Forecasting AE indices using machine learning

Magnus Wik and Peter Wintoft
magnus@lund.irf.se

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 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).

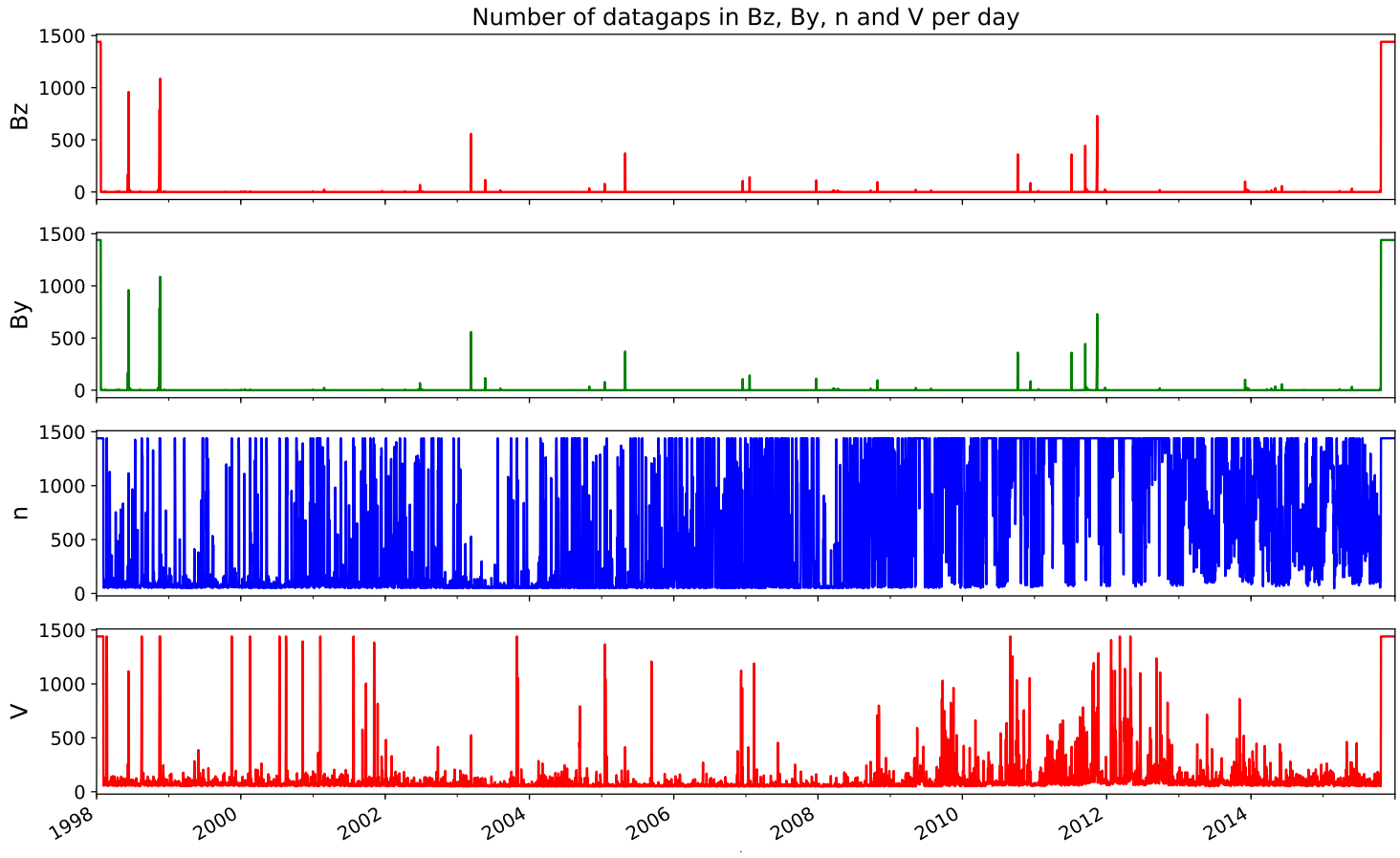
- We have used data from 1998 to 2015, in total 18 years. Training, validation and test sets compromise 10, 4 and 4 years.

