December 2022

A Protocol to Build Trust with Black Box Models

TIMOTHY K. THIELKE
University of Wisconsin-Milwaukee

A PROTOCOL TO BUILD TRUST WITH BLACK BOX MODELS

by

Timothy Thielke

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Atmospheric Science

at

The University of Wisconsin at Milwaukee

December 2022

ABSTRACT

A PROTOCOL TO BUILD TRUST WITH BLACK BOX MODELS

by

Timothy Thielke

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Paul Roebber

Data scientists are more widely using artificial intelligence and machine learning (ML) algorithms today despite the general mistrust associated with them due to the lack of contextual understanding of the domain occurring within the algorithm. Of the many types of ML algorithms, those that use non-linear activation functions are especially regarded with suspicion because of the lack of transparency and intuitive understanding of what is occurring within the black box of the algorithm. In this thesis, we set out to create a protocol to delve into the black box of an ML algorithm set to predict synoptic severe weather patterns and discover if we can more closely observe what is occurring inside the algorithm. In doing so, we prove that despite the lack of context considered when creating the algorithm there can be some recognition of key synoptic features. This protocol is aided by the introduction of a novel visualization tool that acts to peer inside the hidden nodes of an artificial neural network to better diagnose the black box. To show that this protocol and tool have merit, we also consider 5 generalized questions that should be answered to develop trust with ML algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

Machine learning (ML) has become increasingly popular across a wide range of scientific disciplines, such as medical diagnosis (Bera et al. 2019), sports analytics (Hamilton et al. 2014), and financial planning (Camacho-Urriolagoitia et al. 2021). A contributing reason for this is due to its ability to predict complex phenomena and identify patterns within massive datasets. ML has been formally defined by several researchers since the term was first introduced in the 1950s. It is widely considered a subset of artificial intelligence (AI) where AI is a generally a computer algorithm that is meant to emulate human intelligence using logic. ML distinguishes itself from other AI by performing a given task without being given explicit instructions, but instead uses patterns within the dataset to produce an output (Roebber 2022). Further, ML can improve its performance for a given task by learning from its experience, narrowly defined, as suggested by Mitchell (1997) who states, "a computer program that is said to learn from experience E with respect to some class of tasks T and some performance measure P, if its performance on tasks T, as measured by P, improves with experience E".

Considering these ideas, many different types of algorithms used to study a dataset can be considered ML. For example, the popular multivariate linear regression (MLR) can be considered ML if it is created by the modeler to improve upon a given performance metric automatically. Decision tree algorithms are another popular algorithm that can be ML. Decision trees are a technique where a set of IF-THEN rules are used to split and categorize the data to make a prediction. With these two types of algorithms in mind, it can be a straightforward process to understand how and why the model produced the result it did due to their relatively

high transparency. It is that transparency that allows the user to build confidence and trust with those given models (Barredo et al. 2020). Artificial neural networks are another type of ML algorithm which are comprised of a single, or multiple, hidden layer of nodes that can be a mix of either linear or non-linear activation functions. The presence of these non-linear activations within the hidden layers decreases the model's transparency and thus makes it difficult to understand the ANN model. In these instances, the hidden layers are often referred to as an ANN's black box. In fact, without the inclusion of the non-linear functions, the ANN reduces to an MLR. If multiple hidden layers exist, then the ANN can be further classified as a deep neural network (DNN) adding more complexity to the model and its interpretation. The tradeoff for choosing the ANNs and DNNs more complex and less transparent algorithm is that they are more flexible than their simpler linear counterparts (Barredo et al. 2020). Despite this, without being able to intuitively delve into the black box of the model they are often labeled untrustworthy even if the output of the model is highly accurate and resolves the questions asked of the dataset. This problem is further compounded on by the lack of contextual understanding by the ANN algorithm. Because these algorithms train in a narrow and closely defined framework, they are unable to account for changes in context which does not emulate real world predictions. For these reasons, modelers resort to post-hoc model-agnostic approaches to explain and understand the ANNs inner workings and from there establish trust with the model.

Attempts at building trust with a black box model are ongoing with a large focus on being able to interpret a model's output, explaining why a model predicted X, and why a model predicted X instead of Y (Miller 2019). Barredo et al. (2020) survey the literature of explainable AI and state that to properly explain AI for a given audience, AI must "produce details or reasons

to make its functioning clear or easy to understand." There is a large emphasis on the audience for AI, or ML model, when attempting to explaining its output because the background knowledge of the domain can vary greatly depending on who is interpreting the model. They further discuss that there are generally six categories of post-hoc and model-agnostic techniques used to explain a more opaque type of ML, such as an ANN: text explanation, visual explanation, local explanation, explanations by example, explanations by simplification, and feature relevance explanation.

Text explanations are simply the model's ability to explain its result by generating text and symbols to support its output. Local explanations take a less complex localized segment of the solution space that is relevant to the whole model and use that to explain the model's output. Explanations by example is the process of extracting examples that relate to the result and show the relationships and correlations found by the model. Explanations by simplification represents the subset of explanations that require a simpler proxy model that resembles the original model. For example, in 2016, Ribeiro et al. (a) created a model-agnostic technique referred to as LIME (Local Interpretable Model-Agnostic Explanations) which fits a simple linear model to a set of slightly altered predictor values. From there they interpret the simpler model that acts as a proxy for the more complicated ANN. Similarly, Craven and Shavlik (1995) develop an algorithm that generates a decision tree to approximate the black box of an ANN. These proxy models may produce similar results as the ANN they represent, however, they can be unfaithful to that original model due to differing relationships and correlations the proxy model uses (Rudin 2019). Because of this, proxy models don't explain precisely how the output was determined which can

lead to misleading conclusions but can nonetheless be useful when interpreting a non-linear model.

Visual explanations are considered the best way to represent how input variables interact inside the ANN for those who are not familiar with ML modeling, but the process of generating such imagery is a difficult task (Barredo et al. 2020). For this reason, visual explanation imagery is typically paired with feature relevance techniques which are used to compute a relevance score for, and determine output sensitivity to, input variables. Krause et al. (2016) developed a post-hoc interactable visualization tool that generates a partial dependence biplot designed to show the relationship between input variables and can be used on several different types of ML models. Several examples of feature relevance techniques can be found in McGovern et al. (2019) where they test multiple model interpretation and visualization methods with a specific focus on meteorological problems and studies. One such method is impurity importance (Louppe et al. 2013; Breiman 2001) which has been applied to an ensemble approach to decision tree models, random forest, predicting convective storm mode and winter precipitation type. Impurity importance ranks the variables to determine relevance in the model by the average amount of times a feature within the dataset splits incorrectly across all the trees in the ensemble (Louppe et al. 2013). Permutation importance is another method in which variable importance is determined, by measuring how much a model deteriorates when permuting an input variable for all examples. The permutation importance measure is determined by comparing the permutation's performance to that of the original model. McGovern uses single-pass (Breiman 2001) and multi-pass (Lakshmanan et al. 2015) permutation importance on decision tree models, support-vector

machines, and convolutional neural network (CNN) models predicting convective storm mode, winter precipitation, and tornadogenesis.

After creating their visualization tool, Krause et al. (2016) interviewed users and developed 5 questions that their tool helped to aid. We consider these questions here, but have adjusted them to fit a more generalized topic:

1. What impact does an input variable have on the output prediction?
2. Does the model behave correctly on a case-to-case basis?
3. What are the most important features for a given output prediction?
4. Why are certain cases not being accurately predicated?
5. Can we identify high impact actionable features?

The purpose of this thesis is to create a protocol that, if followed, will answer these questions about non-linear black box models and in doing so will build user trust with that model. To accomplish this, we consider a simple, single layer ANN and analyze it using established, and hybrid adaptations of, feature relevance techniques, alongside a novel visualization method. Additionally, we generate a linear model using the same variable inputs as our ANN model for comparison. Both models will consider the entire testing dataset as well as a subset of the testing dataset that represents the primary mode of a given severe weather event along with the synoptic setting as classified by Miller (1972). These aspects will be discussed in greater detail in section 2. The results of these analysis and studies of specific events where the models did not perform well will be presented in section 3. Concluding remarks to follow in section 4.

II. METHODOLOGY

MODEL & DOMAIN

The data for this study comes from the ERA 5 reanalysis (Hersbach et al. 2020). Considering computational constraints, we selected data points that were spatially located between 38-44°N and 97-110°W which we hereafter refer to as the northern plains of the United States (Fig. 1). Further, we restricted the period of interests to March-September in 2010 and 2011.

As in Miller (1972), we will focus on synoptic scale features that promote severe convective thunderstorms. Considering this, we start with hourly geopotential height (GP), specific humidity (SH), temperature (TP), vertical velocity (WW), and both the U- and V-components (UU; VV) of winds at 250, 500, 700, and 850 hPa pressure surfaces as inputs in our model. Additionally, we included 2-meter temperature and dewpoint (DP) temperature, mean sea level pressure (MP), and 10-meter u- and v- component winds. We limit our study to just these input variables because they are the foundations for synoptic features. For example, convective available potential energy (CAPE) is an important measure to determine areas of instability and it can be calculated from the variables above. Adding additional, derived variables from the foundational measures can lead to collinearity and may confound interpretation. After some analysis, discussed later in this thesis, we further limited our model to only 14 variables that maintained the key features important to our model to simplify interpretation.

Using SPC's archived storm reports, we considered a 24-hour period (00 UTC to 23 UTC) to be severe if the number and concentration of reports suggested the presence of organized convective storms. Although the number of severe days to non-severe days are not equally observed, we did filter our data in such a way that half the days were severe and the other half were non-severe. We do this because the ratio of severe to non-severe days can vary based on geographical location, month, season, etc. and having too many "null" (non-severe) exemplars can produce poor training in ML models. However, we will take care to consider the possibility that this procedure will produce bias in the predictions.

To remove dimensionality and simplify our inputs to the ANN, we use k-means clustering on the complete set of variables collected from the ERA 5 reanalysis. The k-means clustering algorithm is another type of ML that is classified as unsupervised because it does not require human supervision to identify clusters. The algorithm iteratively finds k-number of mean cluster centers, where k is determined by the modeler. The algorithm then optimizes the cluster center selection by determining which value has the least amount of error (Likas et al. 2001). From here the algorithm measures the distance each data point is from each mean cluster center and then allocates the data points to the nearest cluster center. Our criteria for selecting a given k-number of clusters for each variable is that each cluster had at least 10 occurrences, the cubic clustering criterion (CCC; Sarle 1983) value as a function of k is maximized and had a value greater than +2, and all clusters are reasonably compact in 2-D space. The CCC value is determined by the within-cluster sum of squares and when maximized the error is minimized. When CCC is considered as a function of k, we can determine which k-value of clusters best represents the dataset. Occasionally, a CCC value would be maximized, but one or more mean

clusters may have less than 10 occurrences. Since we want to make sure that each cluster is reasonably represented within or model, we would look at the next highest value of k with a CCC still above +2 until all mean clusters had at least 10 occurrences. Finally, once the above two criteria are met, we would plot the data by the first two principal components to analyze the data to ensure that there were no extreme outliers. This clustering process eliminated several variables from consideration. Once this process was complete, each variable field was replaced by the mean cluster ID value for each hour of the dataset. For example, temperature at 500 hPa for a given time would be assigned a value of 30 if the variable field was closest to cluster centroid 30.

Other statistical methods were considered to reduce the dimensionality of our dataset, such as principal component analysis (PCA). PCAs represent a fraction of the entire dataset with which most of the variance is captured by a select number of eigenvectors while k-means clustering finds natural groupings of datapoints. PCAs would theoretically work well as inputs in our ANN algorithm, but because clustering highlights the individual synoptic features, while PCA does not, we are better able to analyze those features to gain further understanding, and therefore more trust, with our ML algorithm.

We use the k-means cluster identification values as categorical inputs in our ANN which was created using JMP Pro 16 software. To prevent time correlation within our model, we used the first two thirds of our 2010 severe weather season dataset for training and the remaining one third of the data for validation. For testing, we consider the whole 2011 severe weather season. Because our inputs are categorical, rather than continuous, each input value is assigned a starting

weight for each node within the hidden layer. Additionally, the sum of squares, the penalty parameter, is set to 0. Next, in an outer iterative loop "a nonzero candidate value of the penalty parameter is then chosen, and a univariate line search on the penalty parameter is undertaken" (Gotwalt 2011). While the search is being undertaken, the best value of likelihood value for the training dataset is recorded by the algorithm. In an inner iterative loop, the likelihood of the validation dataset is being recorded. To keep the model from overfitting, when the likelihood from the inner iteration no longer improves then the algorithm terminates. The model in which the best validation likelihood occurred is kept.

Although our aim was to experiment with techniques to build trust with any black box model, we wanted to use a model that performs well. Considering this, we conducted a trial-and-error process in which over 1,000 models were generated with varying architectures, such as altering input variables, the number of hidden nodes, and hidden node activation functions. For each model, we calculated the critical success index (CSI) and the squared error for both the training and validation datasets. We wanted our CSI value to be at least 0.5 and squared error relatively minimized. We selected the simplest model from those that satisfied our criteria for this study.

Table 1 denotes the long- and short-hand IDs for 14 input variables selected for the model. The selected model had one hidden layer with 6 nodes that all used the hyperbolic tangent (TanH) function. The TanH function acts like a step function but is continuously differentiable which allows the error backpropagation method to be used to adjust the model weights. Each hidden node sums the input variable weights for that given hour and halves that

value before applying the TanH function. The resulting value of each hidden node is then applied to a logit equation to calculate the probability for severe weather at that time.

