Development of a High-Latitude Convection Model by Application of Machine Learning to SuperDARN observations

W. A. Bristow [1], **C. Topliff** [2], **and M. B. Cohen** [2]

[1]Pennsylvania State University
[2]Georgia Institute of Technology

Abstract

A new model of high-latitude convection derived using machine learning (ML) is presented. The ML algorithm random forests regression was applied to a database of velocity observations from the Super Dual Auroral Radar Network (SuperDARN). The features used to train the model were the IMF components B_x, B_y, and B_z; the solar wind velocity, v_{sw}; the auroral indicies, A_u and A_l; and the geomagnetic index, *SYM-H*. The SuperDARN velocities were separated into north-south, and east-west components and sorted into a magnetic local time - magnetic latitude grid that ran from 55° to the magnetic pole with a bin size of 2° in latitude, and 1-hour in MLT. Separate models were created for each velocity component in each bin of the grid. It is found that even though the models in each bin are independent of one another a coherent convection pattern is formed when the models are viewed in aggregate. The resulting convection pattern responds to changes in the auroral indicies by expanding and contracting in a way that is consistent with expectations for a substorm cycle. Further it is found that the mean-squared difference between predictions of the model and observed values of the velocity are substantially lower than the same quantity calculated for an existing climatology that was not formed with ML techniques.

1 Introduction

Climatological convection modeling has been carried out for decades by binning various localized measurements of the ionospheric plasma velocity or electric field collected over the high-latitude regions versus some set of parameters, most often the interplanetary magnetic field (IMF) components. Such models often are used to drive circulation models such as the Thermospere Ionosphere Electrodynamic General Circulation Model (TIEGCM) (Roble & Ridley, 1994) and the Global Ionosphere-Thermosphere Model (GITM) (Ridley et al., 2006). They are also used to constrain the data driven convection patterns produced from SuperDARN data (Ruohoniemi & Baker, 1998). The measurements have been collected using a variety of instruments such as satellite based drift meters (e.g. Heelis et al., 1982) or electric field booms (e.g. Heppner & Maynard, 1987), incoherent-scatter radar (Foster, 1983), ground-based magnetometers (e.g. Papitashvili et al., 1994), and coherent-scatter radars (e.g. Ruohoniemi & Greenwald, 1996). Construction of the models typically has involved grouping observations based upon prevailing IMF conditions and perhaps some other parameter such as the planetary K-index (k_p) (Heppner & Maynard, 1987), or the geomagnetic Auroral Electrojet Index (A_e) (Weimer, 2005), or the dipole

tilt angle (Thomas & Shepherd, 2018), and then using the binned observations to constrain an expansion of the electrostatic potential in a set of orthogonal functions.

The underlying assumption of such a binning is that when repeated, a given set of driving conditions will on average produce the same unique convection pattern. In a general sense, physical reasoning and observations show this to be true. For example, when the IMF is southward, there is magnetic merging on the dayside magnetopause and the near-noon field lines connecting from the ionosphere out to the magnetopause are directly influenced by the electric field across the merging region. Those field lines are convected anti-sunward across the polar cap from noon to midnight where they eventually reconnect to the field lines from the opposite hemisphere. Once reconnected, the demand for magnetic flux on the dayside causes the field lines to return, following a path through the magnetosphere that maps to the ionosphere just equatorward of the polar cap boundary. The total potential drop across the polar cap should be equal to the projection of the solar wind electric field along the magnetopause reconnection region multiplied by the length of that region. Hence, for a set of solar wind/IMF parameters, the electric field would be the same and if the merging line is of the same length, the potential imposed in the ionosphere would be the same. This scenario leads to a two-celled pattern with flow from noon to midnight across the polar cap and return flow in the opposite direction at lower latitudes.

Numerous studies have examined the influence of the driving conditions on various aspects of convection. The 1987 article titled "Empirical High-Laittude Electric Field Models" by J. P. Heppner and N. C. Maynard, provides an excellent summary of the patterns that have been observed and how the IMF influences them. In particular, they highlighted the influence of the sign of the IMF y-component on the location and direction of the flow in the dayside throat. Their study contrasts with most others in that rather than binning the observations on a grid and then constraining a functional expansion of the potential with the binned observations, they examined individual satellite passes and categorized them as signatures or "quasi-signatures" and then sorted them based on the IMF. Their result was a set of three basic patterns that covered the majority of southward IMF situations. Those patterns illustrated sharper features (Harang Discontinuity, dayside throat) than are evident in most other models. In addition, they examined the influence of k_p and A_e, but only by comparing the average total cross polar cap potential drop for ranges of the parameters.

As illustrated by Hepner and Maynard's discussion of k_p and A_e, there several factors that influence convection that are not accounted for by the instantaneous state of the IMF and solar wind alone. For

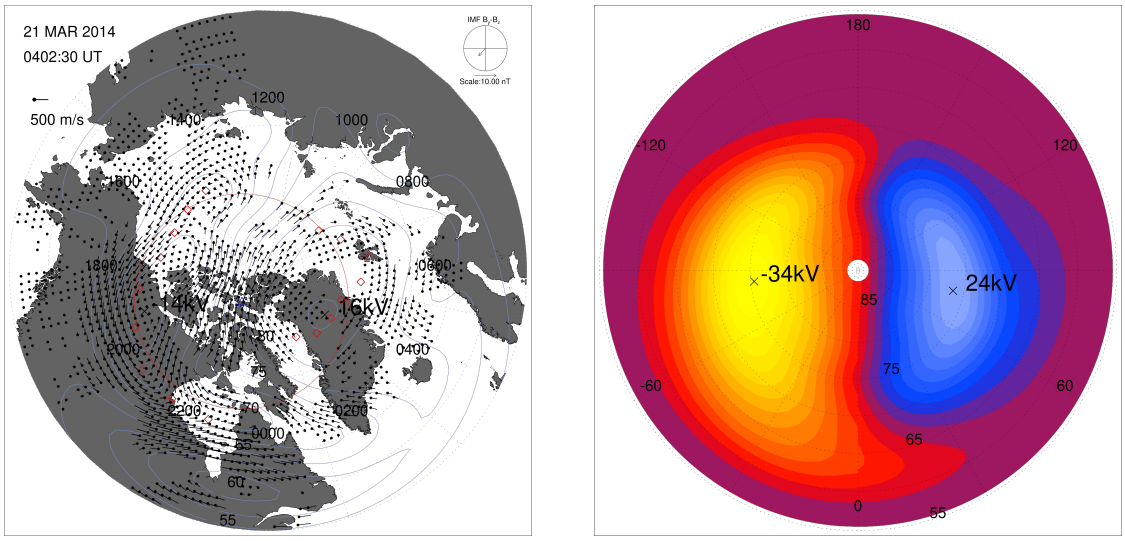
example, while dayside merging converts closed magnetospheric field lines into open polar-cap field lines, night-side merging in the magnetotail converts open field lines to closed, moving them from the polar cap to the magnetosphere. The diameter of the polar cap and hence the latitude of the convection reversal boundary is determined by the total open magnetic flux, which is determined by the balance between the dayside and nightside merging rates (Siscoe & Huang, 1985). That nightside merging rate is highly variable and depends on the internal state of the magnetosphere. In the growth phase of a substorm the nightside merging rate may be substantially lower than the dayside rate, leading to an expanding polar cap and expanding convection pattern. During a substorm expansion phase the opposite is true and rapid night-side merging can lead to a contracting polar cap and convection pattern. Further, features like the enhancement of the Harang Discontinuity during growth phase (Bristow & Jensen, 2007) change the shape of the pattern in addition to its diameter.