- We achieve a serial correlation up to 0.88, and skill score of 0.6 for *AE.*

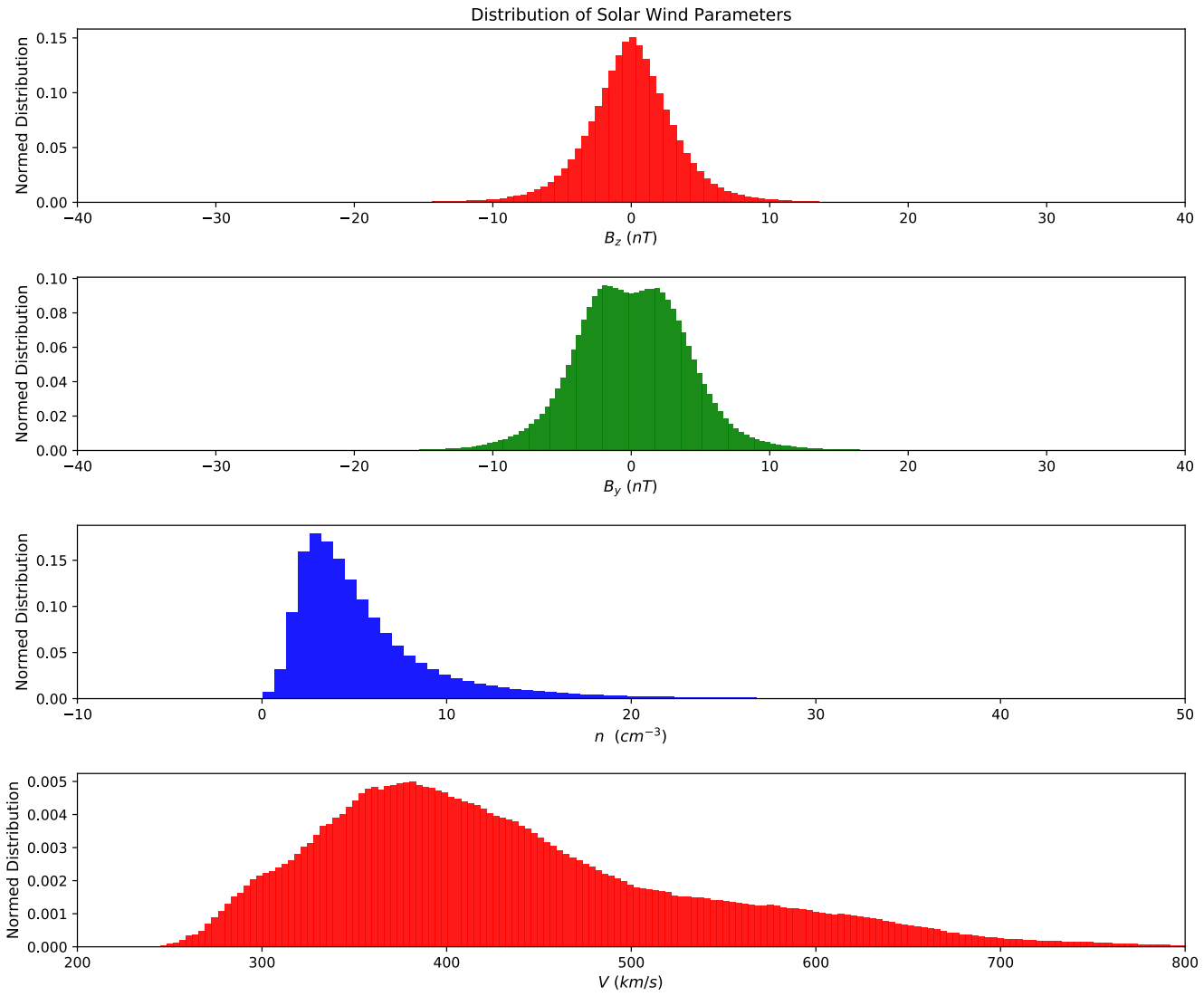
ACE data

- 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 as inputs 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, to try and capture as much as possible of the variance in the data.
- Due to the auto correlation, ~2-3 hours, we can not randomly split data between the data sets. We therefore use yearly data with few overlaps.
- To capture daily and seasonal variation, we also use the sine and cosine of UT and day of year.
- Data were divided into training, validation and test sets. Before training the models, we normalised the data based on the training set.