DOMAIN EXPERTISE

We compare established literature and concepts within the severe weather forecasting domain to our model's output to better rationalize our model's predictions and further build trust with it. Our first concept is to simply observe the severe weather mode for the day in question. For this, we consider archived SPC storm reports and classified each day as either a hail, wind, or tornado day depending on which report was most common. In some cases, there were a near equal amount of two or more types of reports (e.g., Fig. 2). In this case there were a mix of hail and wind reports collected in Nebraska and rather than assigning the day either wind or hail we instead classified it as a wind and hail.

The Miller (1972) report outlines five classic synoptic weather patterns that promote convective weather. Although they consider multiple variables at several pressure levels, each pattern can be connected to specific frontal patterns. With this in mind, we use WPC's archived surface analysis to assign each day to a given pattern. Table 2 briefly describes the frontal patterns we consider when assigning a given Miller classification. Not every severe weather day in our data fits perfectly into a specific Miller classification. In these cases, we used subjective judgement to assign those days to the classification it most closely represented. In both the storm mode and Miller classifications, if the day was non-severe then it was classified as a null day.

LOGISTIC REGRESSION

Rudin (2019) states that a more complicated a model is not necessarily be more accurate one. It is true that in some cases a simple linear model can perform as well as, if not better than, a complicated black box model. To ensure that our model does not fall within this category, we will also generate a logistic regression model using the same dataset and compare the two models using the same methodology and protocol that we will with the ANN model. Further, the logistic regression model can also be used as a simplified model to better understand the implications of our methods and protocol, similar in concept to the LIME approach.

FEATURE RELEVANCE

Feature relevance, or variable importance, are a suite of techniques that determine which input variable, or variables, are most influential in the model's output. In this thesis, we focus on using permutation importance and backward sequential selection, both of which will have single- and multi-pass versions. For all techniques, we use a 10,000 bootstrap analysis considering confidence intervals of 0.05 and we observe the model's CSI and squared error. The reason behind using several ranking techniques and measures of success is twofold. The first is to gain a more holistic understanding of the influence of the variables considered. The second reason can be gleamed from McGovern et al. (2019) where they observe the permutation importance ranking of a random forest model and support-vector machine model. They note that the two models rank variables differently but are statistically similar, and for this reason they can't rely

on one interpretation over the other. Thus, it is important to consider several ranking methods to have a higher degree of confidence in the results.

Permutation importance (PI), as explained earlier, is when you randomly permute one variable across the whole dataset in a way that retains the original distribution of that variable. The reason for doing this is to break the statistical link between the predictor and predictand by assigning an improper value to the predictor. We would then compare scores between the two models, permuted and unpermuted, and if scores deteriorate drastically then we could reason that the variable is important. If the scores do not change by a large enough threshold, then it is either unimportant or redundant with another variable or set of variables. When initially proposed by Breiman (2001), this feature relevance technique was used with the random forests model, but the technique can be applied to many other ML models, including ANNs. PI acts to highlight one or more standalone input variables that can introduce uncertainty in our model like other sensitivity tests, but it does not answer the question of which variables as a group are important. To remedy this, Lakshmanan et al. (2015) adapted a multi-pass version of PI. In their study, they found which single variable was most important through permutation and then keeping that variable permuted while they conducted another systematic permutation of the remaining variables to determine which other variable was second most important and so on. Considering the variables in this way shows which set of variables capture most of the dataset and therefore have the highest influence on the predicted output.

Sequential selection (Stracuzzi and Utgoff 2004) is another technique that ranks the importance of variables by adding or removing input variables to or from a model. In forward

sequential selection, one begins with a model that predicts the mean output, adds input variables to that model, and observes how the model reacts to determine a given variables influence. However, this method does not fit well with our process. Instead, we will consider the opposite version, backward sequential selection (BSS), where we start with our original model and sequentially remove variables from consideration and observe how the model scores fluctuate. Like PI, we consider both the BSS scores individually, single-pass, and as a group, multi-pass.

NOVEL IMAGERY

As previously stated, for those who are not familiar with ML models, visual explanations are a preferred method to determine how variables interact (Barredo et al. 2020). Considering this, we developed a novel visualization tool that will help diagnose and highlight variables within each hidden node to understand what is occurring. This tool can be used across the entire dataset but is most effective in local examples. Additionally, this tool can be used as a qualitative source to understand variable importance instead of the quantitative approach discussed earlier. However, it works best when paired with the variable importance techniques we apply in this thesis.

This tool has an external view of the hidden nodes (Fig. 18) and an internal view for each hidden node (Figs. 19-21). The external view is meant to highlight which of the 6 hidden nodes are influencing the final output and the internal views allow a look inside the hidden node to determine which variables are most important to that hidden node. In our given algorithm hidden nodes 1, 5, and 6 are assigned the heaviest weights (Table 3) and therefore those 3 nodes

dominate the forecast in most cases. Further, combinations of which hidden nodes are active are important to consider. For example, when hidden nodes 5 and 6 agree on a deterministic forecast then the forecast will always align with those nodes, but if hidden nodes 5 and 6 disagree then whichever hidden node 1 agrees with will dictate the output. Additionally, the sign of the weight is important. In this algorithm, negative values increase severe probability. Therefore, if the result of the hidden node activation (through TanH function) is negative and the weight assigned to the hidden node is positive then the total value acts to increase severe probability. Once the inside view has been analyzed, one can investigate the important variables individually and conduct a sanity test to understand what is happening with the algorithm.

PROTOCOL

First, we need to ensure that our model is skillful and that it is a better choice than the logistic regression model. Following this, we want to get a base line understanding of each variable's importance across the entire model. From there, we want to determine if the algorithm is highlighting key features within our domain despite the lack of context. For this, we can investigate the Miller (1972) classifications as well as the severe mode. Further, it is important to also consider the null events to have complete understanding of the algorithm. We will then breakdown the influential variables to understand what features within those variables are influencing the algorithm most. Additionally, while considering subsections of our domain, we will also identify several case studies to gain insight into why the model may have performed well or poorly for those settings using the novel imagery previously discussed. This will give us the ability to perform sanity checks and further investigate variable inputs on an individual basis.

Finally, we will consider the 5 questions presented earlier to determine if our protocol and analyses are useful.

III. RESULTS

LOGISTIC REGRESSION

The logistic regression (LogReg) model that was generated using the same dataset as the ANN model suffered from extreme overfitting of the training dataset producing a model with no meaningful skill. The most probably reason for this is that we have too many extraneous variables for the LogReg model to effectively work. To remedy this, we reduced the number of input variables and tested each combination until we found the best performing LogReg model, (variables used for this model are shown in Fig. 3). With these adjustments, the LogReg CSI was 0.3787 compared to 0.4916 for the ANN model on the testing dataset.

MODEL-AGNOSTIC ANALYSIS

The LogReg model is a linear model so we can directly observe the weights assigned to each input. When comparing those weights to the ranking of BSS (Fig. 3) it becomes clear that they share the same order of importance as we would expect. There is also a difference between the two ranking methods supporting the idea of conducting several variable importance techniques. PI is meant to break the statistical link between the predictor and predictand allowing

it to resemble a sensitivity test and in so doing suggests that only 500hPa specific humidity and 850hPa v-wind are important in the LogReg model.

Comparing the different measures of importance and skill for single-pass PI and BSS (Fig. 4) we see 250hPa geopotential height and surface v-wind occurring in the top four in every example (the order of importance varies slightly). We see other variables (e.g., 850hPa geopotential height, 500 hPa temperature, and surface dew point temperature) alternate with the top variables. However, on an individual basis, most variables do add skill to the model. In contrast to that, variables such as 500hPa specific humidity and mean sea level pressure appear to detract from the overall model performance. Further, there is inconsistency between the rankings for several variables. For example, 250hPa v-wind adds skill to probabilistic forecasts, but harms deterministic forecasts in both BSS and PI. This suggests general unimportance since it can improve probability but not enough to overcome the threshold between severe and non-severe. Perhaps the most surprising contrast between the BSS and PI rankings is that of surface temperature. BSS ranks surface temperature amongst the least influential variables, but PI suggests it does provide influence. The likely cause of this is that surface temperature is correlated with another input variable (e.g., 700hPa temperature) since it decreases in value as surface temperature increases.

Additionally, we can speculate further by considering the differences between the two methods. BSS shows the direct relationship to the testing dataset while PI measures a perturbed version of the testing dataset. Therefore, it is possible that the model has a misplaced high

amount of importance on surface temperature features when surface temperatures do not play a significant role in the testing dataset.

Meteorologically, we can understand that model is placing importance on the strength and position of synoptic cyclones, or anticyclones, as well as if there is cooling or warming occurring in the mid-level. Because meridional winds at the surface are important and the region of interest is the northern plains, one can reason that the model is accounting for the role of temperature and moisture advection. However, the effects of these are complex which we will consider later in this analysis.

The first two variables of the multi-pass BSS ranking for the LogReg model are agreed upon by the single-pass version (Fig. 5), but in the third pass 500hPa specific humidity becomes more important. This change means that when combined with 500hPa geopotential height and 250hPa u-wind 500hPa specific humidity has the most important information of the remaining variables despite being less important by itself. PI, on the other hand, agrees entirely with the single-pass version as to the order of ranking. Single- and multi-pass also agree on importance considering the increase in model skill after second pass.

As one might expect, most input variables rank similarly to the way they did in the single-pass version (Fig. 6). This supports the idea that each input variable provides a small but statistically significant amount of important information to the model. One exception is that the surface v-winds appear to be less important or redundant in the multi-pass PI analysis. As with

surface dew point temperature, it appears the drop in rank may be associated with the increase in rank for surface temperatures.

We can consider this in more depth by observing how a specific variable behaves throughout the multi-pass (Figs. 7-9). Surface dewpoint temperature exhibits a general linear decline in importance with each pass for PI-Error and a drastic decrease in importance in the second and third pass in PI-CSI (Fig. 7). Despite this, surface dew point temperature largely remains influential in the model. In both cases, the most drastic decrease in importance is associated with 850hPa geopotential height and 250hPa temperature. Similarly, surface v-winds also show a drastic decrease in importance when associated with those variables in both PI-CSI and PI-Error (Fig. 8). The connection between surface dew point and surface v-wind makes meteorological sense (moisture advection) and the inclusion of 850hPa geopotential height can be explained by connecting the two surface variables to a synoptic scale pressure system. The connection to 250hPa temperature is less clear (we speculate that this value might represent a synoptic system's strength and location). Surface temperature is ranked the fourth most important variable in multi-pass for PI and we see that both the surface v-wind and surface dew point only marginally improve the model after inclusion. With this, we conclude that the surface variables are connected but still provide useful information when considered individually.

PI and BSS disagree on how 250hPa temperature influences the model (Fig. 9). BSS suggests this variable becomes important eventually, but because we are consistently removing variables, its importance is likely the result of being one of the few variables remaining. PI, on the other hand, suggests that 250hPa temperature quickly loses importance after a given pass.

Further investigation reveals that if 500hPa temperature is included, 250hPa temperature does not provide substantial additional information.

HAIL

For hail, the LogReg model provided no additional insight (Figs. 3 and 5). For the ANN model, there is an increase of importance for both the surface v-wind and 500hPa temperature fields (Figs. 10 and 11). Meteorologically, this connects with knowing that hail growth is dependent upon cold temperatures in the mid-levels. Further, the inclusion of moisture and temperature advection at the surface can also be connected to convective updraft strength, which also plays a critical role in hailstone growth. Otherwise, we see that variables that were important across the whole model (e.g., 250hPa geopotential height) remain influential while uninfluential variables (500hPa specific humidity and surface mean sea level pressure) stay that way as well.

To support the above explanations, one can look further into surface v-wind and 500hPa temperature by observing the BSS change for both CSI and squared error on an individual cluster level to determine if the algorithm is identifying key features for hail growth.

Although surface v-wind and 500hPa temperature are important and one can assume that the key features mentioned early are in fact playing a role in the algorithm's output, we can't be certain. To verify if this is the case, we conducted another BSS analysis (single-cluster BSS) considering only 500hPa temperature (or surface v-wind, respectively). Rather than removing the entire variable across the whole column, we remove only a single cluster from the given variable

and retain the rest to determine what single cluster is the most influential for the given variable. For the 500hPa temperature field (Figs. 12-14), the three most influential clusters (34, 21, and 48) have regions of relatively cold air in the mid-levels supporting our cold core as a key feature theory. Surface v-wind (Figs. 15-17) results are harder to interpret. Each pattern could explain a frontal boundary based on the reversal of wind direction, but the more surprising feature is that clusters 9 and 17 are associated with a core of relatively strong northerly flow. This suggests that rather than warm air and moisture advection, the algorithm considers synoptic scale cold air advection to be a key feature for hail growth.