To account for some of the dependence on conditions beyond the solar wind and IMF, a new convection model was constructed using parameters that provide some indication of the state of the magnetosphere in addition to the solar wind and IMF parameters. Specifically, the auroral indicies A_u and A_l, and the mid-latitude geomagnetic index $Sym{-}H$. The auroral indicies give an indication of the strength of convection (A_u) and of the level of substorm activity (A_l), while the $Sym{-}H$ index gives an indication of the strength of the ring current, which has been shown to influence the diameter of the auroral oval (Schulz, 1997). These indicies are readily available for use from the NASA OMNI database (King & Papitashvili, 2005), which also provides the solar wind and IMF parameters aligned in time to reflect solar wind propagation delays from the point of observation to the Earth's bow shock. Including the magnetospheric parameters increases the dimension of the parameter space to seven, which is fairly large for traditional method of binning the observations. In addition, as will be demonstrated the dependence of the convection velocities on some of the parameters is nonlinear. Because of these two factors, machine learning (ML) was used to form the model.

The paper begins with discussion of the SuperDARN data and how the database influenced the choice of algorithms for generating the model. The form of the data motivated producing independent models for the velocity at the points of a latitiude-MLT grid rather than an orthogonal function expansion of the global-scale potential pattern. That discussion is followed by examination the output of the model at a single location and comparison to an existing climatology. Next, the individual models are combined to generate global-scale potential patterns that could be used in the same way as existing clima-

tologies. Finally, it's demonstrated that the resulting model predicts observations with a lower error than the climatology.

2 Machine learning implementation



(a) (b)

Figure 1: (a) Example convection map from March 21, 2014 created using observations from 0400 UT to 0405 UT, and (b) potential map generated by the Thomas and Shepherd 2018 climatological model for the IMF conditions at the time.

The convection model was based on observations from the Super Dual Auroral Radar Network (SuperDARN) (Greenwald et al., 1995) processed with the potential mapping technique, Map-Potential (Ruohoniemi & Baker, 1998). Potential maps were generated for every 5-minute interval from January 1, 2013 to December 31, 2017. The Thomas-Shepherd 2018 (TS18) statistical model (Thomas & Shepherd, 2018) was used as the background for Map-Potential, with the IMF from the OMNI database used to select the patterns. An example pattern from 0400 UT to 0405 UT on March 21, 2014, is shown in Figure 1. The map shows contours of the electrostatic potential along with flow vectors at locations where radar observations constrained the fit. At the time of the plot, the IMF was weekly southward with a negative y-component. Figure 1b shows the potential contours predicted by the TS18 model for the con-

ditions observed at the time from the observed IMF conditions. There are several subtle differences between observed pattern and that predicted by the model. First, the total cross-cap potential drop is significantly lower in the observations. The drop is only 30 kV, while the predicted value was 58 kV. The observed pattern is shifted toward midnight and rotated slightly toward dusk. That is, the flow across the polar cap in the observation is from pre-noon to pre-midnight, while the model predicted flow directly from noon to midnight. Such differences are typical for any given interval and simply illustrates the inherent variability of the convection pattern and the need for using observations whenever possible rather than relying on climatologies.

Several methods were considered for generating the ML model. The simplest implementation conceptually would have been to bin the line-of-sight observations from the individual radars similarly to the way they have been used to construct previous models. This works well for constraining global models of the potential since the plasma velocity in the F-region ionosphere is very nearly the so called $\mathbf{E}\times\mathbf{B}$ velocity and the observation can be written as the negative gradient of the potential ($\mathbf{E} = -\nabla V$) crossed with the local magnetic field and projected along the line-of-sight. Having an ensemble of observations at different locations is sufficient to constrain the fit for the entire high-latitude region. A second alternative considered was to form potential patterns and treat them as images. Convolutional neural networks are adept at using such observations, however because the pattern obtained at any given time can be dominated by the influence of the statistical model, any ML model trained on such patterns would tend simply to reproduce the statistical model.

Another method considered was to use vectors formed at locations where two or more radars provided measurements. These merged vectors are not influenced by a model, however they would have provided rather sparse coverage and would have required discarding data from locations where only one radar had an observation. In addition, vectors formed in this way can be noisy because the LOS observations from the two radars can potentially be separated in time by tens of seconds which often means they are observing significantly different conditions. Small changes in the azimuth of a flow can result in significant changes in the LOS projections, which are amplified when they are recombined to form a vector, especially when the viewing angle between the lines of sight is small.

The method we chose was to use the vectors from the SuperDARN potential mapping obtained at locations where there were there were one or more observations contributing to the fit. While the vectors are influenced by the statistical model, having the observation at a location significantly lessens

that influence. In addition the since the fitted vectors are consistent with an electrostatic potential, which is a strong physical constraint, using them minimizes the impact of noise and radar-to-radar inconsistencies in the observations. The method had the added benefit of decreasing the size of the database from what would have been required if we had used the observations from the individual radars. One disadvantage of using the data in this way is that values at any given location are not continuous in time, which limited the ML algorithms that could be applied. Individual radars do not have continuous observations and the longitudinal distribution of radars ls not uniform so some longitude ranges are not covered. The Russian sector is the most obvious illustration of the gap in coverage.

The fitted vectors were binned on a magnetic latitude-local-time (MLT) grid from 55° magnetic latitude to the pole with a cell size of 2° in latitude and 1-hour in MLT. The vectors were resolved into north-south and east-west components and written to comma-separated-value files along with the associated value of the IMF vector (B_x, B_y, and B_z), the solar wind velocity (v_{sw}), the Auroral Electrojet Indicies (A_u, A_l), and the Sym-H index, all from the OMNI database.

Figure 2 shows the components of the database for a representative 3-day period for the bin at 65° magnetic latitude and 2000 MLT. As discussed above the data are not continuous in time, which limits the type of ML techniques that can be employed. While it might be desirable to use something like the Long-Short-Term Memory technique (LSTM) since it is a good technique for predicting time evolution based on time-series drivers, continuous observations are necessary to train such a model.