ACE data



Number of datagaps for ACE parameters

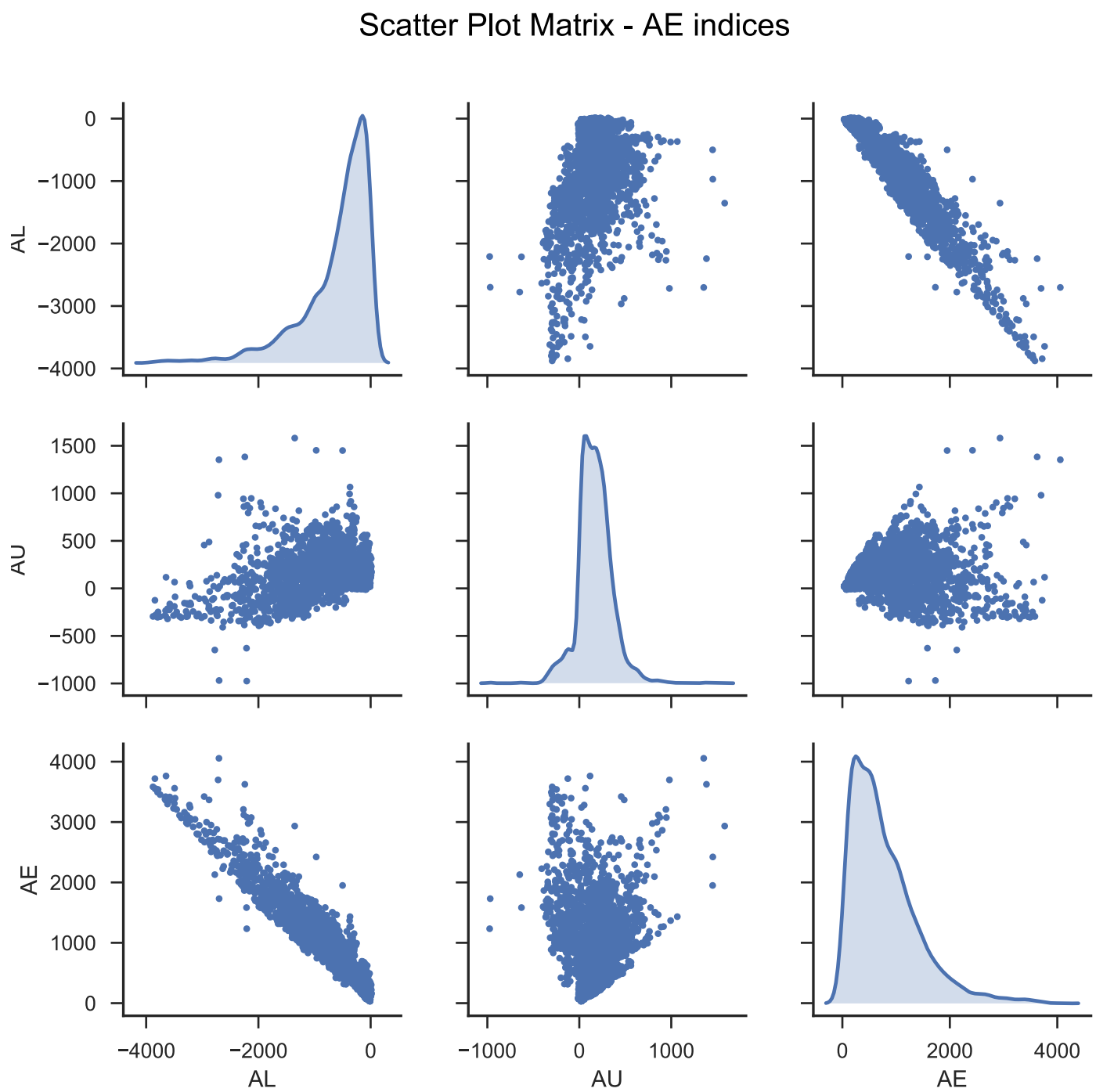


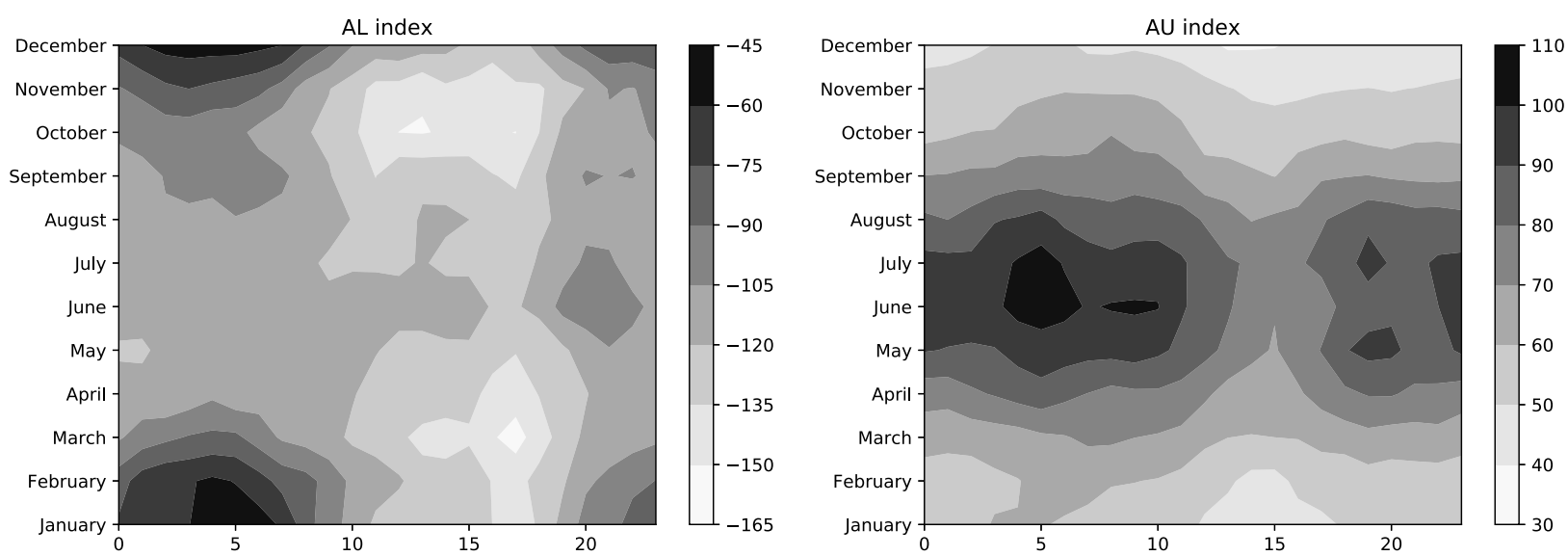
The distribution, for the years 1998 to 2015, of the solar wind parameters B_Z , B_Y , n and speed, V.

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.
- 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, as described earlier.

