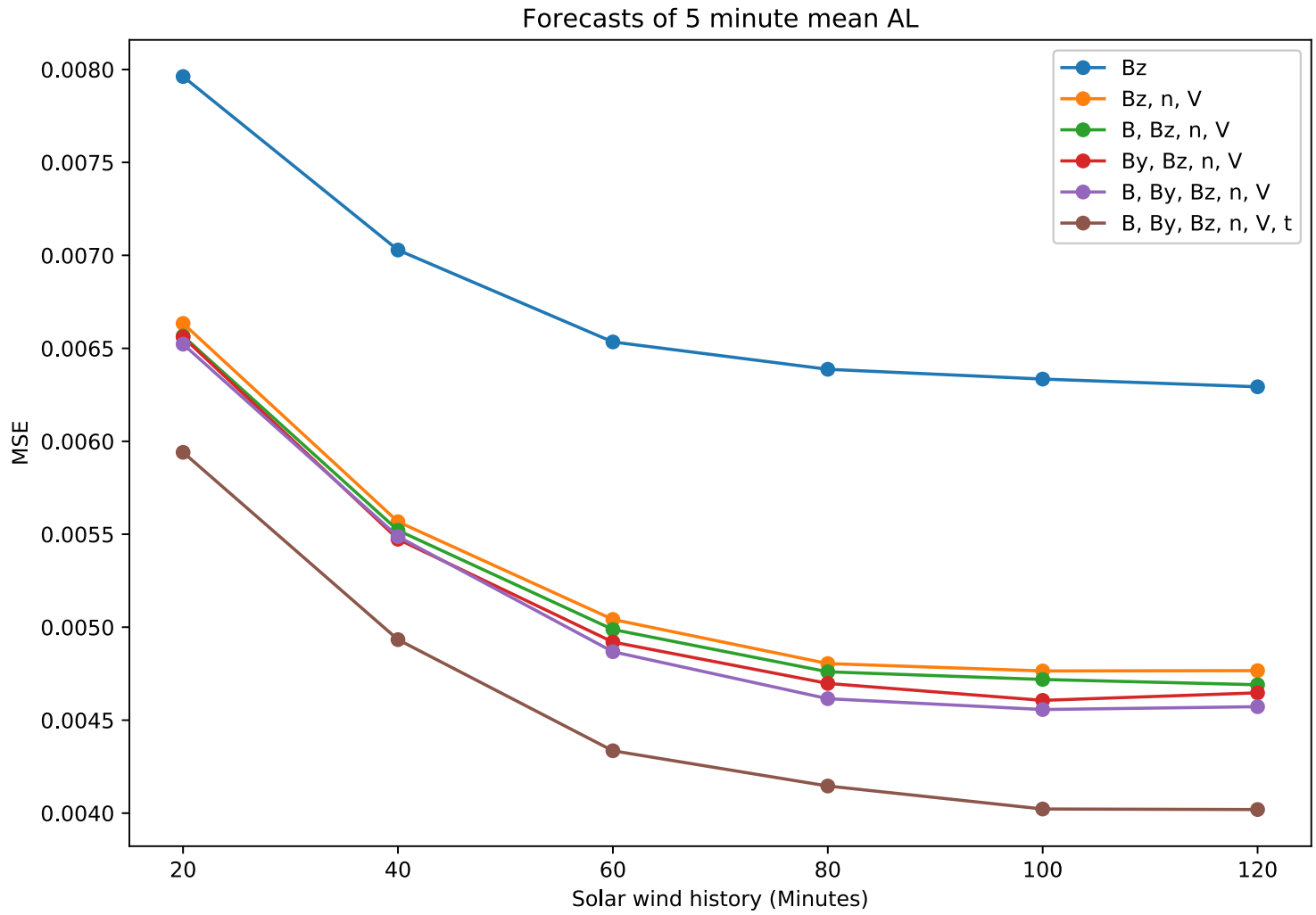
- The ACE data are first propagated to the location of the magnetopause. The data are resampled to 5 or 30 minutes, and we introduce time delays up to 120 minutes, to capture the dynamics (memory). For the *AE* indices we resample the original 1-minute data to 5 or 30 minutes.
- All data is rescaled, based on the training set, to be in the range [-1,1].
- The data were divided into three independent sets, the training, validation and test sets, consisting of 10, 4 and 4 years evenly distributed for 1998 to 2015.
- It is important to separate the training, validation and test set by at least the autocorrelation length of the input parameters.
- The autocorrelation drops to zero after about 2-3 hours at most. This have implications when selecting the rows of input data. Each row in our training data, consists of up to 120 minutes back in time.
- It is therefore not possible to randomly split any rows into the training, validation and test sets. We decided to use yearly data, with few and minimal overlaps between the data splits.
- For training we use the Back-propagation algorithm with the hyperbolic tangent (tanh) activation function in the hidden layers and the Adam optimizer.
- The networks consists of 1-3 layers, and up to 20 hidden nodes. For the model studies we ran 10 models, and for the final models we ran 20 models. The models with lowest validation error were then selected.