Consider a specific case. On July 28, 2011, a shortwave trough passed through central North Dakota and into northeastern South Dakota during the mid to late afternoon (20-23 UTC) generating several discrete convective storms that produced scattered severe hail and isolated severe wind risks. As mentioned previously, hidden nodes 1, 5, and 6 have the heaviest weights and each supports a severe forecast for the given time (Fig. 18). Meanwhile, hidden nodes 2, 3, and 4 show mixed results, but because they have low weights in the final output the mixed results do not present in the final forecast. Focusing on the most influential hidden nodes (Figs. 19-21), 500hPa temperature plays an important role, as we would expect, by supporting a severe forecast in each hidden node. Of note, surface v-wind is not helpful in this case. In hidden node 5 (Fig. 20) it supports a severe forecast, but in hidden node 6 (Fig. 21) it opposes a severe forecast. We speculate that this is due to the lack of frontal boundary associated with this case. This conflict turns out to be a moot point because the result of the total weights supports a severe forecast for both nodes. For this specific case, the 500hPa temperature cluster that has such a high influence on the forecast is cluster 34 (Fig. 12). As we noted before, cluster 34 is amongst

the most influential individual clusters for hail forecasts and it depicts a cold core as a key feature giving the final output some meteorological support for trustworthiness.

MILLER TYPE B SYNOTPCI SETTING

Of the 5 synoptic settings presented by Miller (1972) the type B setting represents the frontal patterns subset. Typically associated with this synoptic setting is a strong surface low pressure center with warm and cold frontal boundaries and strong cold air advection occurring behind the cold front. A warm low level jet transports moisture from the south and there is a well-defined dry intrusion in the mid-levels. With this setting, frontal, and pre-frontal, squall lines typically form along and ahead of the cold front.

Like the hail analysis, the LogReg model did not provide new insight when considering only days where a type B synoptic setting occurred. The ANN, on the other hand, highlighted a few changes to which variables were most important (Figs. 22-23). Unlike the hail and full dataset analyses, there is a greater separation between the most influential variables and the lesser and uninfluential variables which can be seen in all the graphs other than the BSS multi-pass analyses. While 250hPa geopotential height may have the greatest influence (Fig. 22), BSS suggests that surface dew point temperature is the next most important variable and PI indicates that 500hPa geopotential height is more influential. Both the BSS and PI analysis may be right if we consider the importance of the strong low-pressure center and low-level moisture needed for this synoptic setting. Of interest, the multi-pass PI analysis greatly increases the influence of 500hPa specific humidity which could be an attempt to highlight the mid-level dry intrusion. We

conducted a single-cluster BSS on 250hPa geopotential height and found that both zonal patterns (Fig. 24) and patterns of approaching upper-level troughs (Fig. 25) are important. We conducted the single-cluster BSS analysis on other important variables (e.g., 500hPa temperature in Fig. 26, 500hPa geopotential height in Fig. 27) and see similar, but more defined, features which support what we know typically occurs with a Miller (1972) type B synoptic setting.

On June 25th, 2011, a cold front moves across North Dakota forming a squall line at 2100 UTC. As the front and squall line cross the region it gains intensity at 0000 UTC on the 26th of June and becomes severe, producing a mix of severe hail and wind reports. Unlike the hail case, our ANN did not perform well on this day (Fig. 29). Hidden node 1 and 5 appear to be consistent in indicating that severe weather is not likely while hidden node 6 provides a more mixed indication. This results from the 250 hPa geopotential height (Figs. 30-32). As previously noted, 250hPa geopotential height is in general an influential variable. This can be understood by noting the weight differential between 250hPa geopotential height and the other variables. At 1900 UTC, 250hPa geopotential height changes from cluster 13 (Fig. 33) to cluster 55 (Fig. 34), which lowers the weight in both hidden node 1 and 5 and explains why the forecast changed for those nodes. In hidden node 6 the change acts to increase severe probability, but not enough to change the hidden nodes output to severe. By comparing the 250hPa geopotential height cluster 55 to 13, it makes sense that change would decrease severe probability, as it did in hidden nodes 1 and 5, because there is an increase in geopotential height between the two clusters. Unfortunately, the other variables were not weighted in a way to maintain a severe forecast after this change.

NULL FORECASTS

It is also important to understand what features the algorithm indicates are influential in a non-severe forecast. As with hail and Miller type B synoptic settings, we restrict our data to only days that are categorized as non-severe for this portion of the analysis. Because our deterministic score requires a correct positive forecast (hit) we only consider the probabilistic score here rather than change the metrics we have used up to this point.

Somewhat surprisingly, we find that there is a reversal of importance in this context. This is particularly the case with the single-pass methods for surface temperature (Fig. 35), which was considered an uninfluential variable in all the previous analyses. Similarly, 250hPa geopotential height decreases in ranking and importance to the point where it is an unimportant variable when considering the PI method. On the other hand, we see that some of "mixed importance" variables, such as 850hPa geopotential height, remain that way. Given this, it is likely that certain variables drastically increase or decrease severe probability. Originally, one would expect surface temperature to be an important variable for severe convective weather, but the algorithm reversed the role and has it decreasing severe probability. For the multi-pass (Fig. 36), the BSS is similar to that seen in the single pass, while PI focuses on the same top 3 before a drop in influence of the other remaining variables.

Of all surface temperature clusters, cluster 49 (Fig. 37) ranks as the most influential for decreasing severe probability in our null cases. Unfortunately, cluster 49 resembles many of the other mean cluster stamps with relatively cold temperatures to the west and relatively warm temperatures to the east giving the appearance of a frontal boundary. Since most surface

temperature clusters look similar, this may also explain why this variable was not selected to increase severe probability. To understand why cluster 49 stands out, we took the differences between cluster 49 and the mean surface temperature field from the dataset (Fig. 38), and it becomes clear that this cluster has a weaker temperature gradient across the entire field. Despite many of the clusters resembling each other, our algorithm did distinguish the slight differences in the surface temperature fields to assign appropriate weights for the severe probability.

On June 4th, 2011, a high-pressure system was established over the northern plains behind a cold front that has passed through the region the day before (producing several severe wind risks). Despite this transition, the algorithm has trouble switching from the severe risk of the day before (Fig. 39). This is due to the conflicting indications from hidden nodes 5 and 6, with the former suggesting a severe forecast and the latter suggesting a non-severe forecast. Because of the way the nodes are weighted this allows hidden node 1 to drive the result, which is an incorrect positive (false alarm). It is not until hidden node 5 changes to a non-severe forecast that the algorithm is corrected. With hidden node 5 (Fig. 40), the variable weights are more balanced then in previous figures, but the variable change that corrected the forecast for hidden node 5 was surface temperature. At the time of interest, the surface temperature transitioned from cluster 26 to cluster 35. In Figure 41, we can see that this change acted to weaken the surface temperature gradient and decrease severe probability.

IV. CONCLUSION AND DISCUSSION

ANNs with non-linear activation functions are considered untrustworthy due to the lack of transparency and intuitive understanding of what is occurring within the black box, but also because there is no contextual understanding occurring within the algorithm. In this thesis, we set out create a protocol to delve into the black box and discover if we could more closely observe what is occurring inside the algorithm. In doing so, we hoped to prove that despite the lack of context considered when creating the algorithm there can be some recognition of context that comes in the shape of highlighting key features. Our protocol and associated use of the novel visualization technique introduced here helped to accomplish this task by identifying key variables and associated features in the domain that are important to specific synoptic settings and highlight areas within the model's hidden nodes that can be investigated to explain the model's success or failure on a case-to-case basis.

Outside the scope of this thesis, but still important to consider, is a more in-depth analysis of correlation between variables. We theorized several correlations by observing single- and multi-pass behaviors, but a more complete analysis into the correlations would be beneficial. In meteorology, many variables are interconnected to each other as can be observed through physical equations and principles, such as the Law of Thermodynamics and the quasi-geostrophic equations. For one to trust if a non-linear black box algorithm can pick up on contextual clues of the synoptic setting, we would want to believe that these correlations existed in the algorithm too. Additionally, it could prove to be beneficial to test this protocol on an algorithm that was trained and tested on a synthetic dataset, like created by Mamalakis et. al (2022). By using a dataset where the ground truth is known a priori we could determine if the algorithm correctly picks up on key features. Finally, this is a highly simplified algorithm and it

would be beneficial to test it on deep neural networks or different black models (e.g., random tree forest). Despite all of this, our protocol and tool shows promise as a base for modelers to use in order to develop and build trust with a black box model.

V. FIGURES

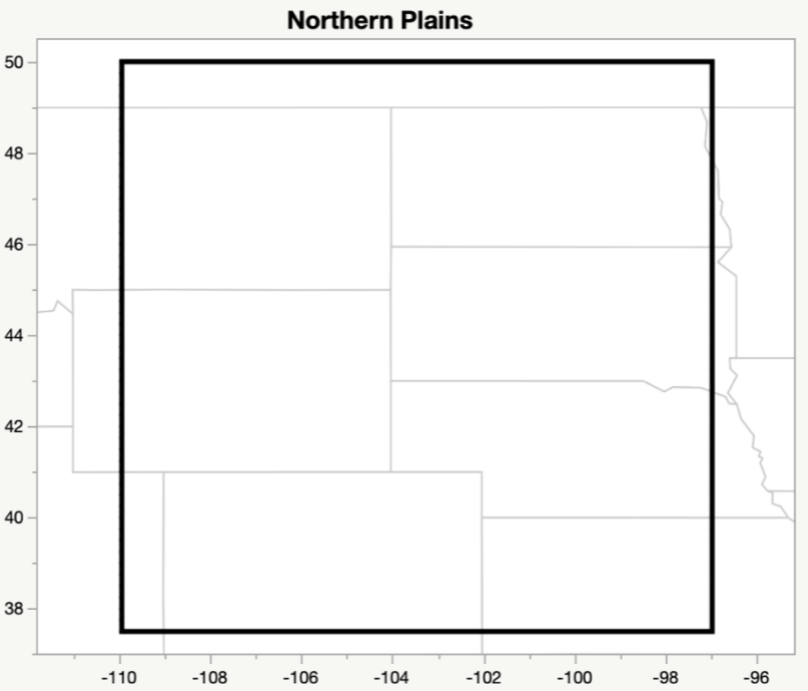

Figure 1. Northern plains region set between 38 N to 44 N latitude and 97W to 110W longitude.

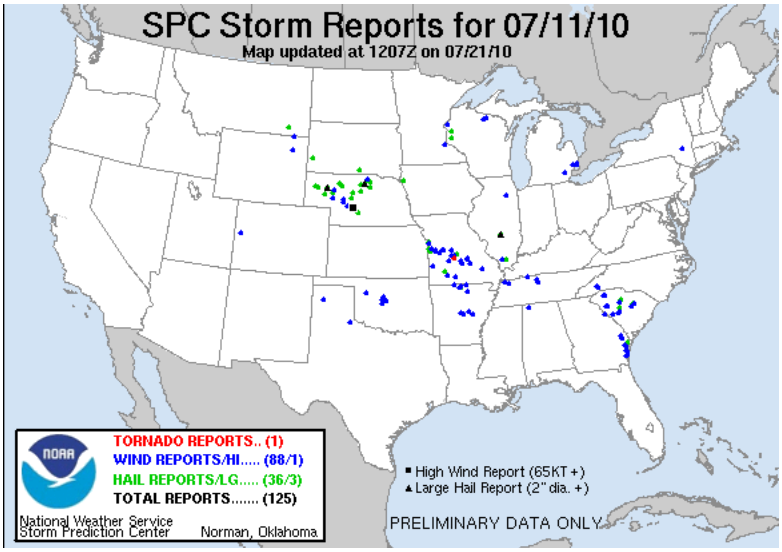

Figure 2. Archived SPC report of July 11, 2010 depicting a mix of hail and wind reports over Nebraska warranting a unique classifier code for a mix of severe hail and wind. Collected from https://www.spc.noaa.gov/climo/reports/100711_rpts.html

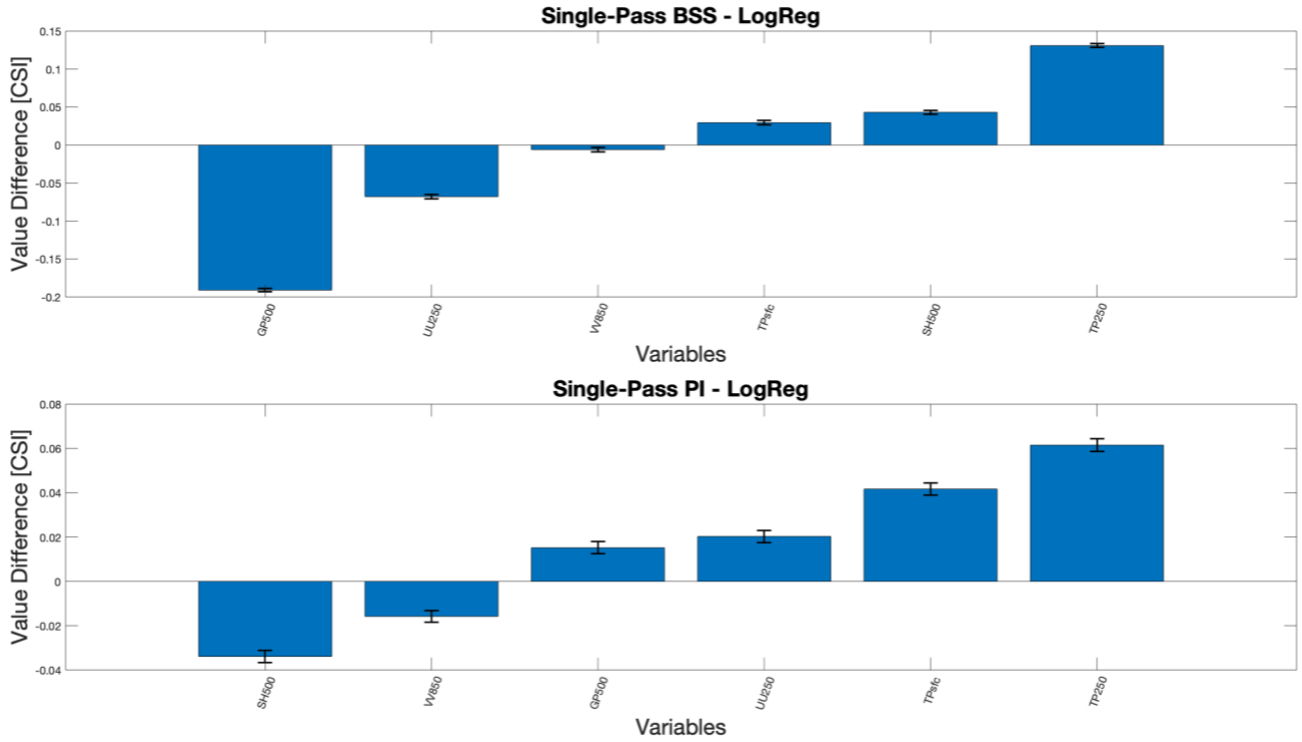

Figure 3. Comparison of single-pass rankings of CSI for both permutation importance (bottom) and backward sequential selection (top) for all LogReg input variables. Values are the difference between the original model score and the score of the model after the change to the input variable with negative values indicating importance and positive values indicating unimportance. Error bars show the 5th and 95th percentiles for 10,000 bootstrap replicates of the testing dataset.