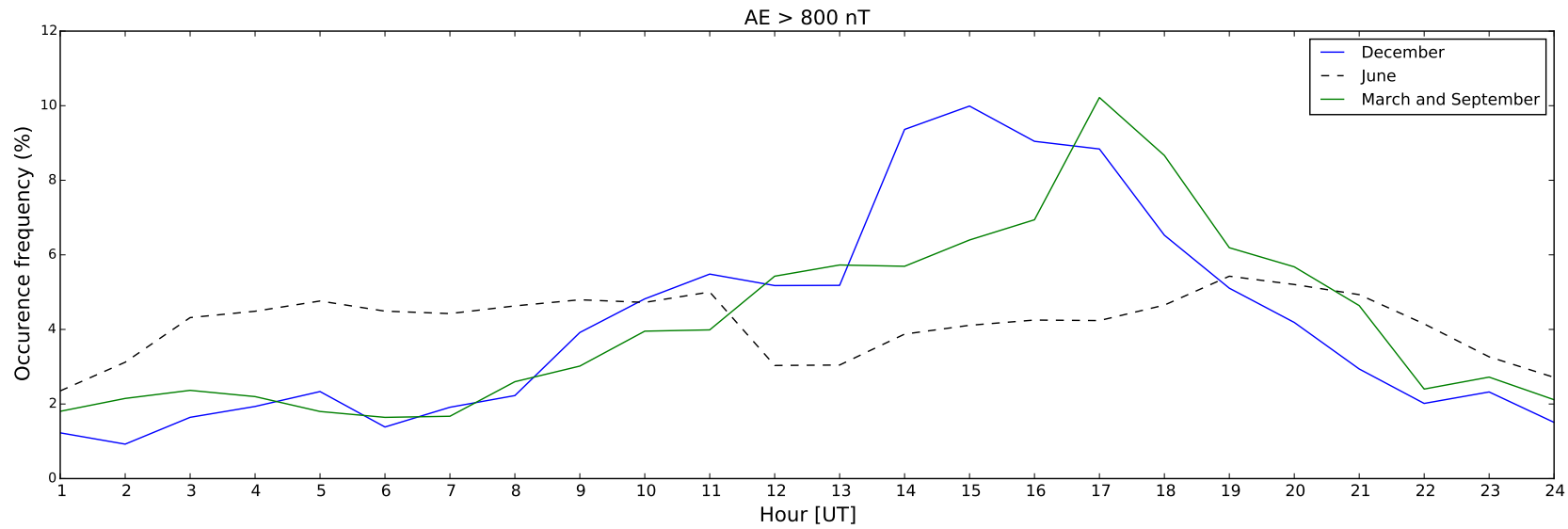
AE indices

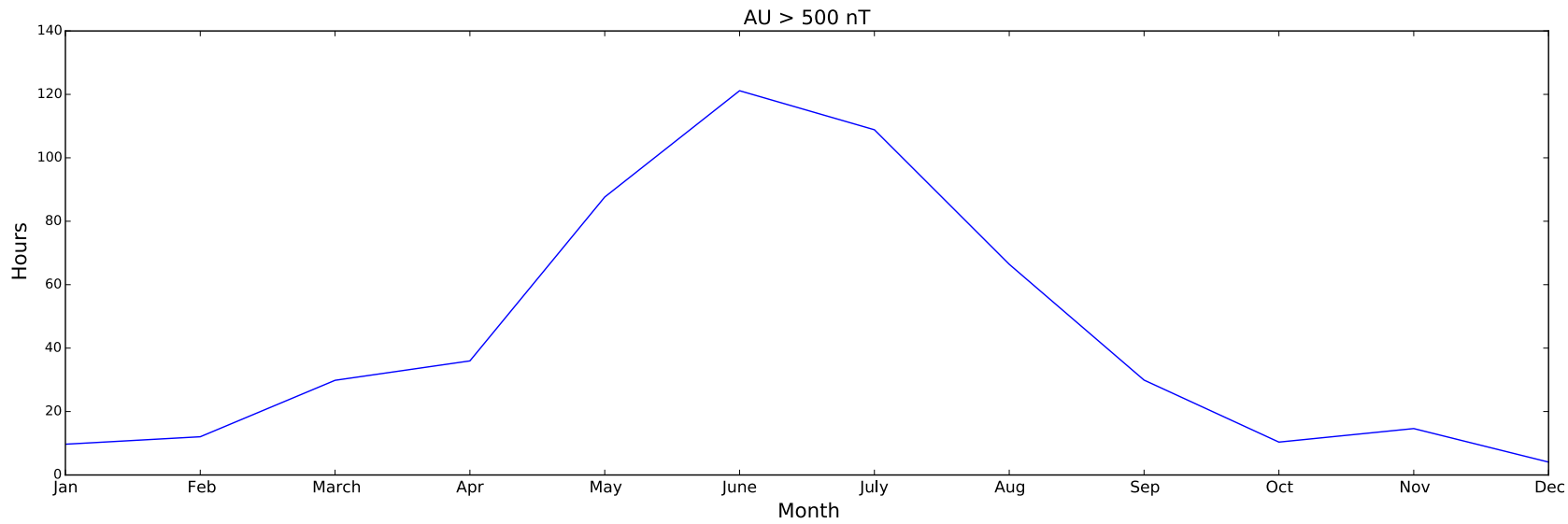


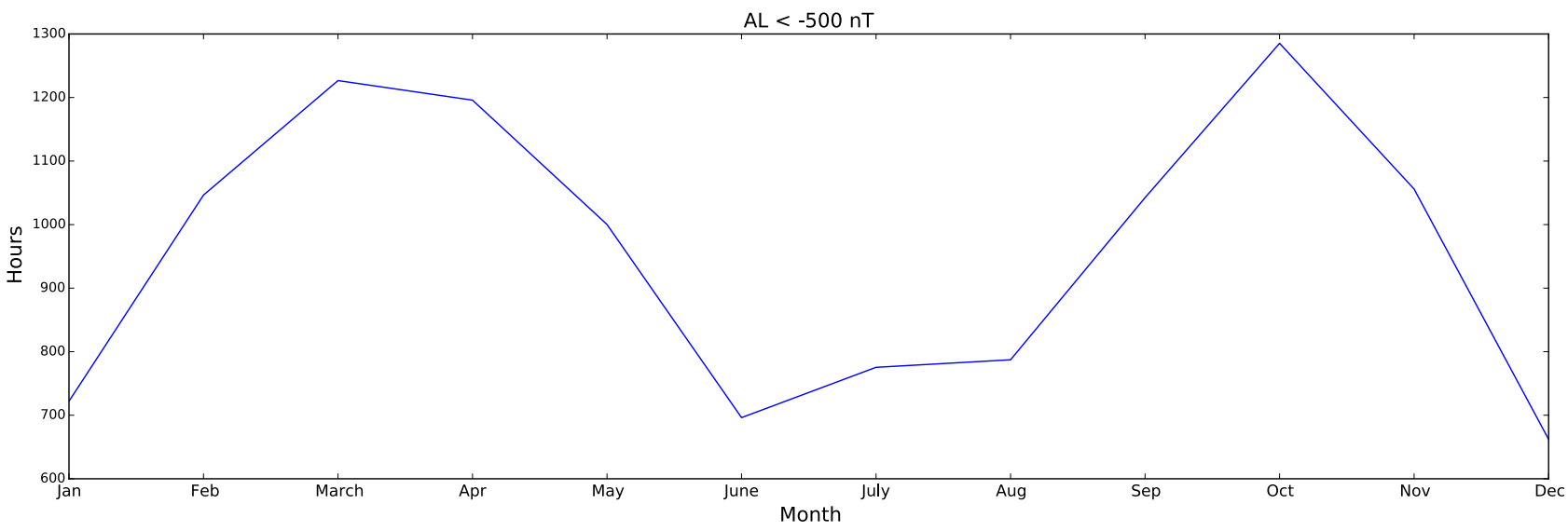


A contour plot of the average *AL* and *AU* as function of UT and month.