While not continuous, the database is relatively large. Figure 3a shows the number of data points in each grid cell. The highest values illustrated in the figure are in excess of 1-million, however at the lowest latitudes some cells have less than 100,000 observations. These locations are often equatorward of the convection zone so there are no usable observations available. The low-latitude extent of significant convection is highly variable in time, like most aspects of convection, and is determined by the magnetospheric drivers and state. In general convection is confined to latitudes above a low-latitude convection boundary referred to as the Heppner-Maynard Boundary (HMB) (Shepherd & Ruohoniemi, 2000). Fortunately, by examining the observations from the entire SuperDARN network at a given time it is possible to identify this boundary with some confidence. With that determination, it is possible to assign a zero value at those times to the velocity in cells that lie at latitudes below the HMB. Figure 3b shows the density of points including the assignment of zero velocity when a bin is at a lower latitude than the HMB. With this assignment, there are in excess of 400,000 points in all bins between

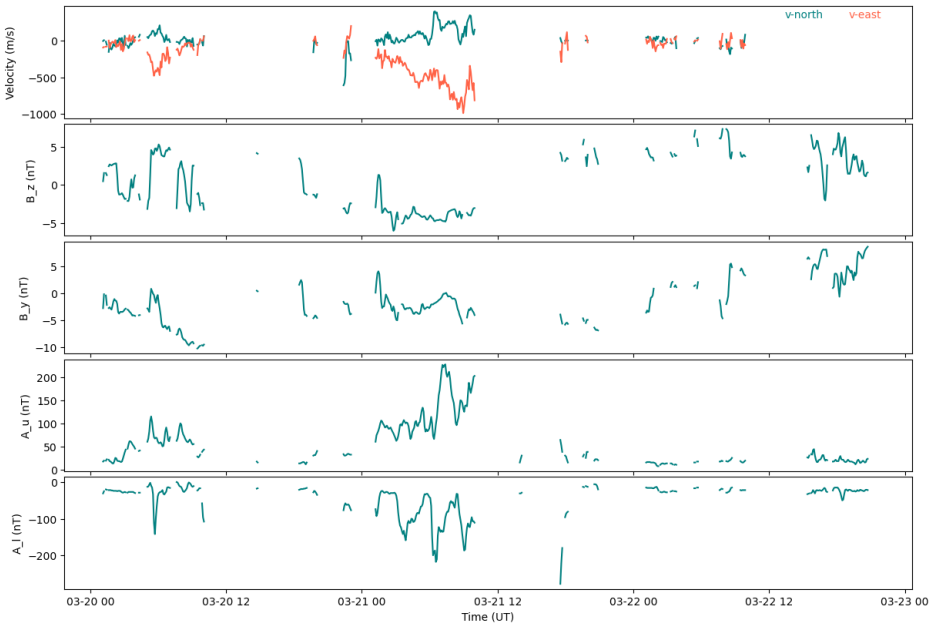
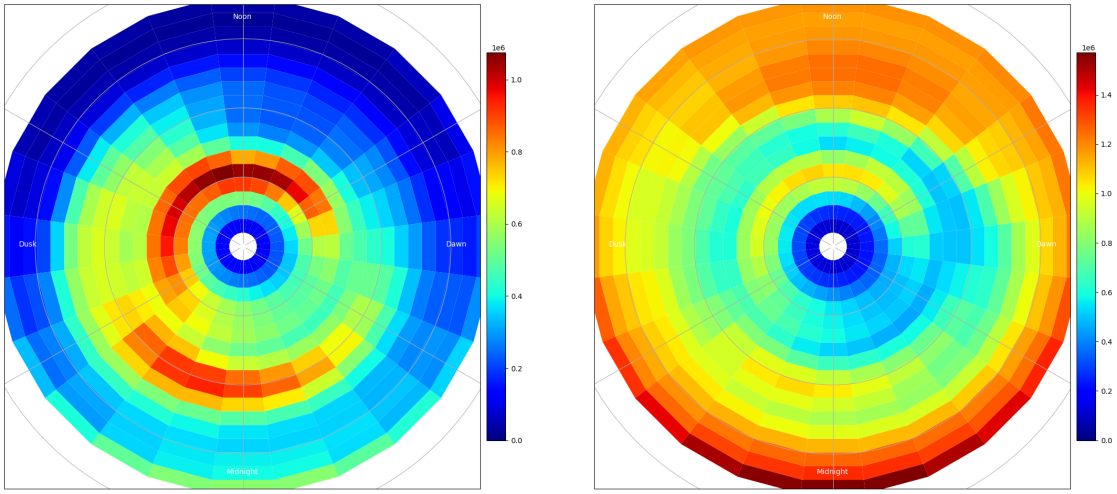


Figure 2: Time series of observations and feature values in the grid cell at 65° magnetic latitude and 2000 MLT for the interval March 20 - 22, 2014.

55° and 80°. Above about 85° there are not many observations, so uncertainties will be larger there than in other regions.

Figure 4 illustrates the relationships between the velocity components in the grid cell at 67° magnetic latitude and 1800 MLT, and some of the parameters from the database. In each frame of the figure, the vertical axis is one of the velocity components (v_{ns} or v_{ew}) and the horizontal axis is one of the database parameters. Pixel color indicates the density of points in a bin. Solid black dots indicate the average velocity in each of the parameter bins. No color scale is provided since the goal is to examine trends and not to extract quantitative information. The purpose of examining the data in this way is to select parameters for inclusion as features for training the ML model. If the velocities were uncorrelated with any of the parameters it would be possible to exclude them and decrease the complexity of the model. With that in mind, it is still interesting to examine the trends that the plots show. As would be expected for the auroral zone latitude dusk MLT location, the magnitude of the north-south component (v_{ns}) is significantly smaller than the east-west component (v_{ew}). v_{ew} shows a strong dependence on each of the selected parameters, though the dependence is clearly nonlinear for A_u and A_l. Frame 4a illustrates that v_{ew} is negative (westward) for the vast majority of the data, indicating



(a) Number of observations

(b) Observations + assigned zeros

Figure 3: Density of points in the database. Color corresponds to a) number of observations in cell, or b) number of observations plus the number of assigned zero values.

that the location remains equatorward of the convection-reversal boundary under most conditions. The plot shows significant scatter of velocity values for all IMF values, though the average shows a roughly linear trend of increasing westward velocity with increasing negative B_z magnitude. The average values of the velocity are offset from the highest density of points indicated by the color contours, which shows that the distributions are non-Gaussian. Frame 4b shows that v_{ns} also has significant scatter, however it remains small for all values of B_z. It demonstrates a nearly linear trend of increase with increasing negative B_z magnitude, however the trend is small and the spread of velocities is significantly larger than the trend. 4c shows v_{ew} vs the auroral index A_l which is an indicator of substorm activity. Again, there is significant scatter in the velocity values for all values of A_l. There is also a clear nonlinear dependence of the velocity on the index. For small values of A_l, the velocity magnitude increases rapidly with increasingly negative A_l, while at higher index values the velocity increase is small. Similar behavior is illustrated for the dependence of v_{ew} on the A_u index (4d).

Figure 5 is the same format as Figure 4 but for the bin at 81° latitude and 1300 MLT, which lies in the post-noon polar cap under most conditions. The dependencies differ significantly from the auroral zone dusk cell. v_{ns} at this location shows a clear nearly linear dependence on B_z, with positive (antisunward) values for negative B_z and negative (sunward) values for positive average B_z in excess of about 2.5 nT.