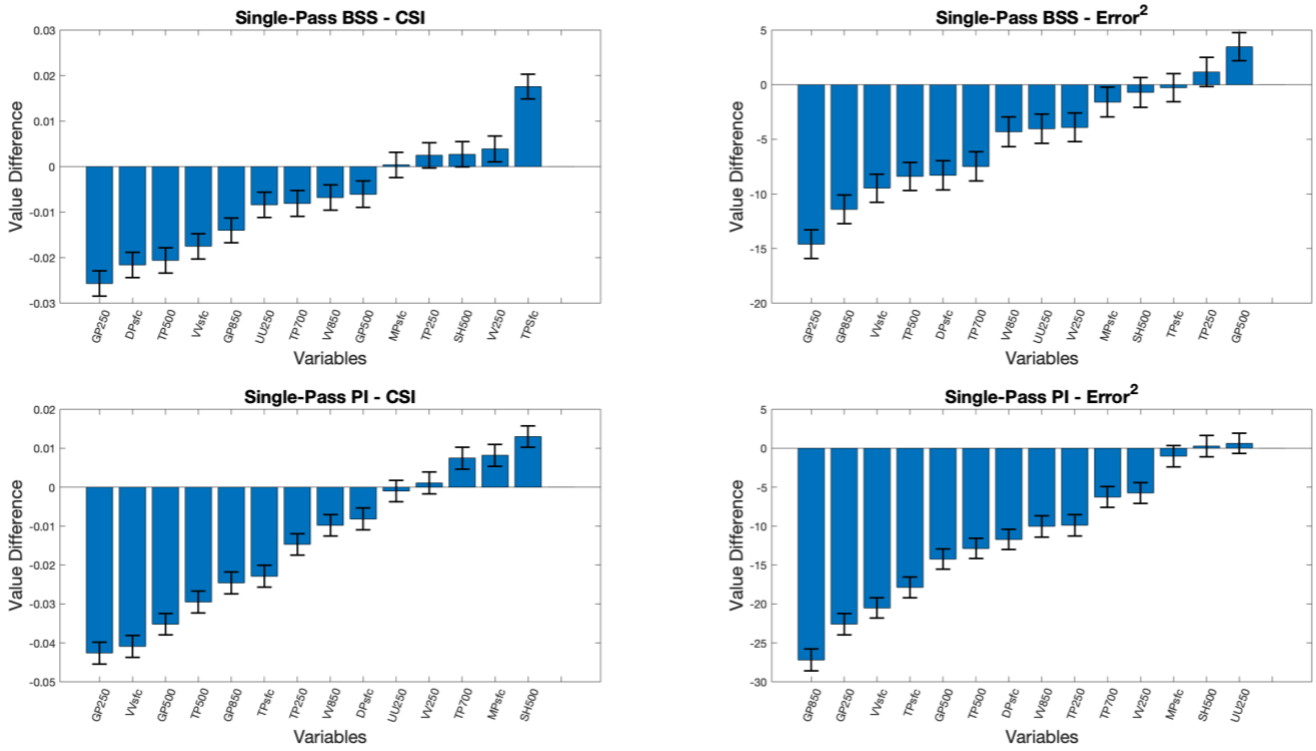

Figure 4. Comparison of single-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all ANN input variables. Values and error bars are the same as they are for Fig. 3.

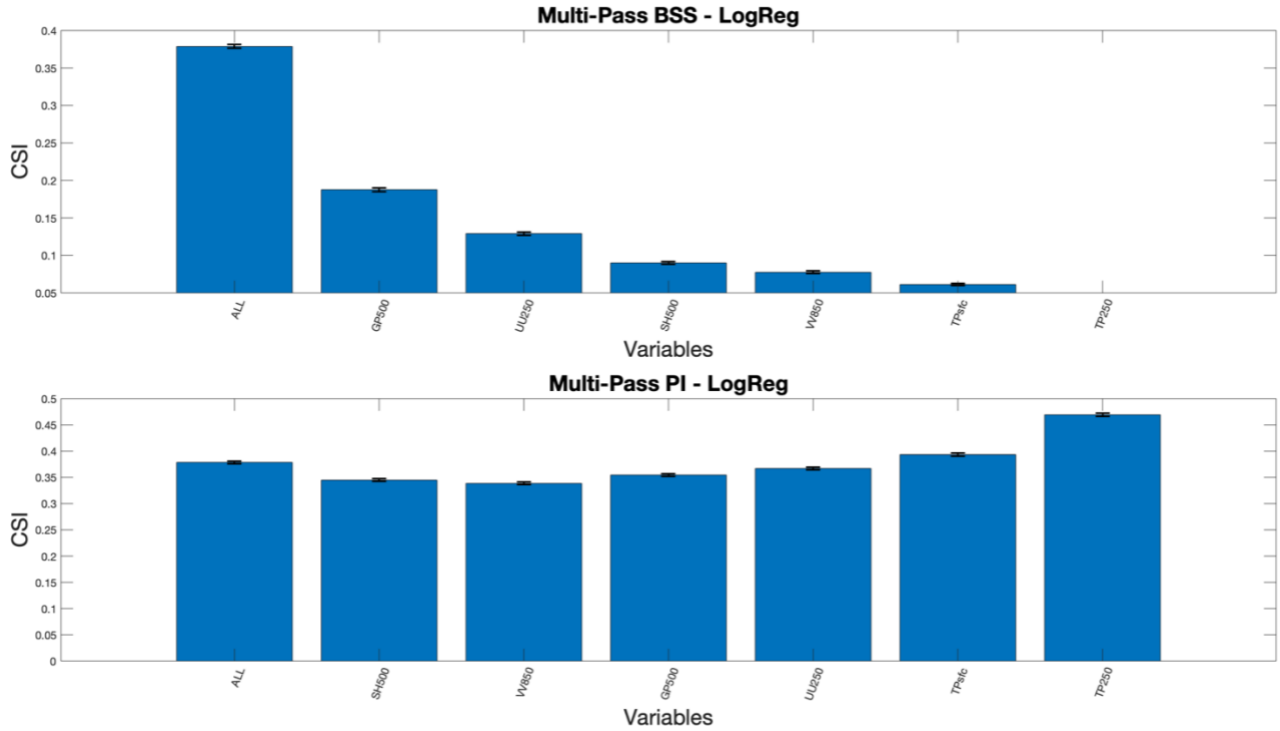

Figure 5. Comparison of multi-pass rankings of CSI for both permutation importance (bottom) and backward sequential selection (top) for all logistic regression input variables. Values are the successive loss in skill as each variable is removed or perturbed. Input variables are altered in order of importance such that early variable removals result in a greater loss of skill. Error bars show the 5th and 95th percentiles for 10,000 bootstrap replicates of the testing dataset.

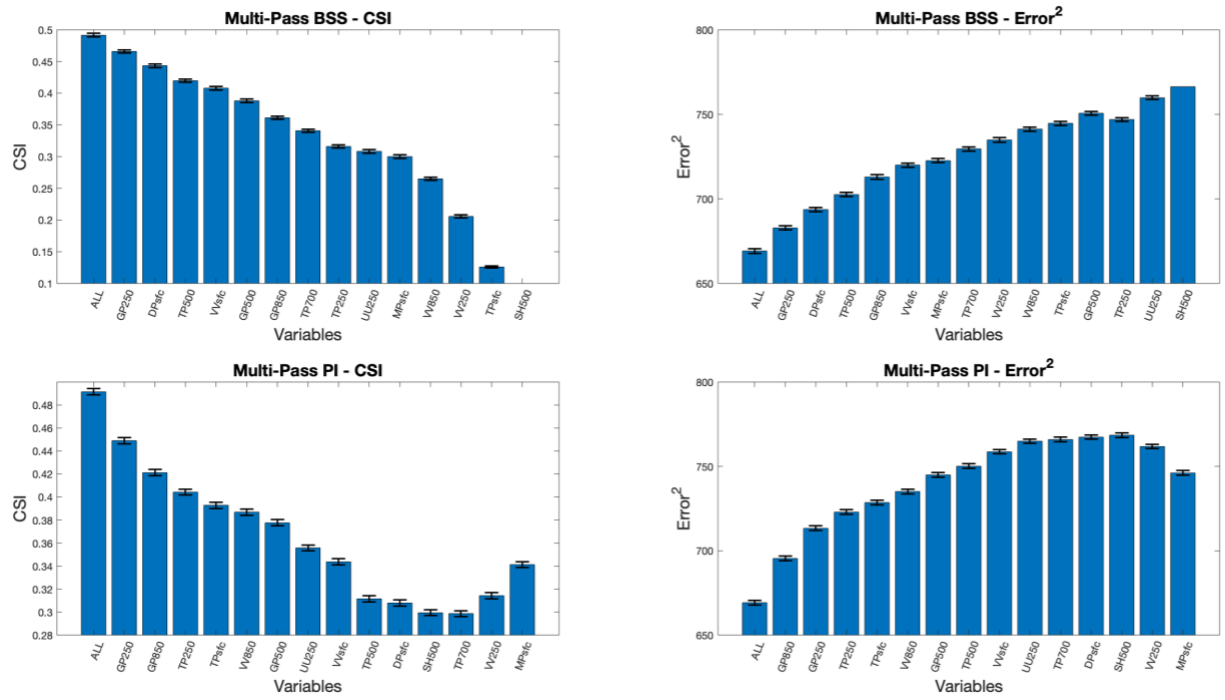

Figure 6. Comparison of multi-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all ANN input variables. Values and error bars are the same as they are for Fig. 5.

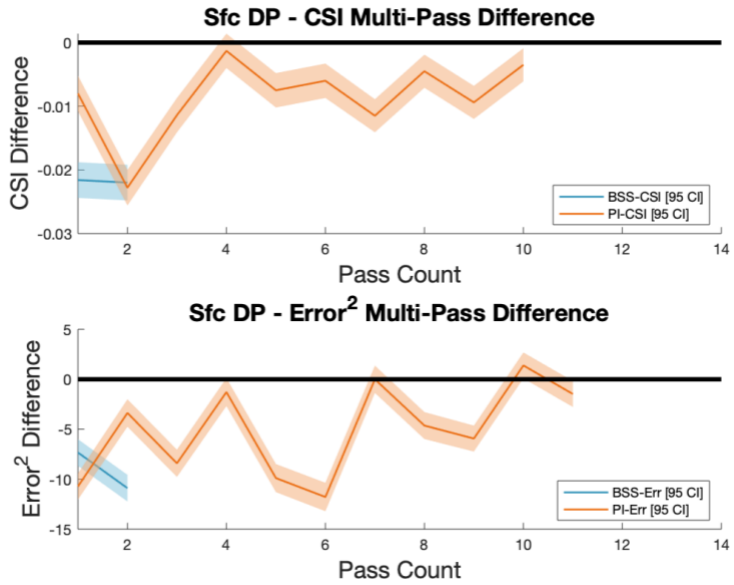

Figure 7. Comparison of surface dew point temperature difference values for multi-pass CSI (top) and squared error (bottom). Values are the difference between the model score of the current pass and the previous pass for BSS (blue) and PI (orange). Negative values suggests the variable is important and positive values means the value is unimportant. Shading represents the 5th and 95th percentiles for 10,000 bootstrap replicates of the testing dataset.