AE indices





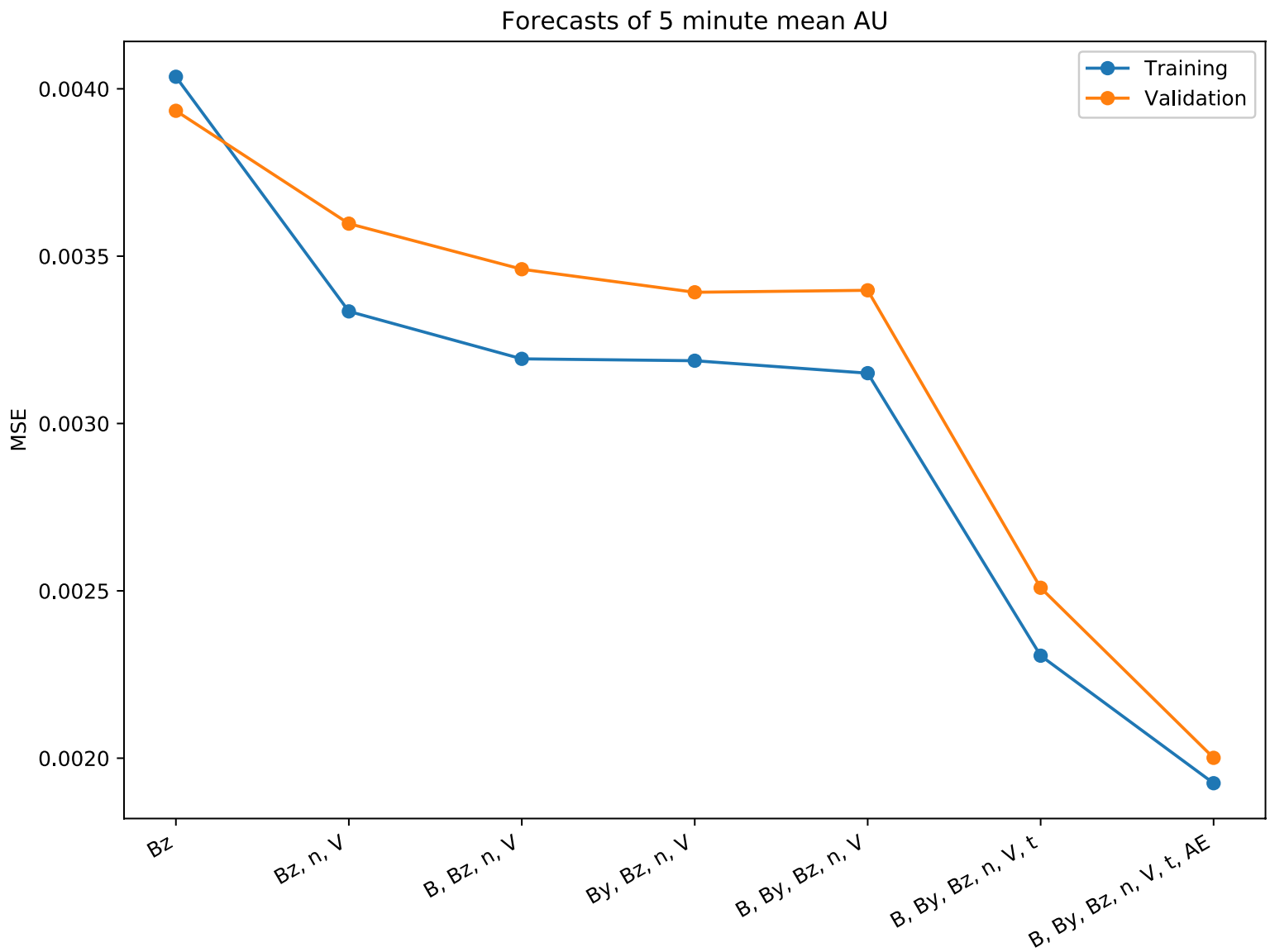


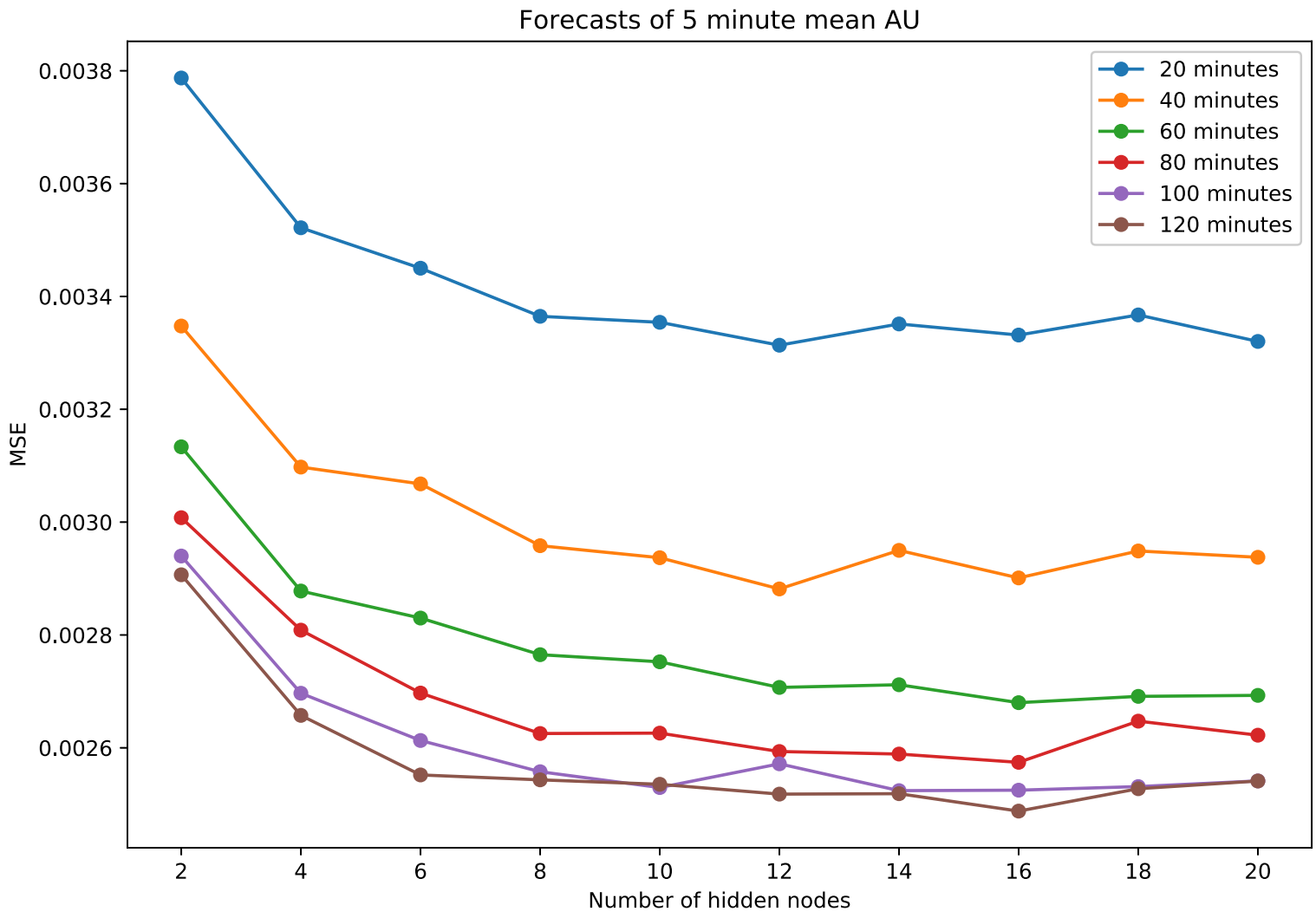
Plots showing the UT and season dependence for the *AE* indices.