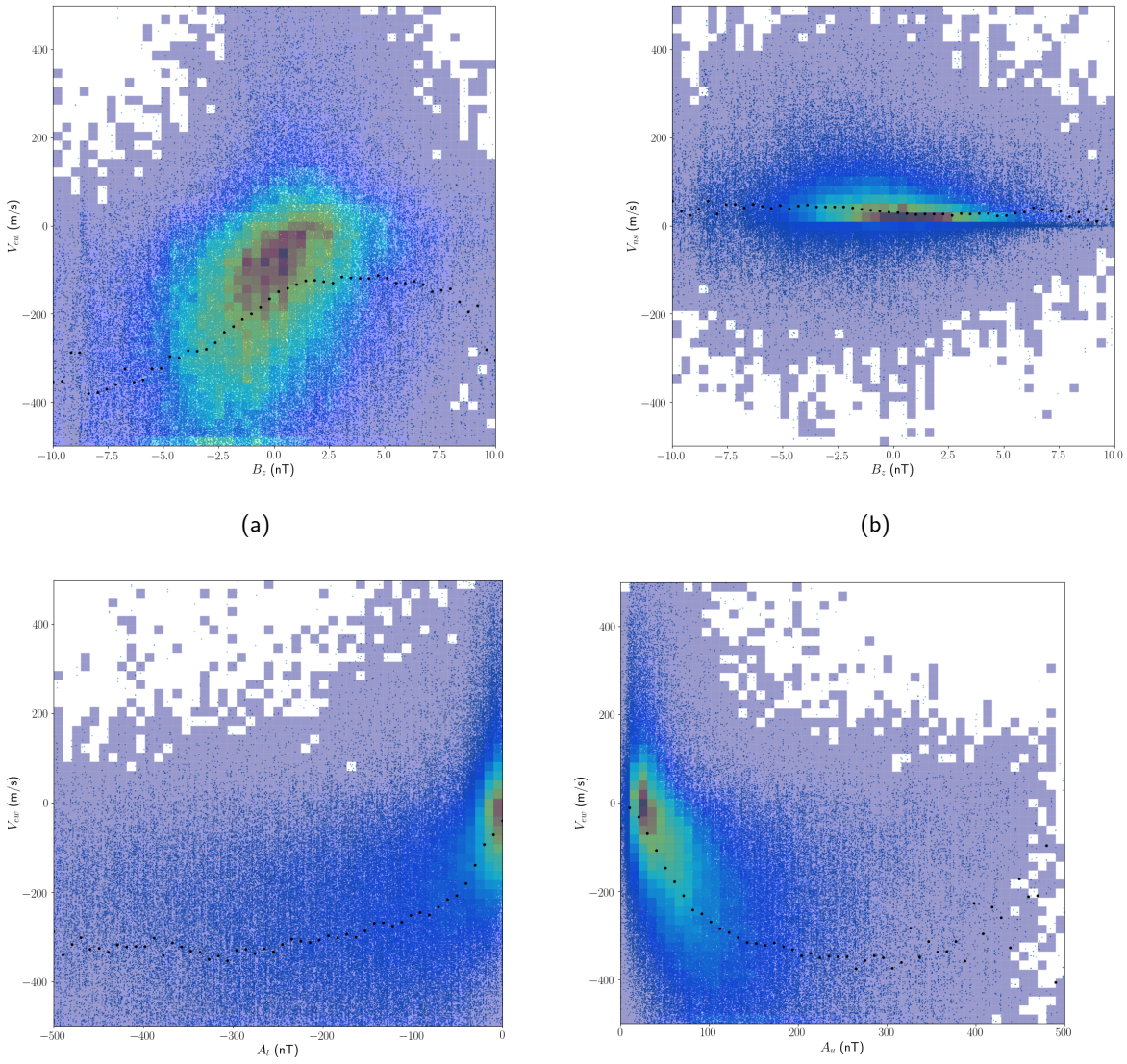


Figure 4: Dependence of velocity components in the bin at 67° magnetic latitude and 1800 MLT versus select parameters from the database. (a) shows the relationship between v_{ew} and the IMF z-component, (b) shows v_{ns} vs IMF z-component, (c) shows v_{ew} vs A_l, and (d) shows v_{ew} vs A_u.

The east-west velocity component is small magnitude and appears weakly correlated with the parameters (B_z, A_u, and A_l).

As the two figures show, the relationship between the velocity components and the database features is complex and varies from place to place. In some regions the velocity may be much more strongly correlated with one parameter than with another, while in another location the opposite is true. Because of this, none of the parameters was eliminated from consideration. All seven parameters were used to train the models.

The database was used to train an independent model of each velocity component (v_{ns} and v_{ew}) in each latitude-MLT grid cell. With the 17 latitude bins and 24 longitude bins, there are 408 grid cells. Fitting the model components separately means that there are a total of 816 independent models. Three algorithms were tested for forming the model. The algorithms were the LinearRegression, DecisionTreeRegressor, RandomForestRegressor provided by the Scikit-Learn software package (Pedregosa et al., 2011). To test each algorithm, the data base was processed with each model in a ten-cell subset of the grid space. To limit over-fitting, the maximum-depth hyperparameter for the Random Forest and Decision Tree model was set to 15. The resulting models were used to predict velocities in a sample of data outside of the training set and the model with the lowest root-mean-squared error (rmse) was selected. In each case, the Random Forest Regressor was substantially better than the others. For example in the bin at 67° latitude 1800 MLT, linear prediction of v_{ew} resulted in a rmse of 169.3 m/s, the decision tree resulted in a rmse of 126.5 m/s, and the random forests resulted in a rmse of 113.2 m/s. In ScikitLearn, the Random Forest model is trained by fitting multiple decision trees to random subsamples of the inuput data and aggregating the predictions of all the trees. This is one way to address the overfitting in addition to controlling the maximum depth of each tree.

After model selection on the subset of grid cells, the full dataset was split into data from the years 2014 to 2017, which was used to train the model, with data from 2013 used as a test set. Figure 6 shows a 2000 sample interval from the model in the bin latitude 67°, MLT 1800. The horizontal axis is sample number from the database which corresponds to time, however because of the data are not continuous multiple time intervals contribute to the plot resulting in discontinuities in the plot traces that do not represent temporal discontinuities of the values. The upper frame of the plot shows the observed values in red, the model predictions in green, and the predictions from the TS18 model in blue. The lower two frames show the IMF y and z components, and the A_u and A_l indicies. For most of the in-