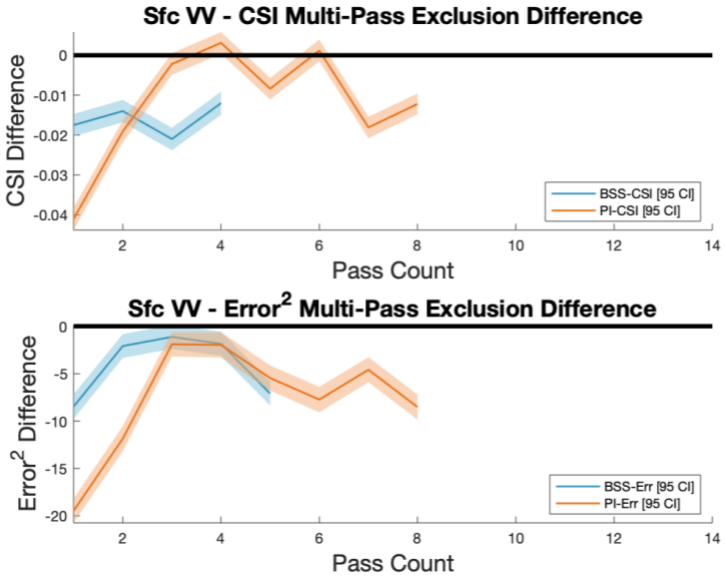

Figure 8. Comparison of surface meridional wind difference values for multi-pass CSI (top) and squared error (bottom). Values, colors, and error as the same as Fig. 7.

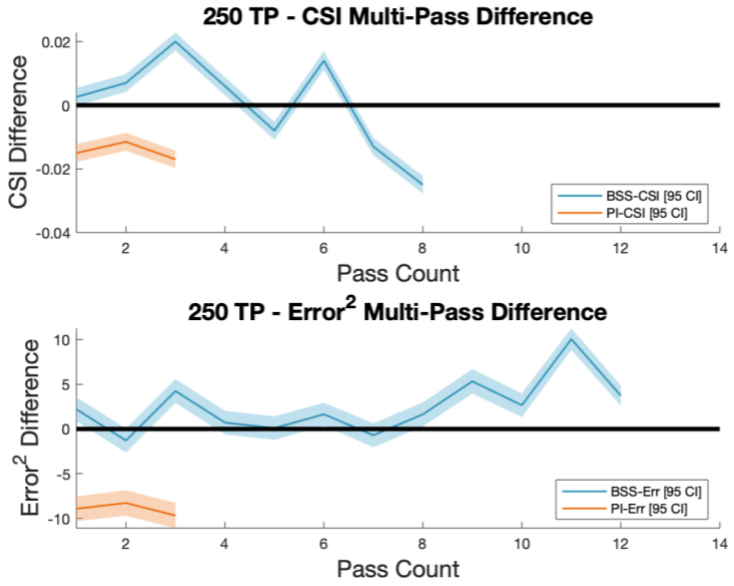

Figure 9. Comparison of 250hPa temperature difference values for multi-pass CSI (top) and squared error (bottom). Values, colors, and error as the same as Fig. 7.

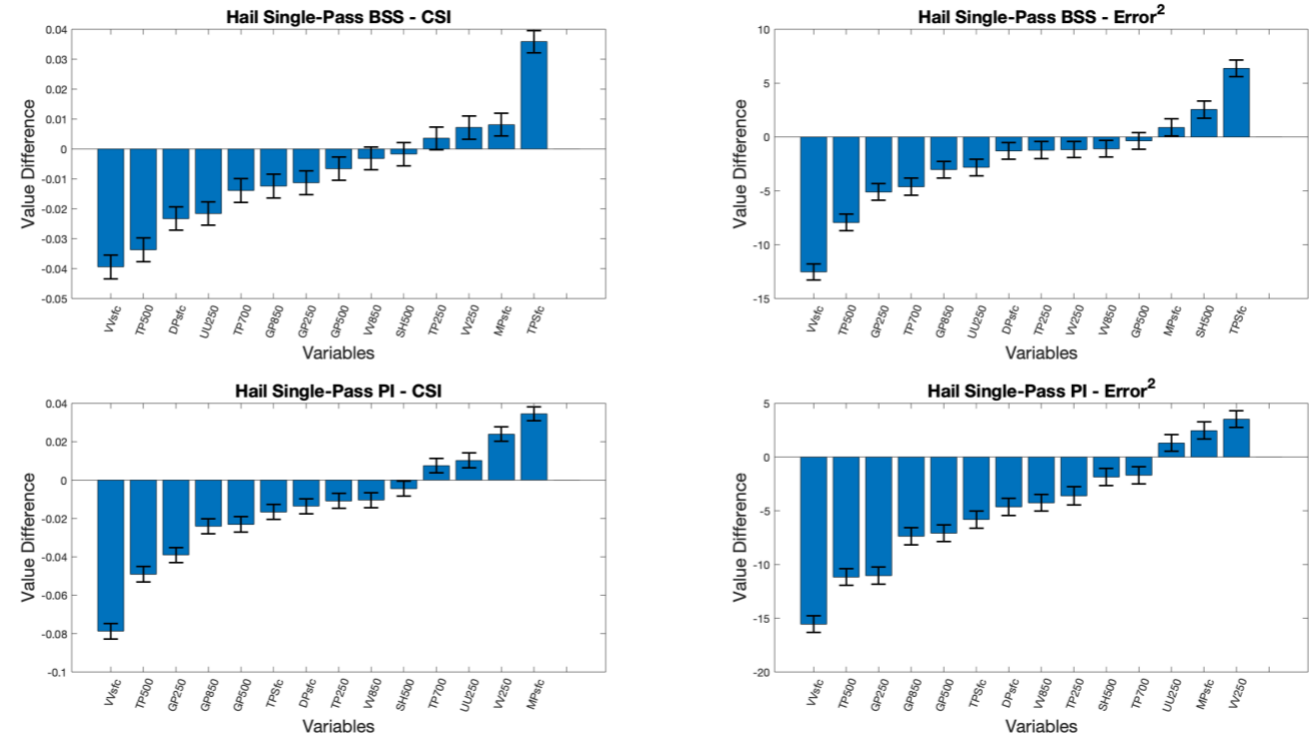

Figure 10. Comparison of single-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all severe hail days. Values and error bars are the same as they are for Fig. 3.

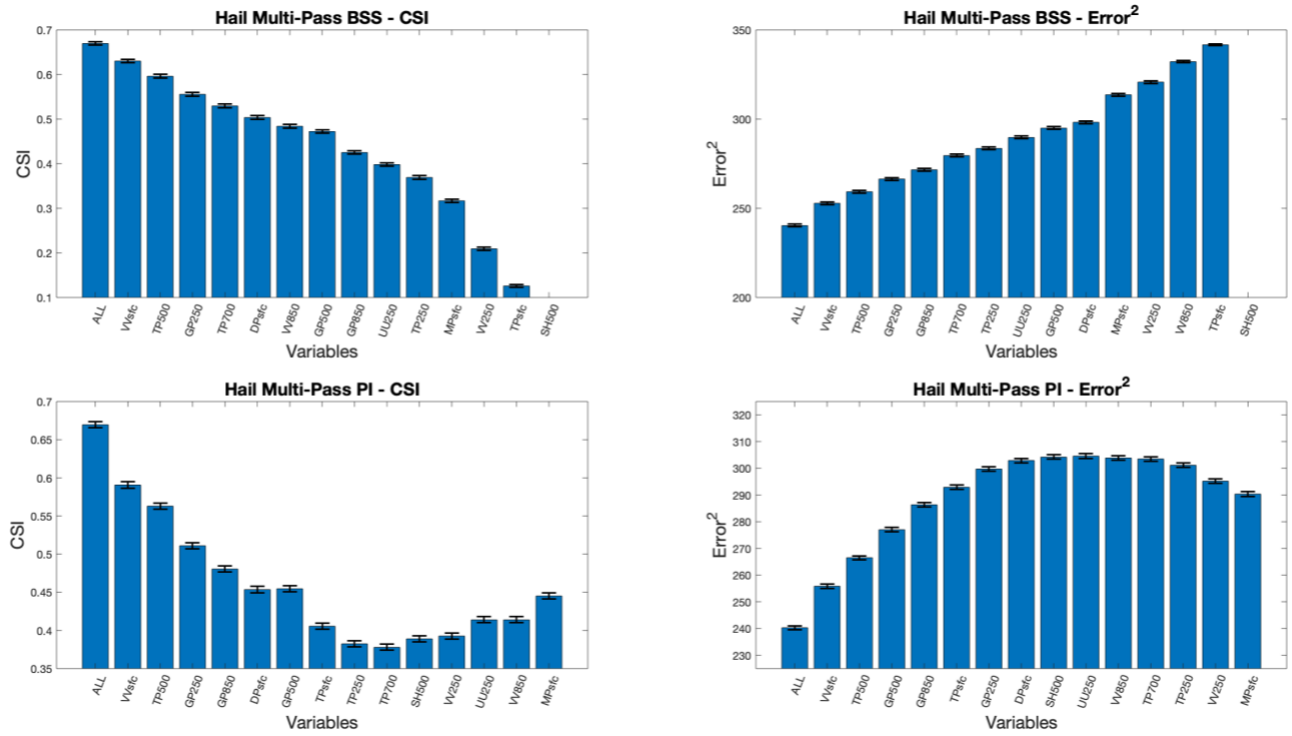

Figure 11. Comparison of multi-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all severe hail days. Values and error bars are the same as they are for Fig. 5.

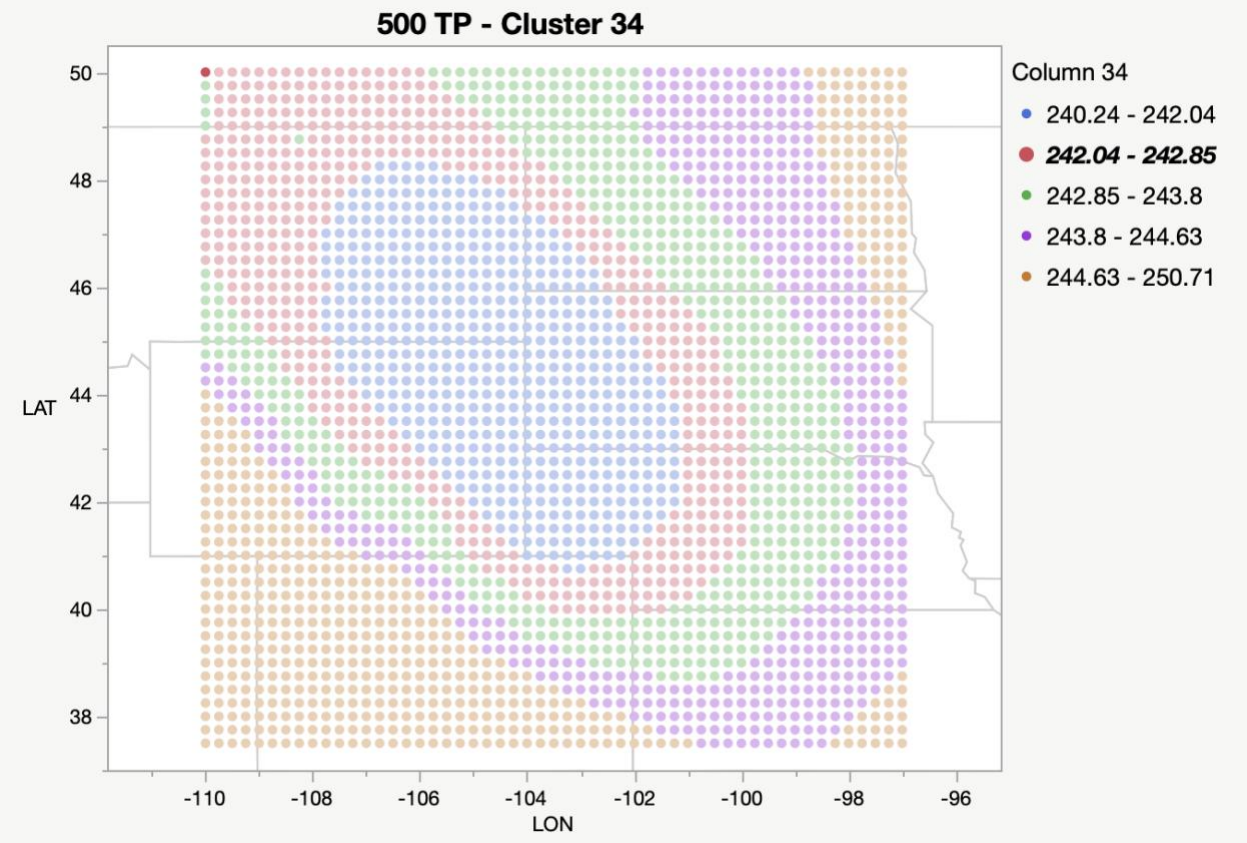

Figure 12. The composite image of 500hPa temperature fields associated with the cluster centroid 34. Temperature is measured in units kelvin (K).