Model Training

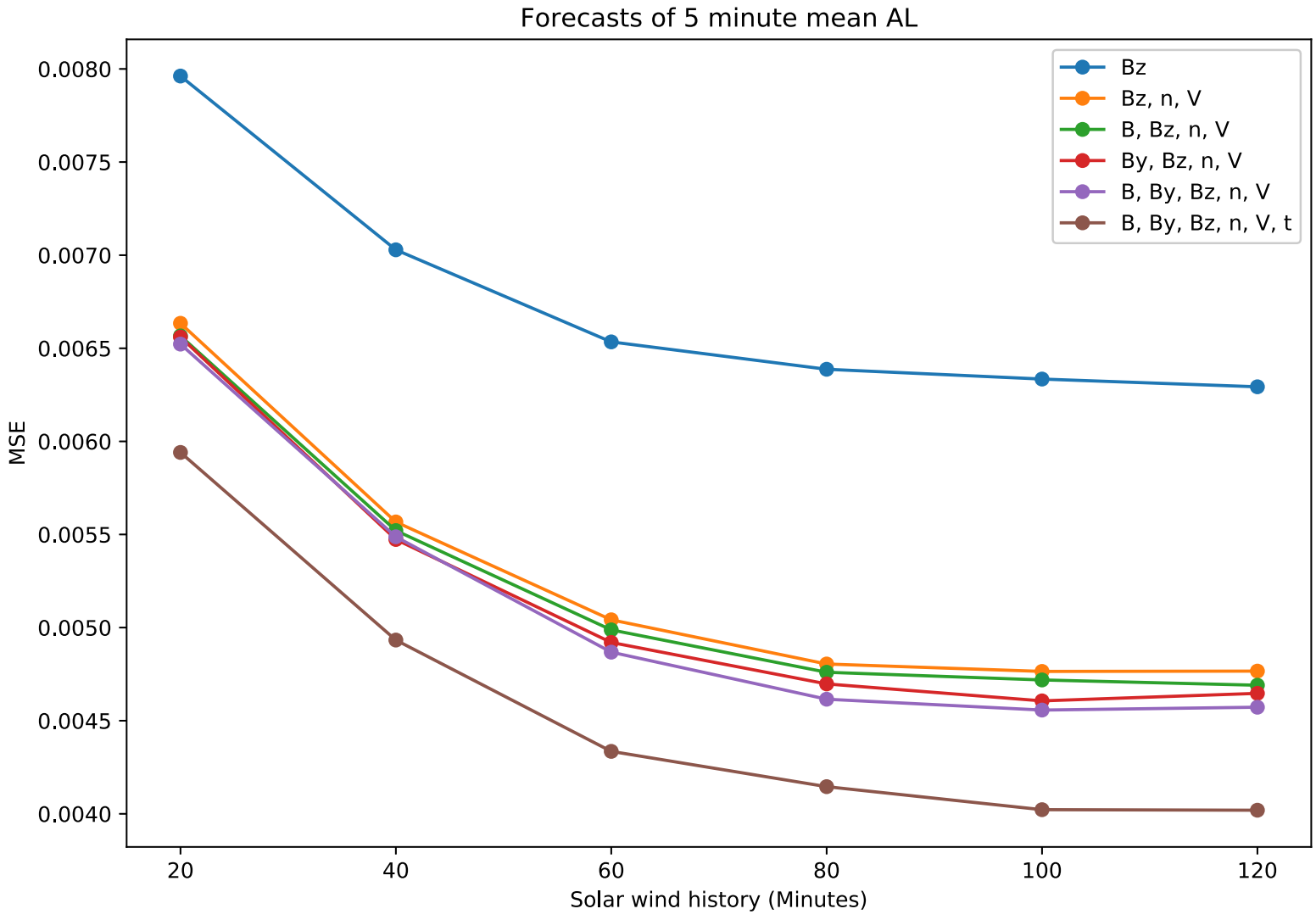
- The ACE data are first propagated to the location of the magnetopause. The data are resampled to 5 minute mean values, and we introduce time delays up to 120 minutes, to capture the dynamics (memory). For the *AE* indices we resample the original 1-minute data to 5 minute mean values.
- All data is rescaled, based on the training set, to be approximately 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.
- 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 optimiser.
- 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 model with lowest validation error was then selected.