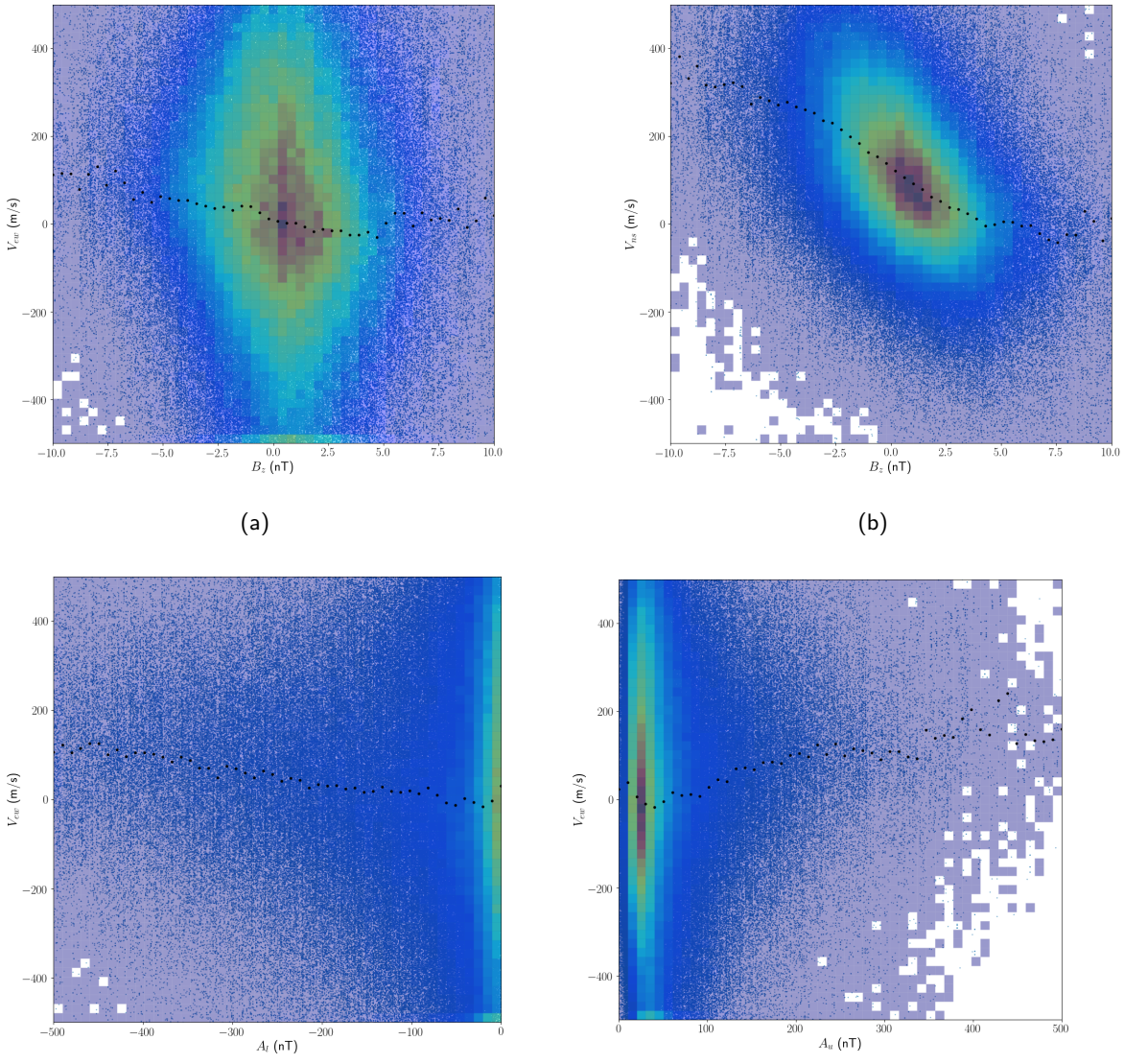


Figure 5: Same as Figure 4 except for the bin at 81° magnetic latitude and 1300 MLT.

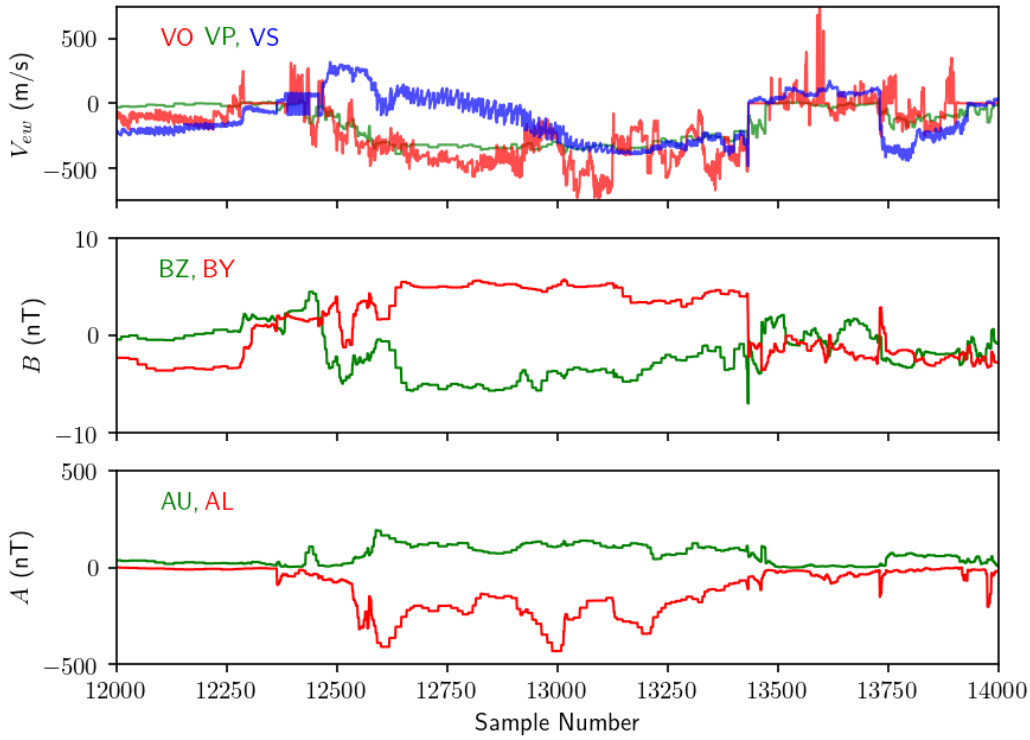


Figure 6: Sampling of the driving features and model predictions for the bin at 67° magnetic latitude and 1800 MLT. The 2000 sample interval is composed of multiple time intervals. In the top panel, the red trace is the observed velocity, the green trace is the velocity predicted by the ML model, and the blue trace is the velocity predicted by the TS18 model.