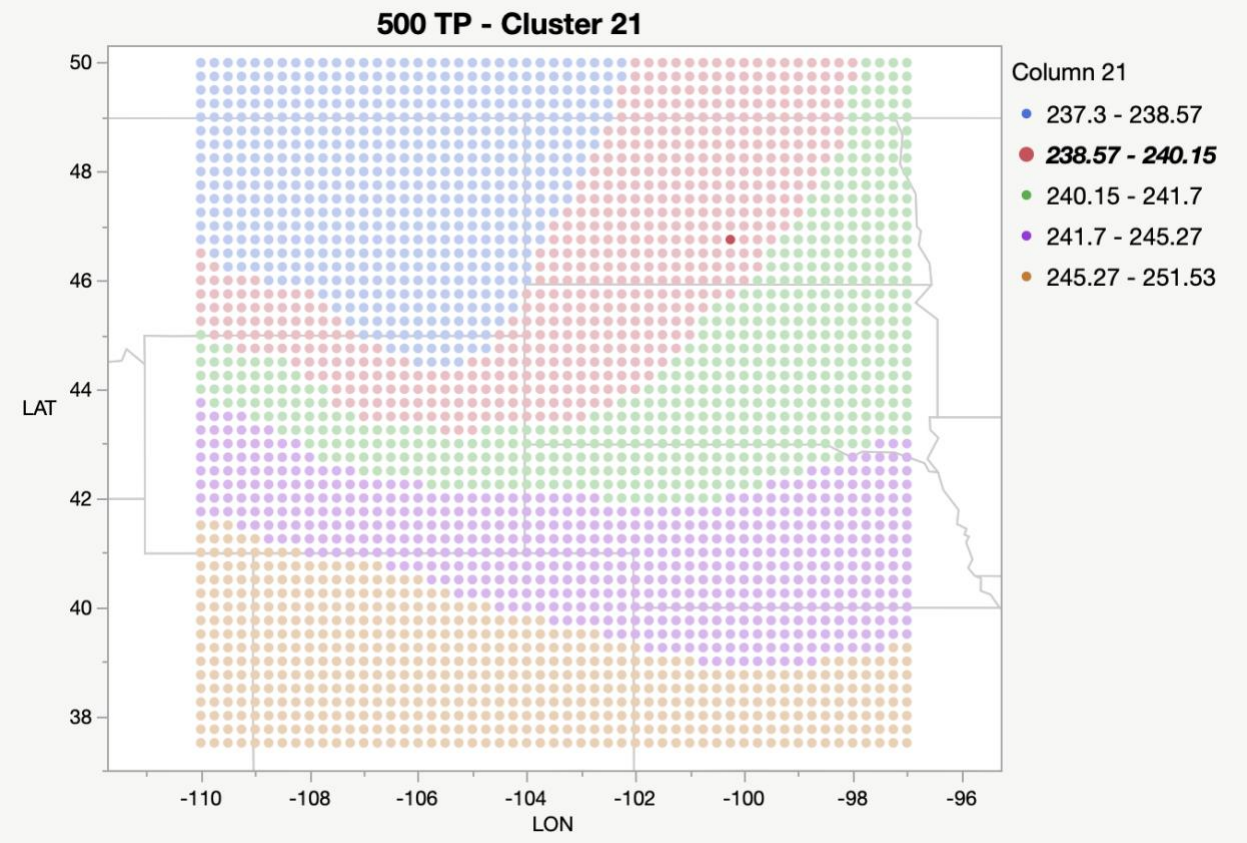

Figure 13. The composite image of 500hPa temperature fields associated with the cluster centroid 21. Temperature is measured in units kelvin (K).

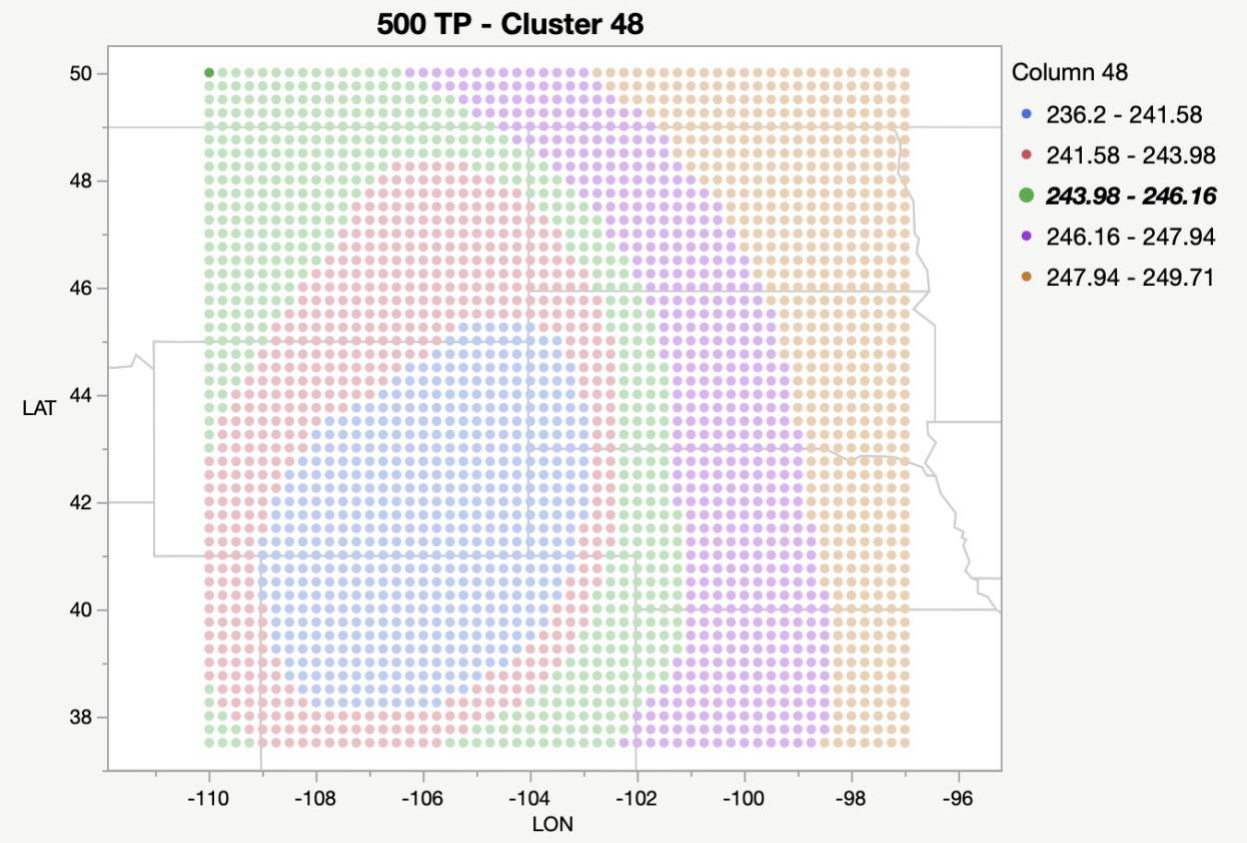

Figure 14. The composite image of 500hPa temperature fields associated with the cluster centroid 48. Temperature is measured in units kelvin (K).

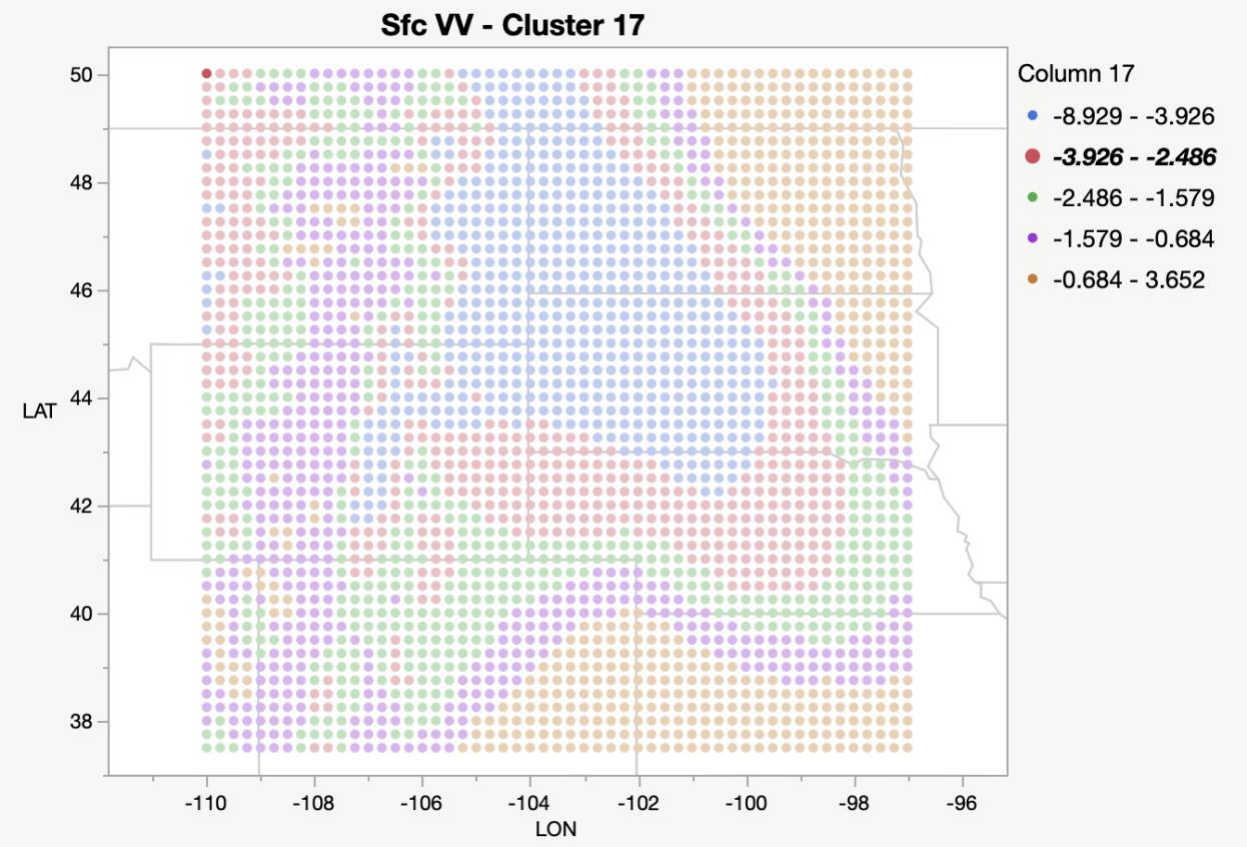

Figure 15. The composite image of surface v-wind fields associated with the cluster centroid 17. Positive values represent south to north flow and negative values represent north to south flow. Velocity is measured in ms^{-1}.

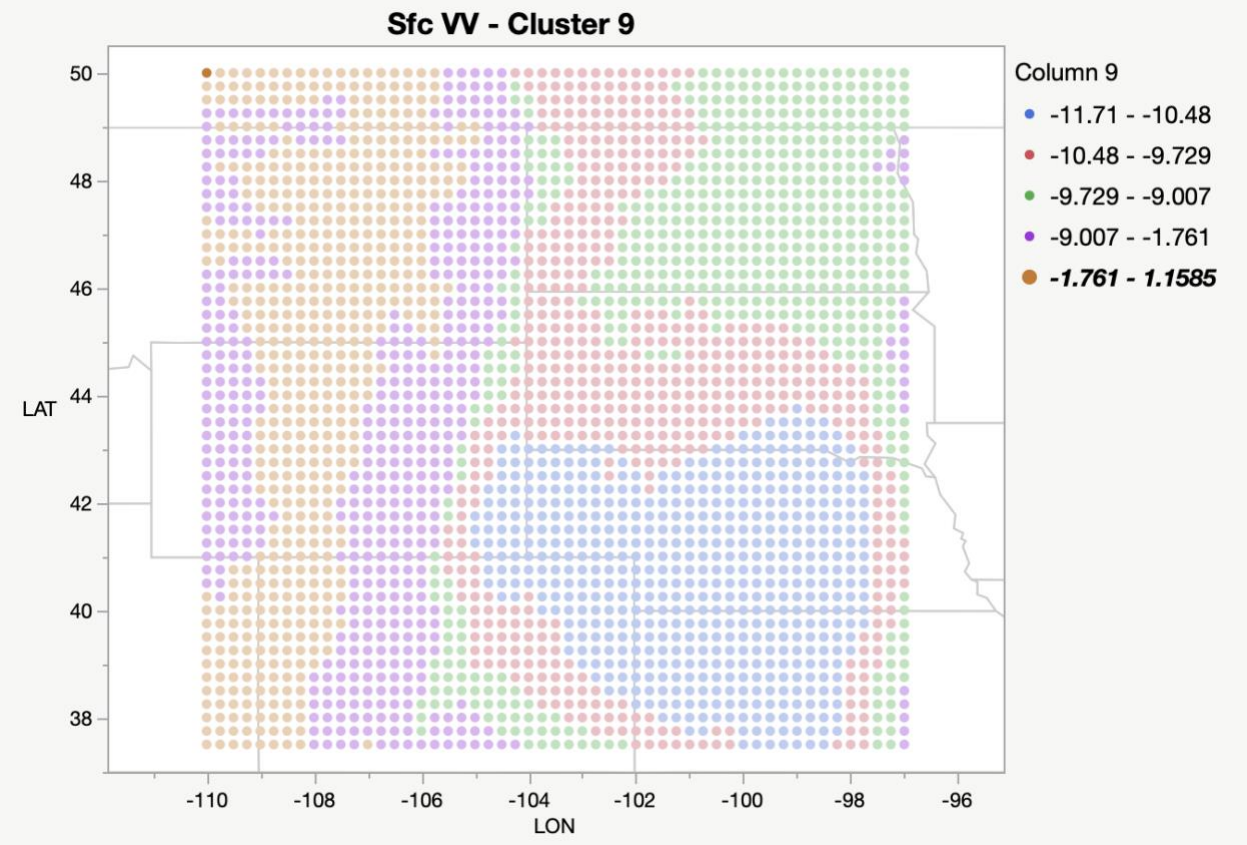

Figure 16. The composite image of surface v-wind fields associated with the cluster centroid 9. Positive values represent south to north flow and negative values represent north to south flow. Velocity is measured in ms^{-1}.