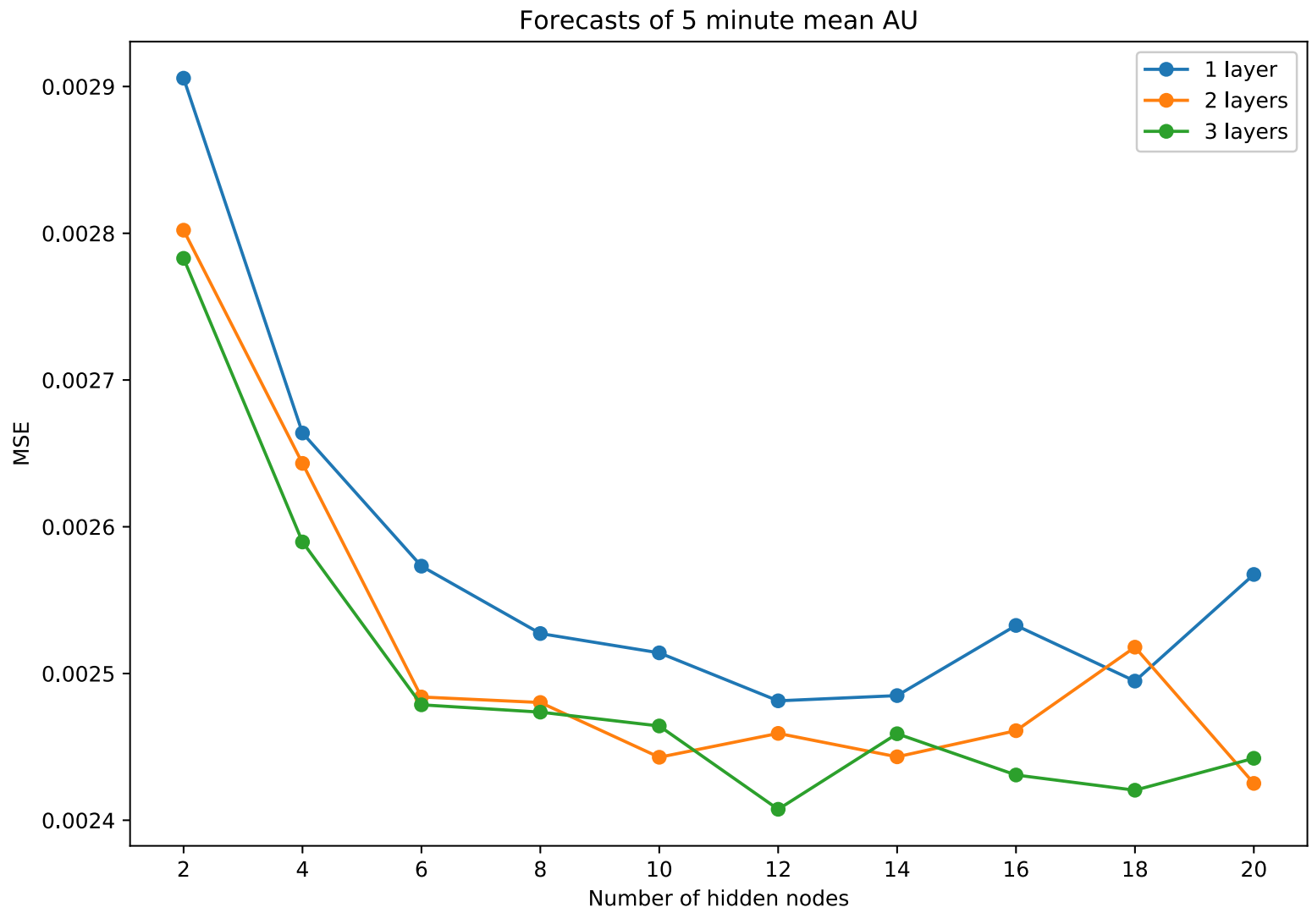
Model Studies



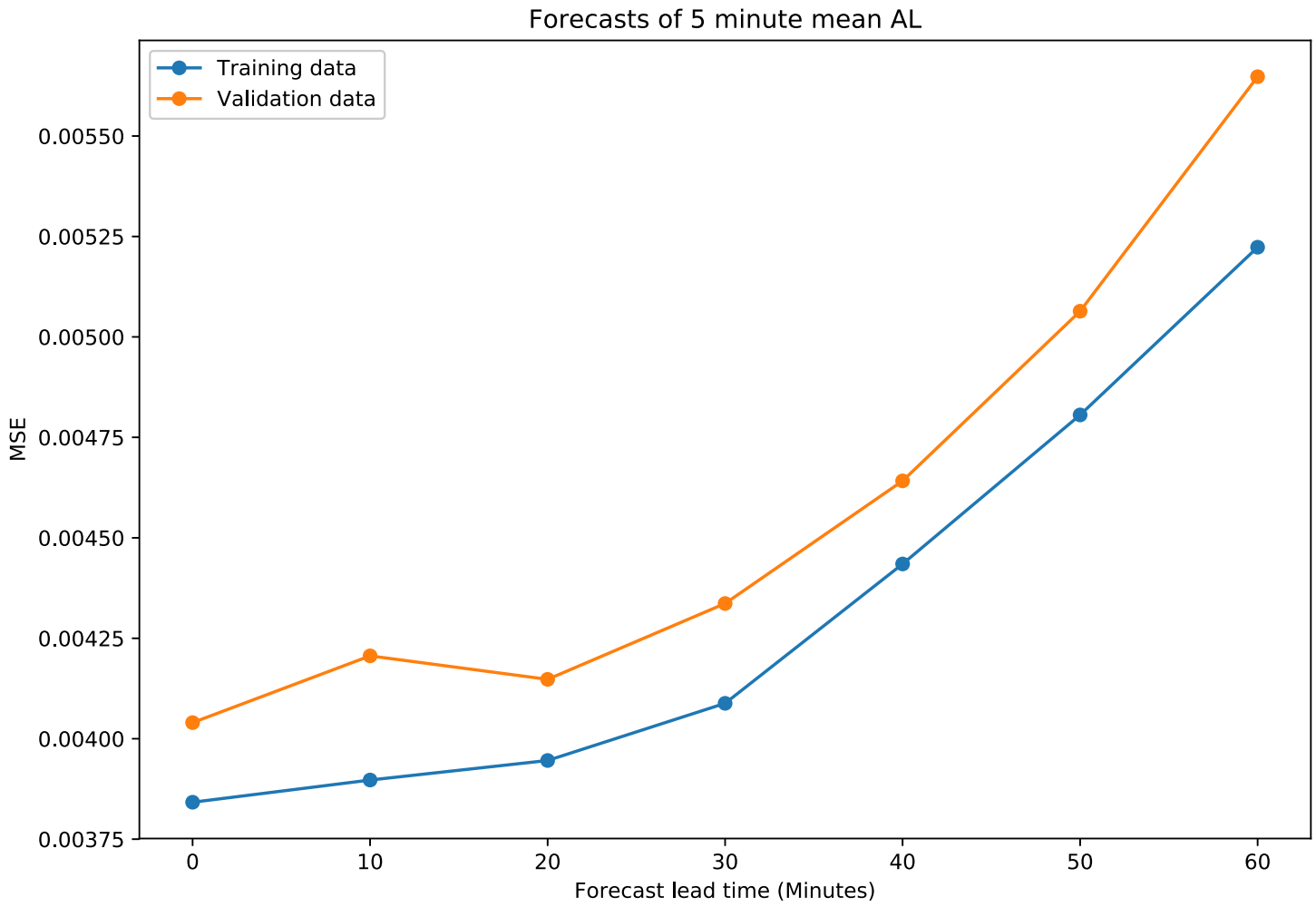


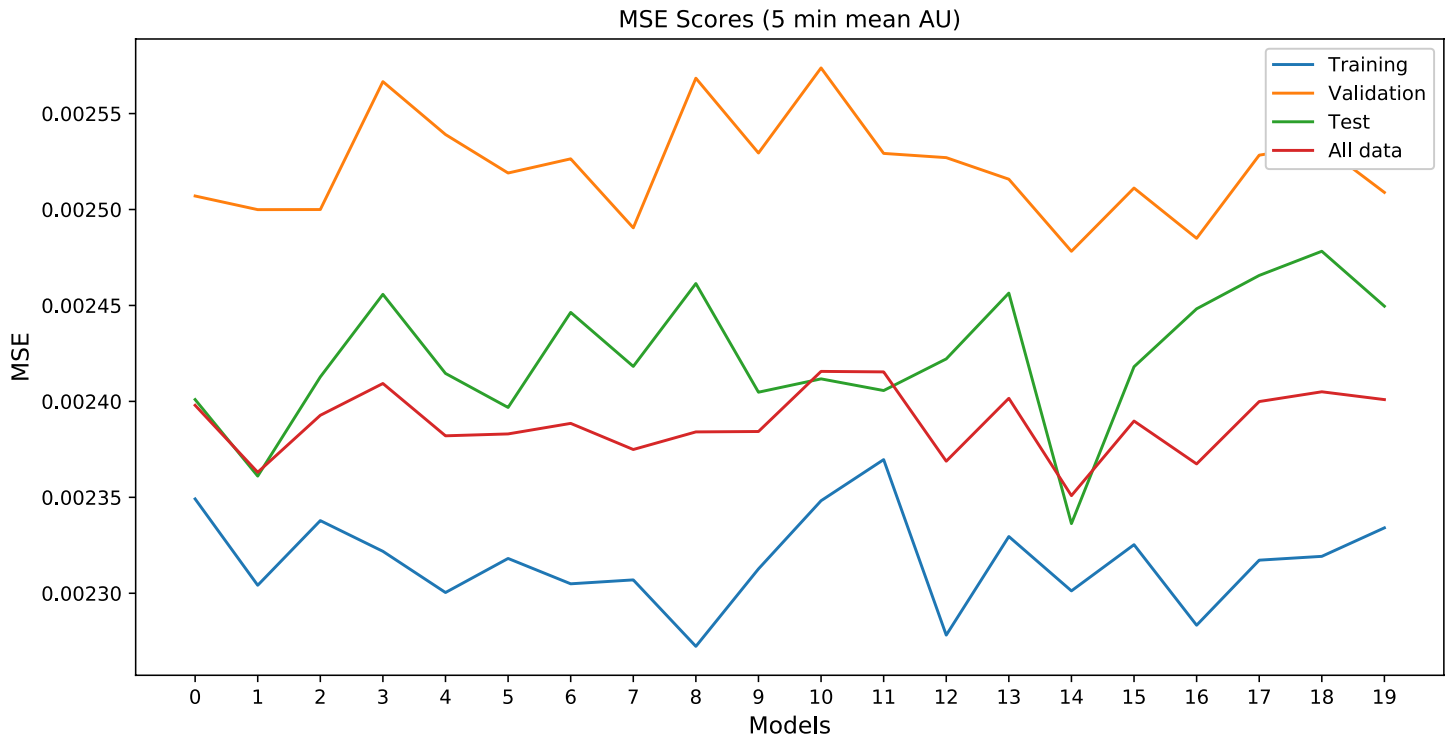
Model Studies



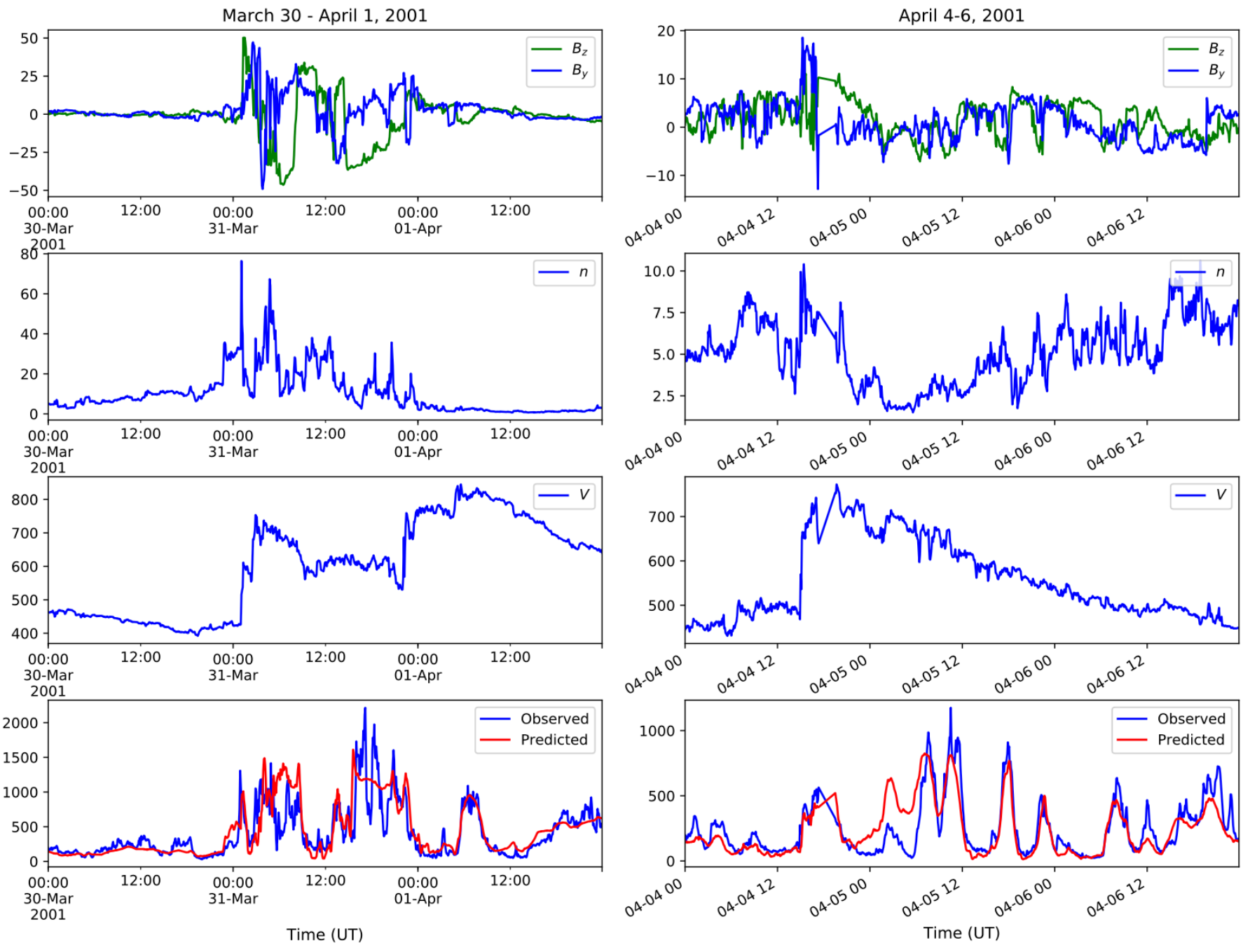


Model Studies





Forecast of 5 minute AE



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*.

Verification

- The lead time from L1, using the flat delay propagation, varies from 19 to 109 minutes, with a mean and median value of 60 minutes, which is used for the persistence model.

- The forecasts are verified using the measures: Bias (or mean error), mean absolute error (MAE), root mean square error(RMSE), linear correlation (Corr) and the mean square error skill score (MSESS) 1-$MSE_{model}/MSE_{persistence}$. The max and min values for the indices are also listed.

- The results are listed for training, validation and test set and all three sets combined. The results for the 60 minute persistence model are also included.

- We achieve the highest correlation, 0.88 for the *AE* index and MSESS of 0.6. This is clearly better than the persistence model.

- Although not shown in the tables, we also predicted the *AE* using predicted *AU* and *AL*. The results have a close agreement with direct predictions of *AE*. This is the reason we used the same data sets for all indices.

Verification