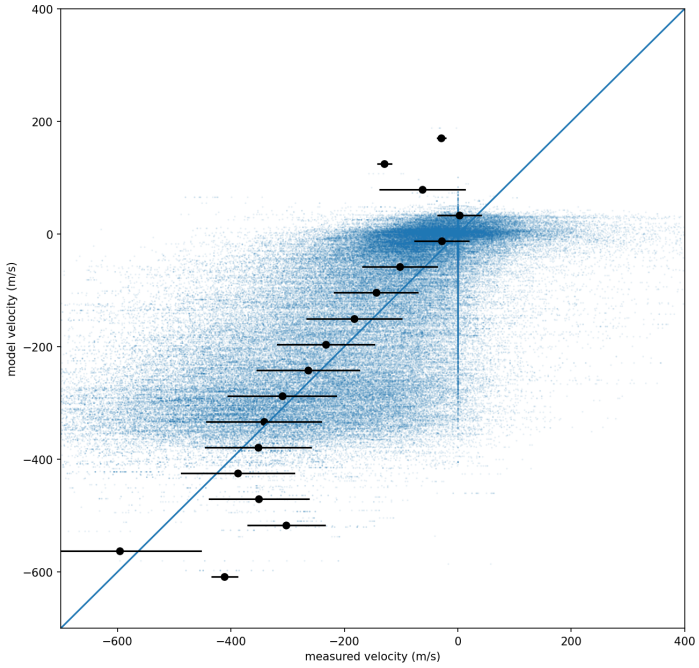
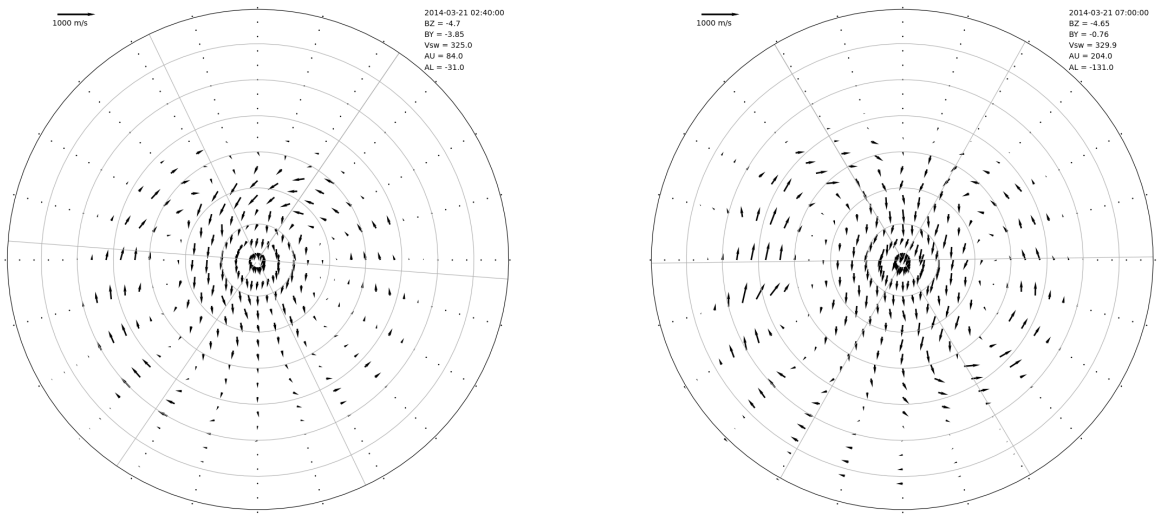


Figure 7: Scatter of model predictions versus observed values of v_{ew} in the bin at 67° latitude 1800 MLT.

terval, the predicted value ls close to the observed value, with differences of less than 100 m/s. The predictions from the TS18 model at times show much larger differences from the observations as is especially well illustrated in the values from samples between 12400 and 13000 where the difference is on the order of 500 m/s.

Figure 7 shows the difference between the predictions and observations for the full year of 2013. The figure shows the scatter of predicted velocity (vertical axis) versus measured velocity (horizontal axis) in the bin at 67° and 1800 MLT. The solid black circles show the average values and the horizontal lines show the average plus and minus one standard deviation. The average values follow the equality line for most of the range, with significant deviation only for values where there are relatively few points. Where the average values lie above the equality line, there is a small bias (< 50 m/s) for the model values to be smaller magnitude than the observed values. While the scatter appears large, the standard deviations demonstrate that the majority of the predictions are within 100 m/s of the observations for all values with a significant number of observations.

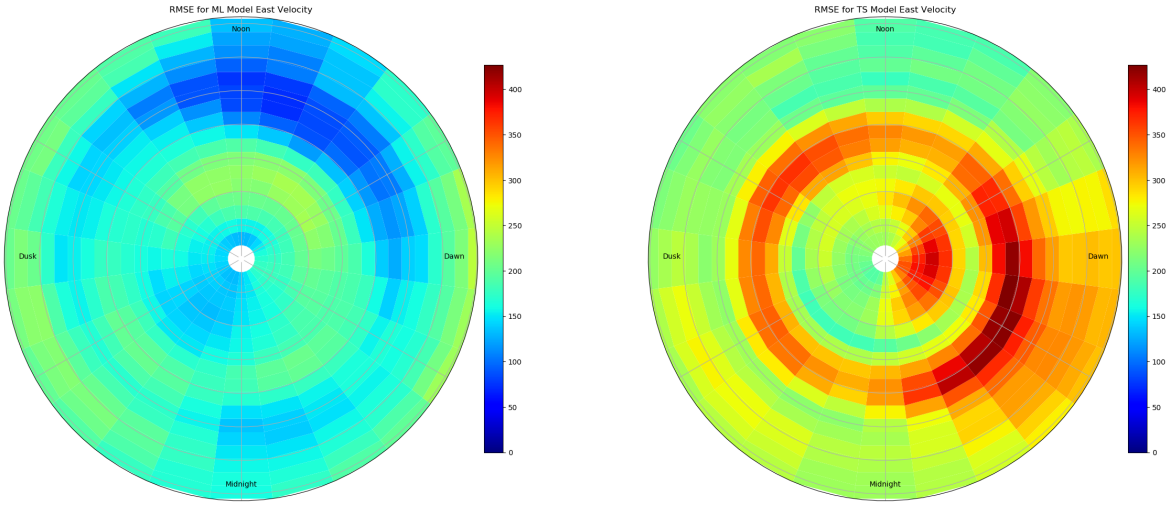
Figure 8 shows the result of running the model for two intervals with similar IMF values but significantly different values for A_u and A_l. The north-south at east-west models were run for each grid cell and then combined to form vectors. Each grid cell is independent, so there was no guarantee that the output would produce a coherent convection pattern. The results do in fact illustrate a well defined coherent convection pattern that is consistent with expectations based upon the observed driving conditions. During the two intervals the IMF was southward with $B_z =$ 4.7 nT and $B_y =$ -3.85 nT in the first interval (8a) and $B_y =$ -0.76 nT in the second interval (8b). In frame 8a, the auroral indicies are small with $A_u =$ 84 nT and $A_l =$ -31 nT. In frame 8b A_u was roughly two and a half times and A_l was roughly four times the value in 8a . While the IMF values are similar, the patterns show significant differences. The most obvious of which are that the pattern driven by the larger values of A_u and A_l extends to lower latitudes and has larger magnitude at nearly all locations. In 8a the convection is confined to latitudes above about 65°, while in 8b it extends to below 60° in the pre-midnight sector. Dayside plasma flows extend to slightly lower latitude in the later interval, though not by as much as the night-side flows. The main difference in the dayside is that the direction and local time of plasma entry to the polar cap reflects the influence of the larger IMF B_y in the earlier interval. The nightside exit of plasma differs significantly between the two plots. In 8a flow near midnight is small magnitude and mainly equatorward before turning to connect to the return flow regions. In 8b the dawn cell is roughly "D" shaped with the flow turning directly from cross-cap to the return flow region, while in the dusk cell, there is the flow rotates first dawnward before rotating back to connect with the dusk return flow. This dusk-cell shear flow illustrates the development of the Harang Discontinuity with increasing auroral activity.