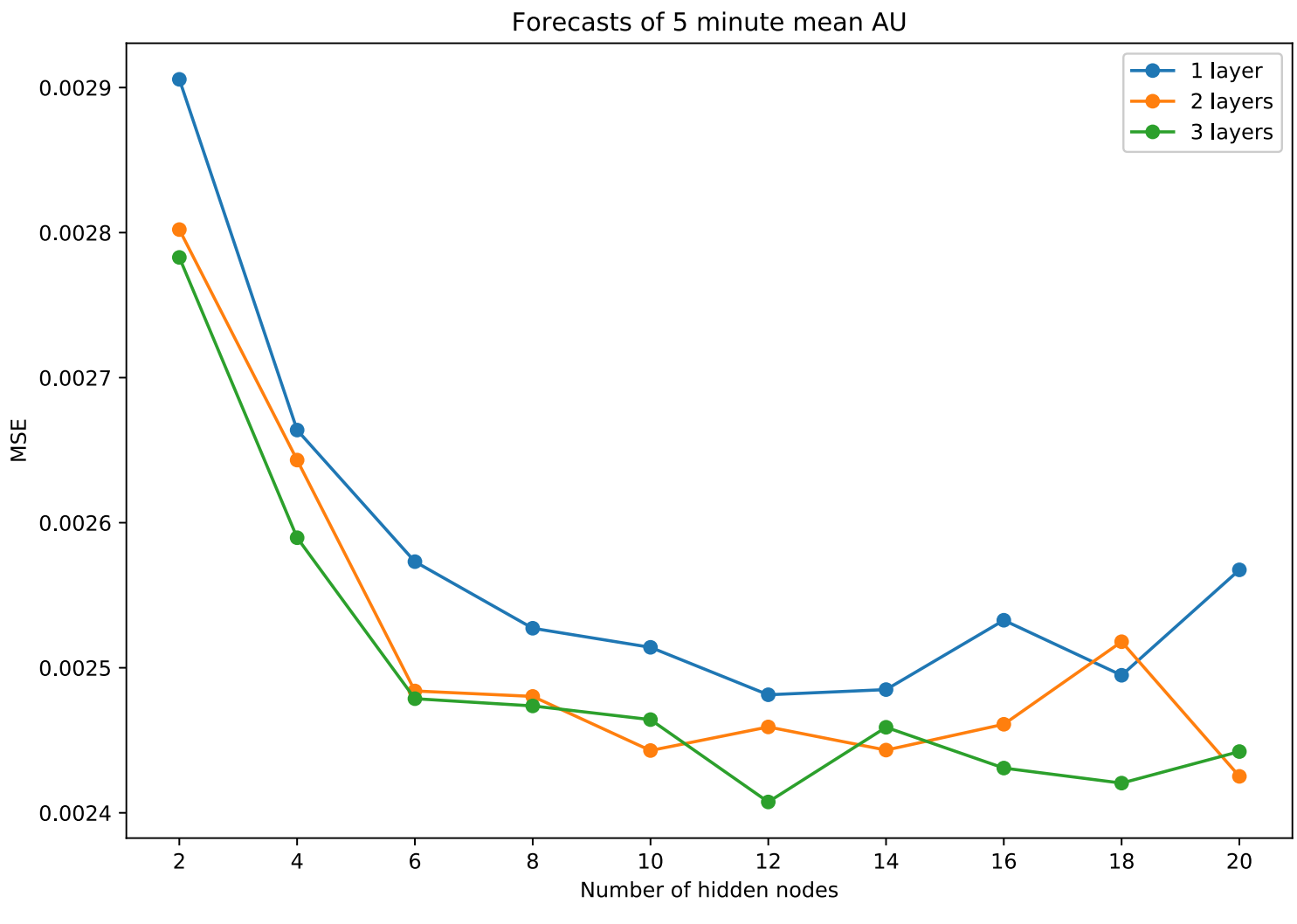
Model Studies



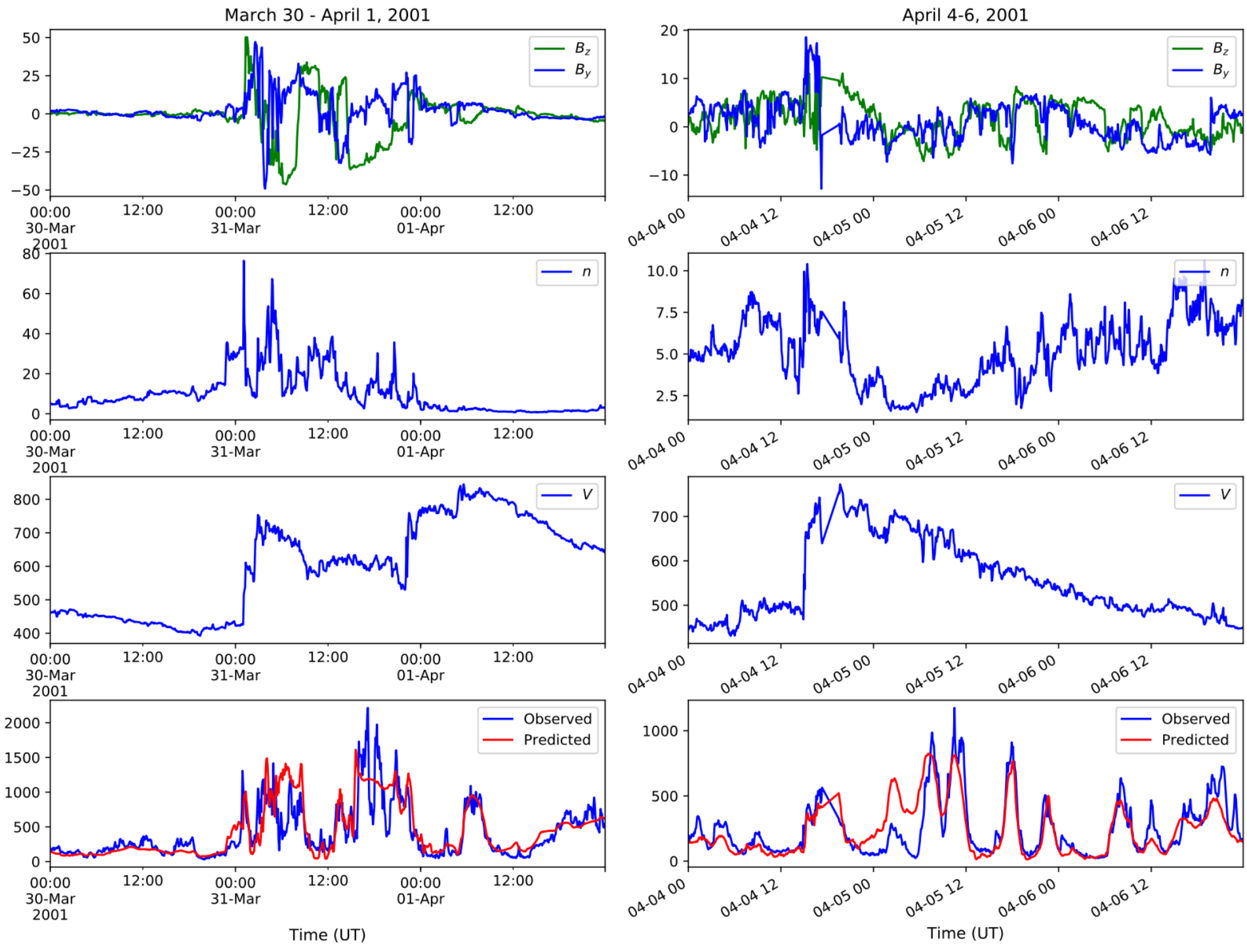


Model Studies



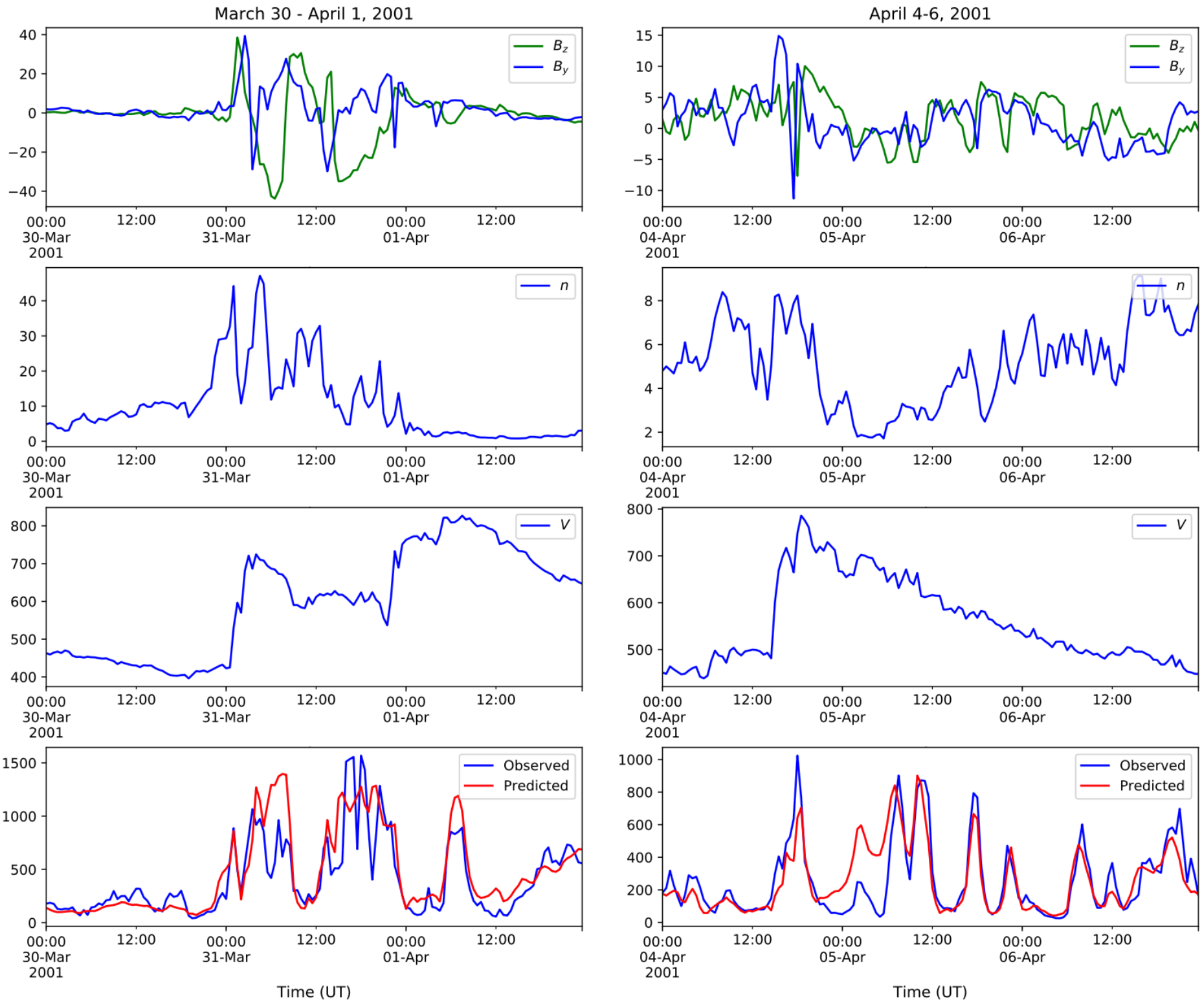


Forecast of 5 minute AE mean



Forecast of 5 minute averaged *AE* during two events, from the test set, in 2001. The top three panels show the propagated 5 minute averaged solar wind magnetic field (B_Z and B_Y), density (n) and speed (V). The bottom panels show observed and predicted *AE*.

Forecast of 30 minute AE mean



Forecast of 30 minute averaged *AE* during two events, from the test set, in 2001. The top three panels show the 30 minute averaged solar wind magnetic field (B_Z and B_y), density (n) and speed (V). The bottom panels show observed and predicted *AE*.

Conclusions

- In this work we have developed, forecast models, using neural networks, for the three geomagnetic indices *AE*, *AL* and *AU*, with time resolutions of 5 and 30 minutes, with serial correlation of up to 0.88 and 0.9.

- Although *AE* (and *AL*, *AU*) are global indices, they show a seasonal and UT dependence. Therefore, *AE* indices data were selected to cover both different seasons, UT and years with both low and high solar activity.

- We performed model studies, with various inputs and different network topology, to find key features and network configuration, for best performance.

- With parameters n, V, B_Z, B_Y, B, UT hour and day of year, we achieve the lowest errors.

- There seem to be a limit in performance at about 100 minutes in time delays, which indicate that the magnetospheric system memory saturates at a time delay of about 100 minutes.

- The results indicate that an optimal network should use 2 hidden layers and at least 12 nodes per hidden layer in this case.

- A major improvement in the performance is seen when we also add time and the index itself as inputs.

- Later we will add error analysis and verification and use other algorithms such as SVR and LSTM.