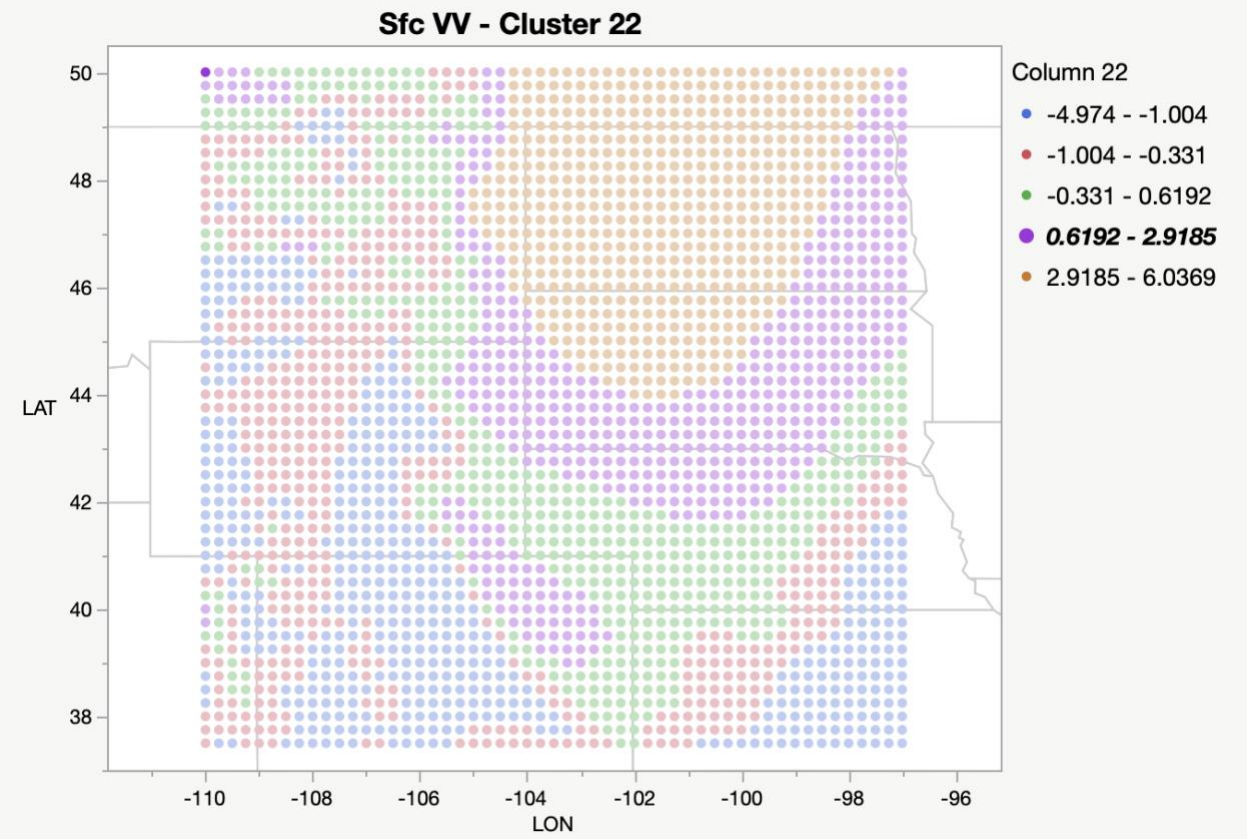

Figure 17. The composite image of surface v-wind fields associated with the cluster centroid 22. Positive values represent south to north flow and negative values represent north to south flow. Velocity is measured in ms^{-1}.

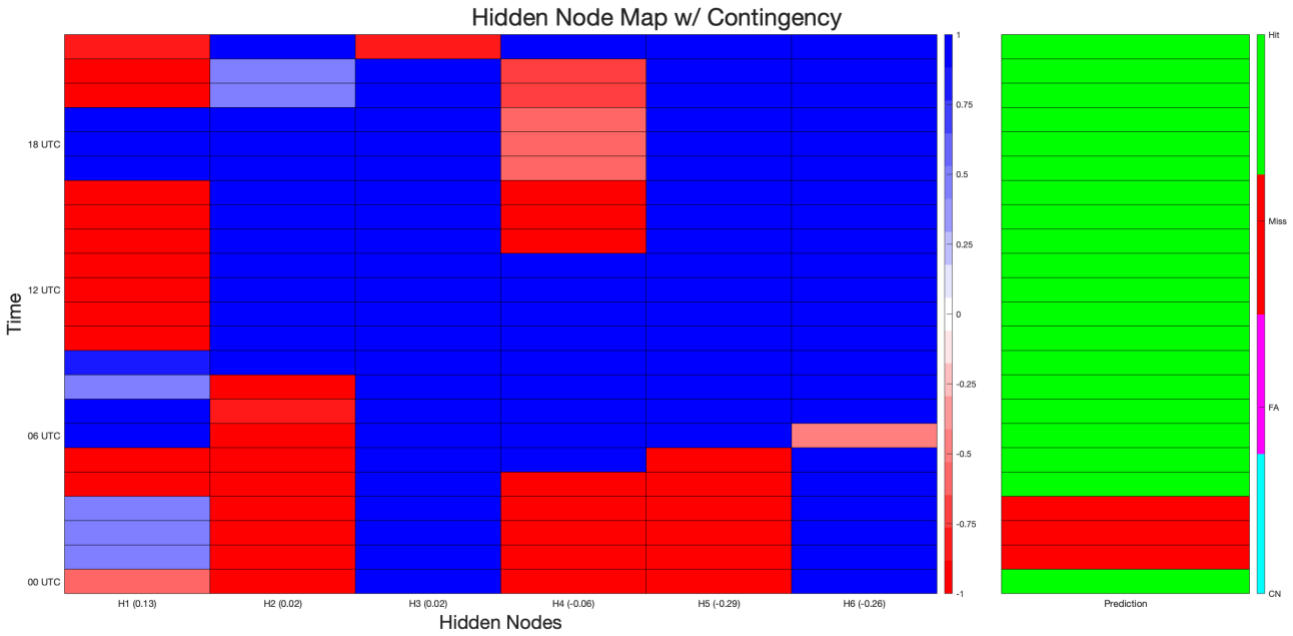

Figure 18. A map depicting the values of the hyperbolic tangent activation functions for each of the six hidden nodes (left) and the forecast contingency (right) for July 28, 2011, from 00 UTC (bottom) to 23 UTC (top). In the left diagram, shades of red are negative outputs of the function and shades of blue are positive outputs of the function. A weight, seen in Table 3, is assigned to the function output. If the total value of the function multiplied by the weight of the hidden node is negative, then severe probability increases and vice versa for positive values. In the right diagram, green represents a correct positive forecast (hit), red represents an incorrect negative forecast (miss), pink represents and incorrect positive forecast (false alarm), and blue represents a correct negative forecast (correct negative).

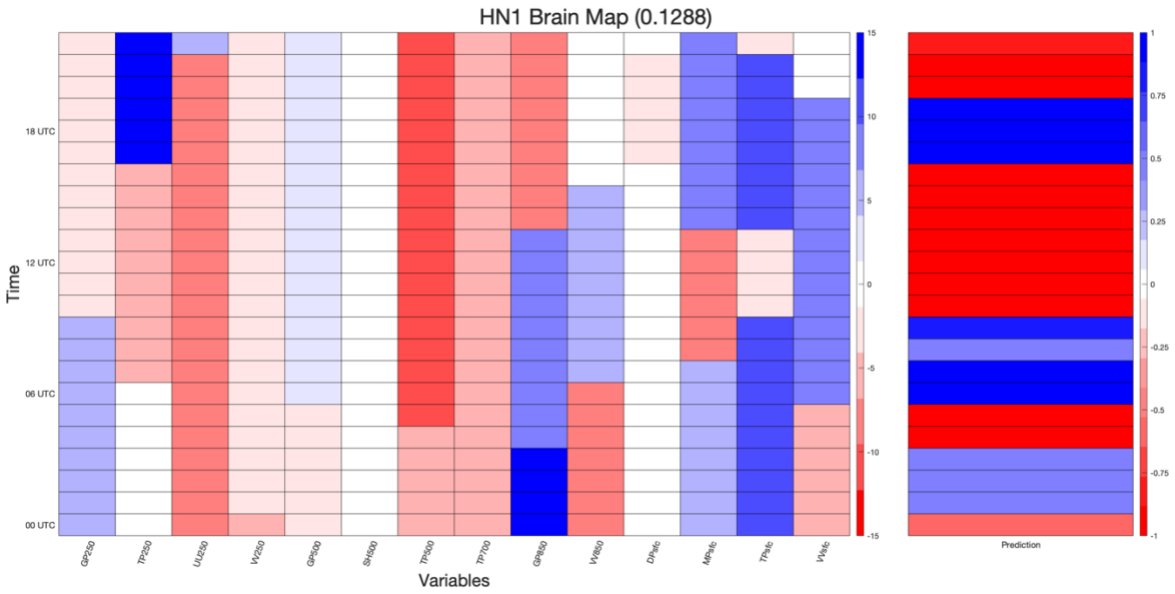

Figure 19. A map depicting the values of each variable within hidden node 1 (left) and value of the hyperbolic tangent activation function for hidden node 1 (right) for July 28, 2011, from 00 UTC (bottom) to 23 UTC (top). In the left diagram, shades of red are negative values and shades of blue are positive values. The scale of the color bar is determined by finding the heaviest weight and setting that as the max and min values. Note that the sign of the weight applied in the logit equation determines if the influence of a variable is to increase severe probability (negative) or decrease severe probability (positive). Hidden node weights can be determined in the Table 3. The right diagram is as it is in the left diagram in Fig. 17.

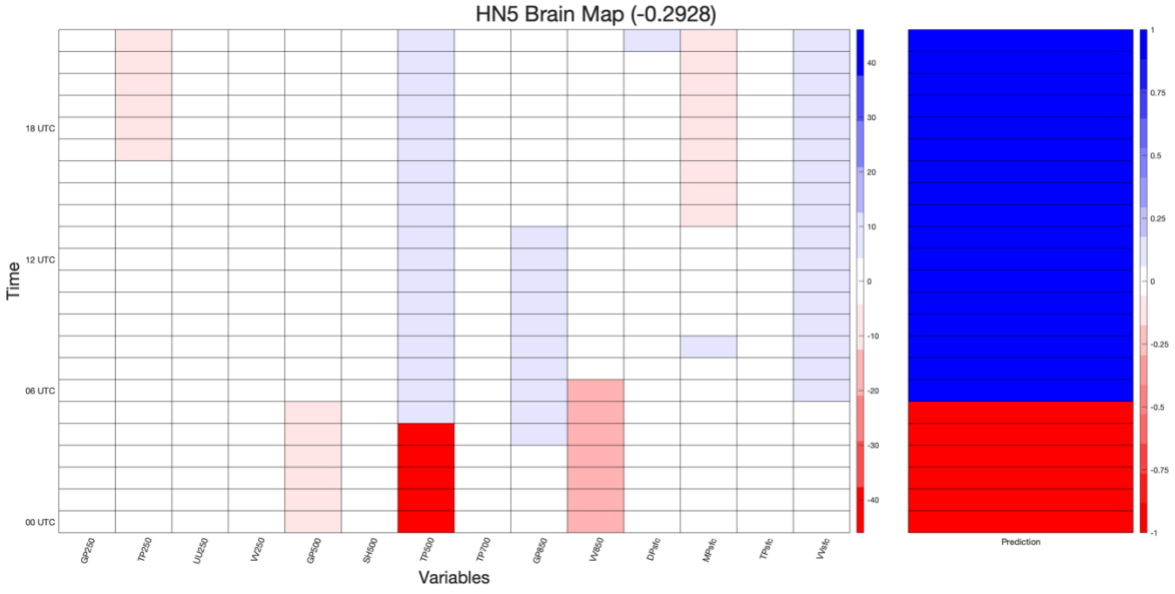

Figure 20. A map depicting the values of each variable within hidden node 5 (left) and value of the hyperbolic tangent activation function for hidden node 5 (right) for July 28, 2011, from 00 UTC (bottom) to 23 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

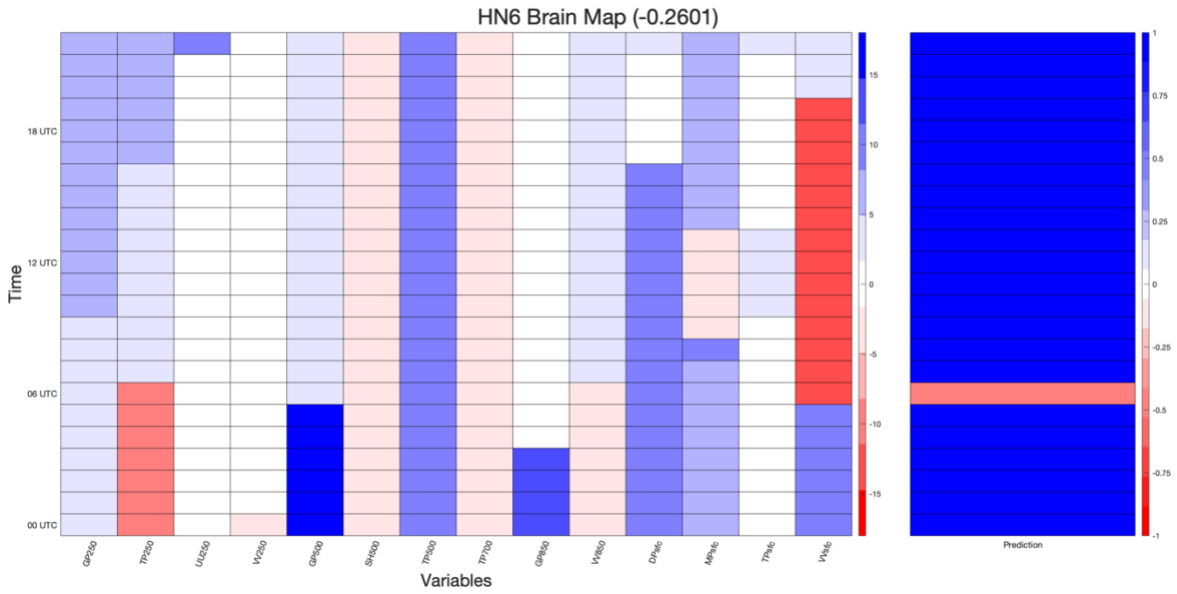

Figure 21. A map depicting the values of each variable within hidden node 6 (left) and value of the hyperbolic tangent activation function for hidden node 6 (right) for July 28, 2011, from 00 UTC (bottom) to 23 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