(a) (b)

Figure 8: Output of the 816 independent models displayed on the latitude-MLT grid for similar IMF conditions but differing auroral indicies a) $B_z =$ -4.53 nT, $B_y =$ -2.45 nT, $A_u = 82$ nT, $A_l =$ -34 nT b) $B_z =$ -4.6 nT, $B_y =$ -1.41 nT, $A_u = 157$ nT, $A_l =$ -128 nT

To examine the accuracy of the model prediction over the entire grid, the root-mean-squared difference between the predictions and observations were calculated in each grid cell for all observations the year of 2013. The models were used to predict the velocity components at each time for which there as an observation in a given grid cell based upon the values in the OMNI database from the time of the observation. For comparison, the TS18 model was used similarly to predict the velocity components and compared to the observations. Figure 9 shows the results for v_{ew} over the grid for both the ML model and the TS18 model.



(a) (b)

Figure 9: Root-mean-squared difference between model predictions and the observations of the east-west velocity component for the year of 2013. a) RMSE from the ML model, b) TS18 model

The figure shows that the RMSE for the ML model is less than about 250 m/s over the entire grid. The lowest values occur on the low-latitude dayside, where velocities are typically low. The highest values occur in the prenoon sector between 75° and 80°, and at the lowest latitude bins near dawn and dusk. In addition errors on the night-side are highest in the region between 70° and 75°. The plot for the TS18 model shows higher RMSE for all bins, with particularly large errors (>350 m/s) near 70° for all local times.

3 Discussion

The need for accurate forecasting of space weather increases on a nearly daily basis. There isn't a better example of this than the requirement for accurate orbit prediction that becomes more critical with the launch of every new low-Earth-orbit satellite. Orbit prediction is based on thermospheric density, which can be predicted using global circulation models driven by convection models such as described in this study. Hence, it is imperative that we have models of convection that accurately capture the variation of the high-latitude potential with IMF and internal magnetospheric state. The ML model pre-

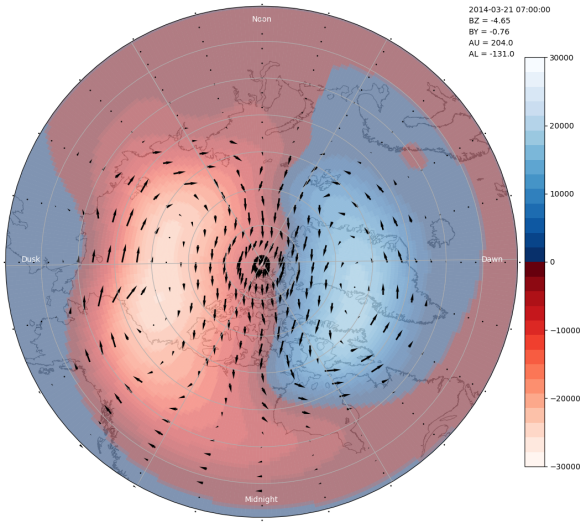


Figure 10: Potential pattern resulting from a spherical harmonic expansion constrained by the ML model

sented here shows a marked improvement over traditional climatological models, however it requires specification of auroral indicies, A_l and A_u, which are based upon magnetometer observations. Recently there have been successful efforts at predicting these and other magnetospheric indicies using machine learning techniques using the solar wind and IMF as inputs. With these predicted indicies it would be possible to predict the convection, and in turn the thermospheric density, several hours into the future (Topliff et al., 2019).

For retrospective studies, observations of the indicies are readily available and can be used to select model patterns for use in GCMs or to serve as a constraint on instantaneous convection patterns generated using MapPotential. GCMs and MapPotential use models of the electrostatic potential rather than velocities. Such maps can be generated from the output of the ML model by using the same technique that has been applied in generating other climatological models. Figure 10 shows the result of expanding the potential in spherical harmonics using the ML model velocity field in Figure 8b as a constraint on the fit. Because the fit is a functional expansion it can be calculated on a fine grid, which gives the smooth variation with position illustrated in the figure. The pattern is similar to that shown in Figure 1b, which was generated from TS18 using similar values of the IMF. The dusk cell is not quite as round in the ML model plot and the flow near midnight extends to lower latitudes.

Figure 8, shows that changes in A_l and A_u result in changes in the convection predicted by the ML model which reflect the expected behavior of the polar cap. Convection extended to lower latitudes in Figure 8b than in Figure 8a in response to the significantly larger values of A_u and A_l. The latitude of the convection reversal is impacted by the changes in the indicies, though the change is not uniform in local time. At dawn the boundary was lower by several degrees in 8b than in 8a, and the development of the Harang Discontinuity appears in 8b, which gives a convection reversal at low latitudes in the premidinght region and extending across 0000 MLT. The extension of the convection reversal across midnight results in the tongue of negative potential extending across midnight illustrated in Figure 10. The dayside convection reversal and that at dusk shows little difference between the two intervals. The dayside differences between the two intervals that do appear are more likely due to the differing values of the IMF B_y.

As illustrated by the comparisons to the TS18 climatology, the ML model represents an improvement over models that do not attempt to capture variability driven by internal magnetospheric processes. Figure 9 showed that the RMSE of the ML model predictions vs SuperDARN observations is substantially lower than that for the TS18 model. The region of large RMSE in the TS18 model concentrated between 70° and 75° is most likely due to the inability of the model to capture the expansion and contraction of the polar cap boundary during substorm cycles. The region of large RMSE is close to the average position of the polar cap boundary, which is close to the latitude of the convection reversal boundary (CRB). The CRB is of course, the latitude separating the antisunward flow in the polar cap and the sunward flow on field lines that map into the magnetosphere. When the polar cap expands so that the CRB latitude is below it's average position for a given set of IMF/solar wind conditions, grid cells just below the boundary would be predicted to lie in magnetosphere and have sunward flow, while in fact they lie in the polar cap and have antisunward flow. Likewise, when the boundary contracts above it's average position, grid cells just poleward of the average boundary would be predicted to have antisunward flow while in fact it is sunward. When either of these conditions happens, the difference between the prediction and the observation would be on the order of double the average magnitude of the velocity in the cell. Since the boundary is so dynamic it is likely that this is a common occurrence, which is reflected by the large average errors that appear in the figure. It should be noted that the largest errors in the ML model also appear in this region, however they are significantly lower magnitude than in the TS18 model indicating that the ML model does a better job of representing the changes in the boundary position.

Another advantage of the ML model over existing climatologies is that the way it was formed allows for easy characterization of the distribution of velocities in each grid cell, which can be used when assimilating the model output. The RMSE is returned for each grid cell as part of the ML regression. If the model is assumed to be unbiased, the RMSE value can be used as the square root of the variance and combined with the model output value to generate a distribution function assuming a Gaussian distribution. A Bayesian assimilation scheme would use the distribution of model as prior information. Having an RMSE in each grid cell contrasts with the information obtained when expanding the potential in orthogonal functions. In such a fit, the function coefficients are returned, which distributes errors over the domain. While it would be possible to generate a covariance matrix for a functional expansion, it requires the extra step of using the model to predict a local velocity and calculating the sample variance around that value.