	Bias	MAE	RMSE	Corr	MSESS	Max	Min
Model (train)	2.287	56.904	90.394	0.844	0.606	35.2	-2894.8
Model (val)	1.687	57.345	93.480	0.830	0.575	32.2	-3330.0
Model (test)	0.830	52.483	87.660	0.837	0.566	41.4	-3747.4
Model (all)	1.824	56.035	90.544	0.839	0.591	41.4	-3747.4
Per (all)	0.006	82.650	141.493	0.639	0.000	41.4	-3747.4

AL index

	Bias	MAE	RMSE	Corr	MSESS	Max	Min
Model (train)	0.586	28.609	43.021	0.842	0.473	1389.4	-404.2
Model (val)	0.745	29.123	44.644	0.821	0.412	1106.2	-442.2
Model (test)	1.980	27.562	43.347	0.836	0.410	1011.0	-352.0
Model (all)	0.930	28.500	43.482	0.836	0.446	1389.4	-442.2
Per (all)	-0.003	35.952	58.396	0.729	0.000	1389.4	-442.2

AU index

	Bias	MAE	RMSE	Corr	MSESS	Max	Min
Model (train)	-2.094	69.786	107.216	0.883	0.617	2987.2	2.8
Model (val)	-1.813	69.874	108.579	0.871	0.589	3260.0	2.6
Model (test)	-0.443	64.594	103.355	0.883	0.579	3407.2	2.6
Model (all)	-1.664	68.664	106.705	0.880	0.603	3407.2	2.6
Per (all)	-0.009	103.484	169.339	0.717	0.000	3407.2	2.6

AE index

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 minutes, with serial correlation of up to 0.88 for *AE*.

- 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 dependence and the index itself as inputs.

- Later we plan to include F10.7 and use other algorithms such as SVR or LSTM and cross validation.