4 Conclusions

This paper describes a climatological model of high-latitude convection derived using machine learning (ML) techniques applied to observations from the SuperDARN radar network. The model was generated from a database of four years of observations and tested over a separate fifth year. SuperDARN convection patterns were generated for every five minute period over the five year period. From those patterns, velocity vectors were calculated at locations where there was at least one radar contributing observations. Those velocities were separated into north-south, and east-west components and sorted into a magnetic local time - magnetic latitude grid that ran from 55° to the magnetic pole with a bin size of 2°, and MLT bins of 1-hour.

In each MLT-MLAT bin, the two velocity components were used separately to train a ML model using random forests regression. Random forests was selected after testing three different ML algorithms to find the one that produced the lowest RMSE in a subset of the points in the grid. The features used to train the model were the IMF components B_x, B_y, and B_z; the solar wind velocity, v_{sw}; the auroral indicies, A_u and A_l; and the geomagnetic index, *SYM-H*.

After the model was trained on data from the years 2014 to 2018 (inclusive), it was tested using data from the year 2013. Predictions from the model were compared to the SuperDARN observations and distributions of predicted versus observed velocity were examined. While there was significant scatter

of the predictions around the line of equality with the observations, the average of the distributions tracked the average measured velocity well with a small bias to lower values. The standard deviation of the model predictions was less than 100 m/s for all bins where there were a significant number of observations.

RMSE values for the model were compared to those from the TS18 model in each bin of the grid. The ML model exhibited smaller errors than TS18 at all locations. In particular, errors in the ML showed the largest improvement over TS18 in bins that are near the average latitude of the convection reversal boundary. It is likely that the improvement was due to the ML model's ability to expand and contract in latitude in response to changes of A_l and A_u.

The software for generating the model is free and available for download from the scikit-learn web site. The web site has links to numerous examples and tutorials for application of the various algorithms it provides. The software is simple to use even by senior investigators with no prior experience with ML techniques. Despite the simplicity, good results can be obtained with some time spent reading the tutorials.

5 Data Availability

The raw SuperDARN data are available from the British Antarctic Survey (BAS) SuperDARN data server (https://www.bas.ac.uk/project/superdarn).

Acknowledgments: This work is supported by the Defense Advanced Research Projects Agency (DARPA) through US Department of the Interior award D19AC00009 to the Georgia Institute of Technology and subaward to The Pennsylvania State University. SuperDARN operations and research at Pennsylvania State University are supported under NSF Grants PLR-1443504 from the Office of Polar Programs, and AGS-1934419 from the Geospace Section of NSF Division of Atmospheric and Geospace Sciences. The authors acknowledge the use of SuperDARN data. SuperDARN is a collection of radars funded by national scientific funding agencies of Australia, Canada, China, France, Italy, Japan, Norway, South Africa, United Kingdom and the United States of America. We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service, and OMNI data.

References

Bristow, W. A., & Jensen, P. (2007). A superposed epoch study of superdarn convection observations during substorms. *Journal of Geophysical Research: Space Physics*, *112*(A6). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006JA012049 doi: 10.1029/2006JA012049

Foster, J. C. (1983). An empirical electric field model derived from chatanika radar data. *Journal of Geophysical Research: Space Physics*, *88*(A2), 981-987. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA088iA02p00981 doi: 10.1029/JA088iA02p00981

Greenwald, R. A., et al. (1995). (1995), DARN/SuperDARN: A global view of high-latitude convection. *Space Sci. Rev*, *71*, 763-796.

Heelis, R. A., Lowell, J. K., & Spiro, R. W. (1982). A model of the high-latitude ionospheric convection pattern. *Journal of Geophysical Research: Space Physics*, *87*(A8), 6339-6345. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA087iA08p06339 doi: 10.1029/JA087iA08p06339

Heppner, J. P., & Maynard, N. C. (1987). Empirical high-latitude electric field models. *Journal of Geophysical Research: Space Physics*, *92*(A5), 4467-4489. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA092iA05p04467 doi: 10.1029/JA092iA05p04467

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010649 doi: 10.1029/2004JA010649

Papitashvili, V. O., Belov, B. A., Faermark, D. S., Feldstein, Y. I., Golyshev, S. A., Gromova, L. I., & Levitin, A. E. (1994). Electric potential patterns in the northern and southern polar regions parameterized by the interplanetary magnetic field. *Journal of Geophysical Research: Space Physics*, *99*(A7), 13251-13262. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JA00822 doi: 10.1029/94JA00822

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ridley, A. J., Deng, Y., & Tóth, G. (2006, May). The global ionosphere thermosphere model. *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*(8), 839-864. doi: 10.1016/j.jastp.2006.01.008

Roble, R. G., & Ridley, E. C. (1994). A thermosphere-ionosphere-mesosphere-electrodynamics general circulation model (time-gcm): Equinox solar cycle minimum simulations (30–500 km). *Geophysical Research Letters*, *21*(6), 417-420. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93GL03391 doi: 10.1029/93GL03391

Ruohoniemi, J. M., & Baker, K. B. (1998). Large-scale imaging of high-latitude convection with Super Dual Auroral Radar Network HF radar observations. *J. Geophys. Res*, *103*, 20,797.

Ruohoniemi, J. M., & Greenwald, R. A. (1996). Statistical patterns of high-latitude convection obtained from goose bay hf radar observations. *Journal of Geophysical Research: Space Physics*, *101*(A10), 21743-21763. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JA01584 doi: 10.1029/96JA01584

Schulz, M. (1997). Direct influence of ring current on auroral oval diameter. *Journal of Geophysical Research: Space Physics*, *102*(A7), 14149-14154. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JA00827 doi: 10.1029/97JA00827

Shepherd, S. G., & Ruohoniemi, J. M. (2000). Electrostatic potential patterns in the high-latitude ionosphere constrained by superdarn measurements. *Journal of Geophysical Research: Space Physics*, *105*(A10), 23005-23014. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JA000171 doi: https://doi.org/10.1029/2000JA000171

Siscoe, G. L., & Huang, T. S. (1985). Polar cap inflation and deflation. *Journal of Geophysical Research: Space Physics*, *90*(A1), 543-547. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA090iA01p00543 doi: 10.1029/JA090iA01p00543

Thomas, E. G., & Shepherd, S. G. (2018). Statistical patterns of ionospheric convection derived from mid-latitude, high-latitude, and polar superdarn hf radar observations. *Journal of Geophysical Research: Space Physics*, *123*(4), 3196-3216. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2018JA025280 doi: 10.1002/2018JA025280

Topliff, C., Cohen, M., & Bristow, W. (2019). Simultaneously forecasting global geomagnetic activity using recurrent networks. *Machine Learning and the Physical Sciences Workshop, Advances in Neural Information Processing Systems*.

Weimer, D. R. (2005). Improved ionospheric electrodynamic models and application to calculating joule heating rates. *Journal of Geophysical Research: Space Physics*, *110*(A5). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010884 doi: 10.1029/2004JA010884