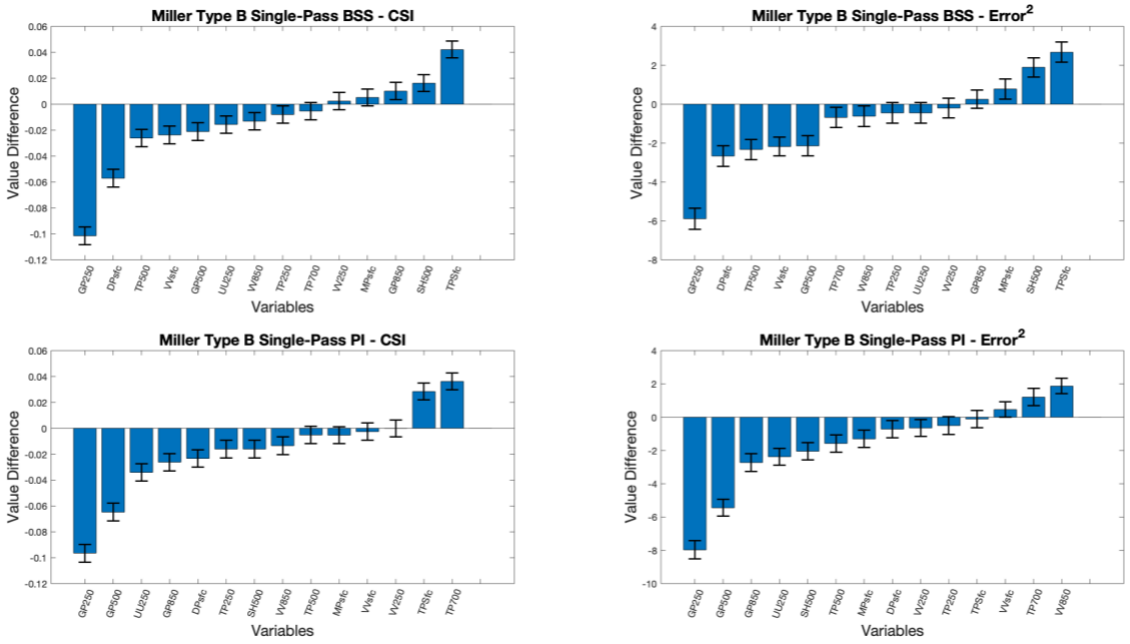

Figure 22. Comparison of single-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all Miller Type B synoptic settings. Values and error bars are the same as they are for Fig. 3.

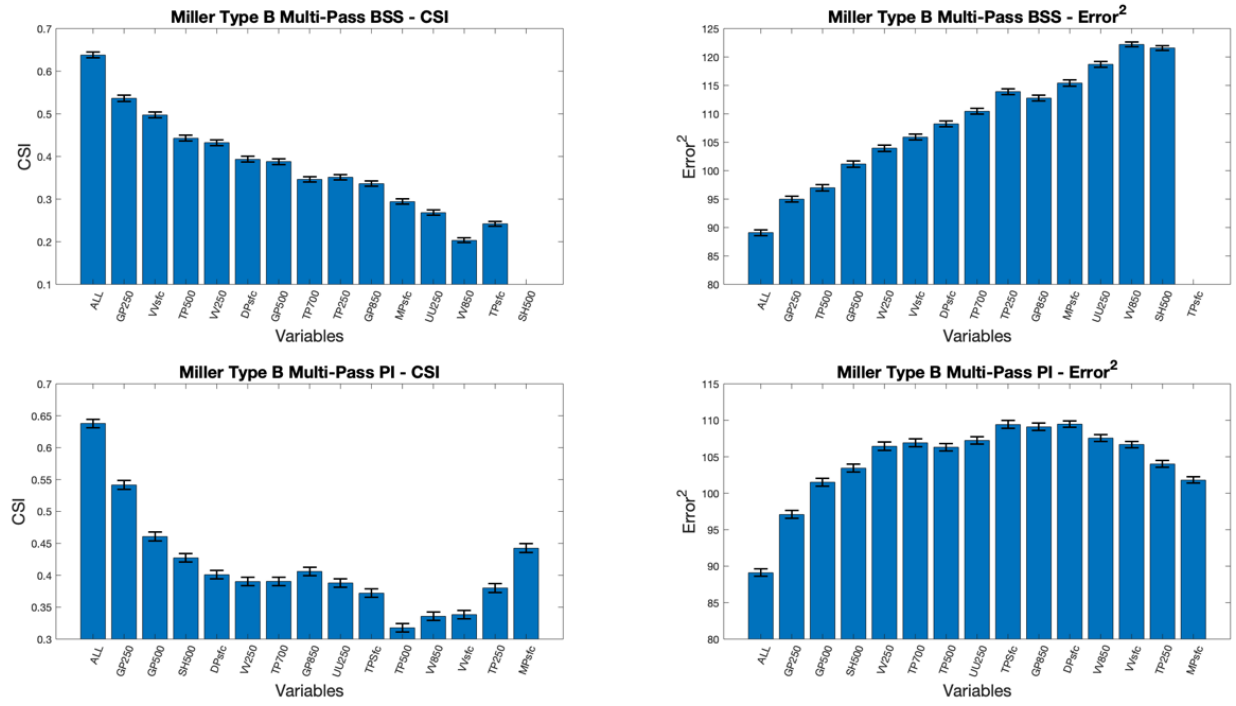

Figure 23. Comparison of multi-pass rankings of CSI (left) and squared error (right) for both permutation importance (bottom) and backward sequential selection (top) for all Miller Type B synoptic settings. Values and error bars are the same as they are for Fig. 5.

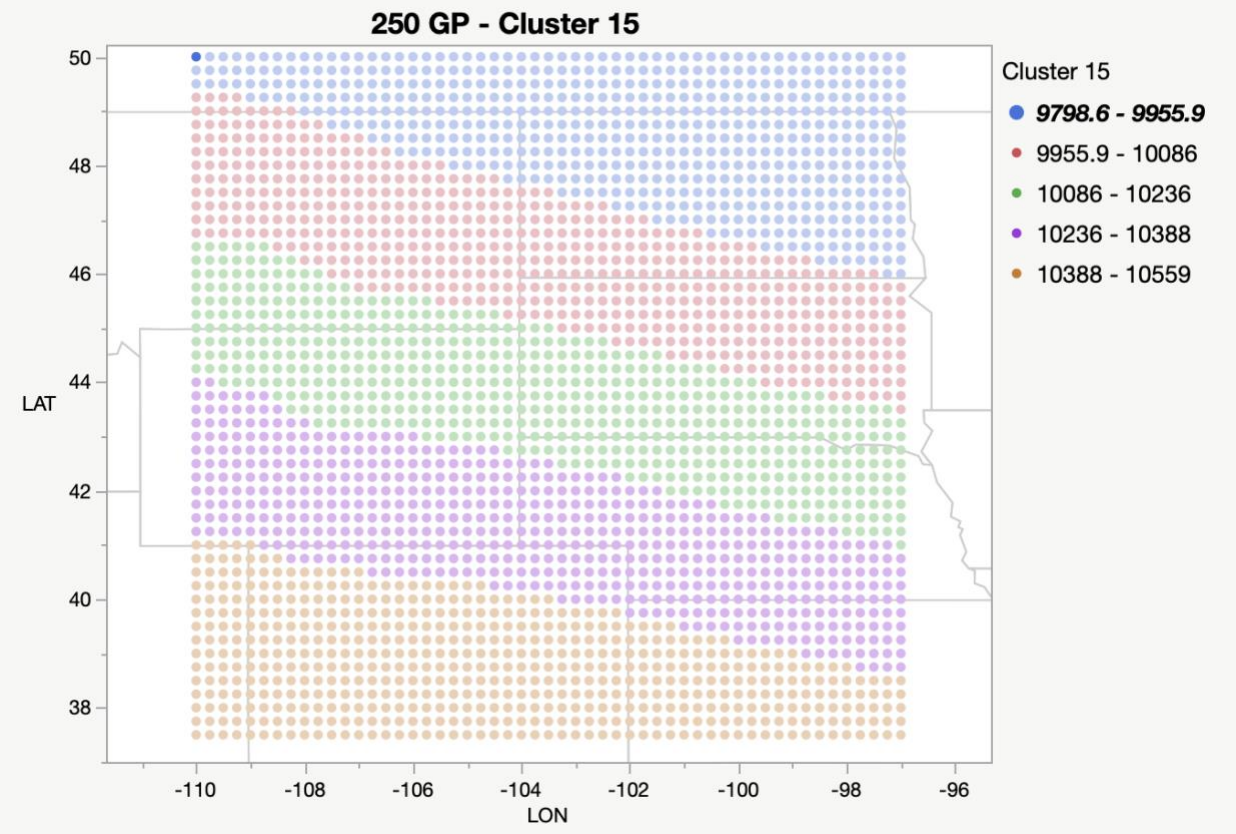

Figure 24. The composite image of 250hPa geopotential height fields associated with the cluster centroid 15. Height is measured in m.

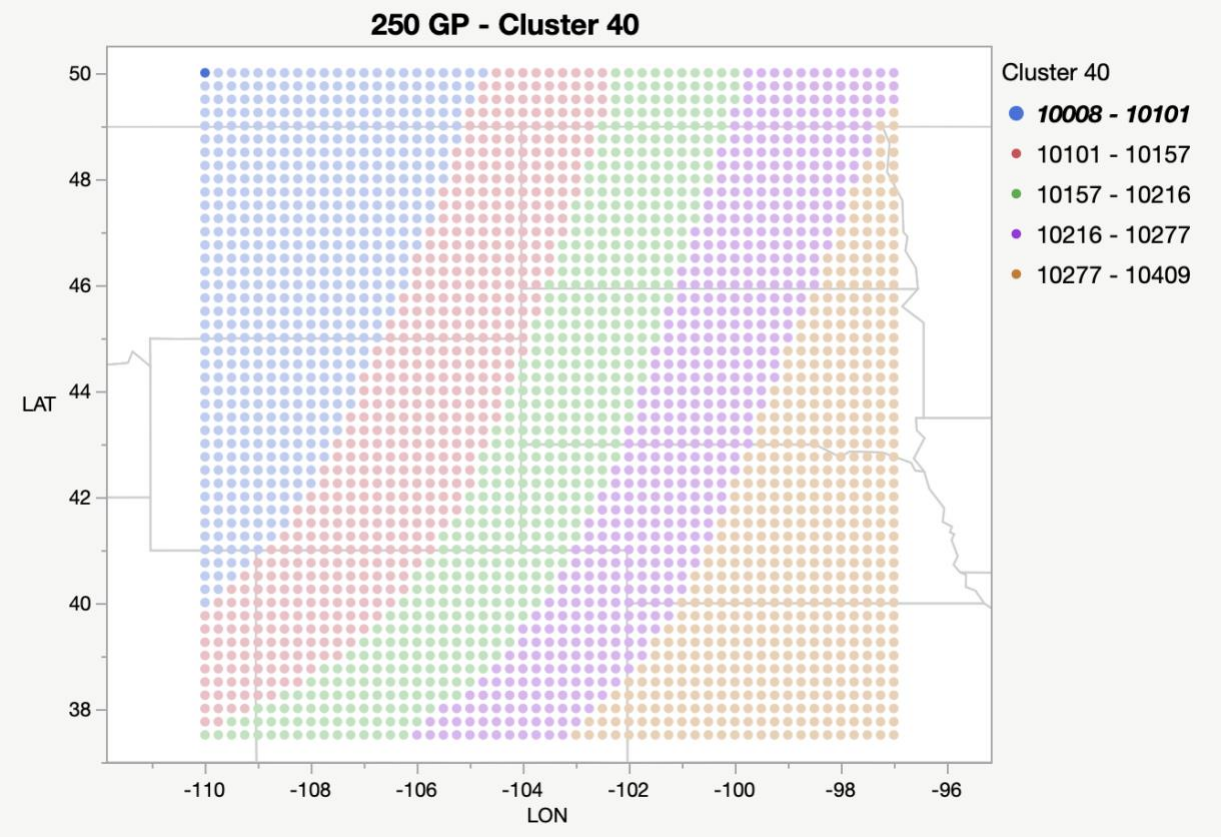

Figure 25. The composite image of 250hPa geopotential height fields associated with the cluster centroid 40. Height is measured in m.

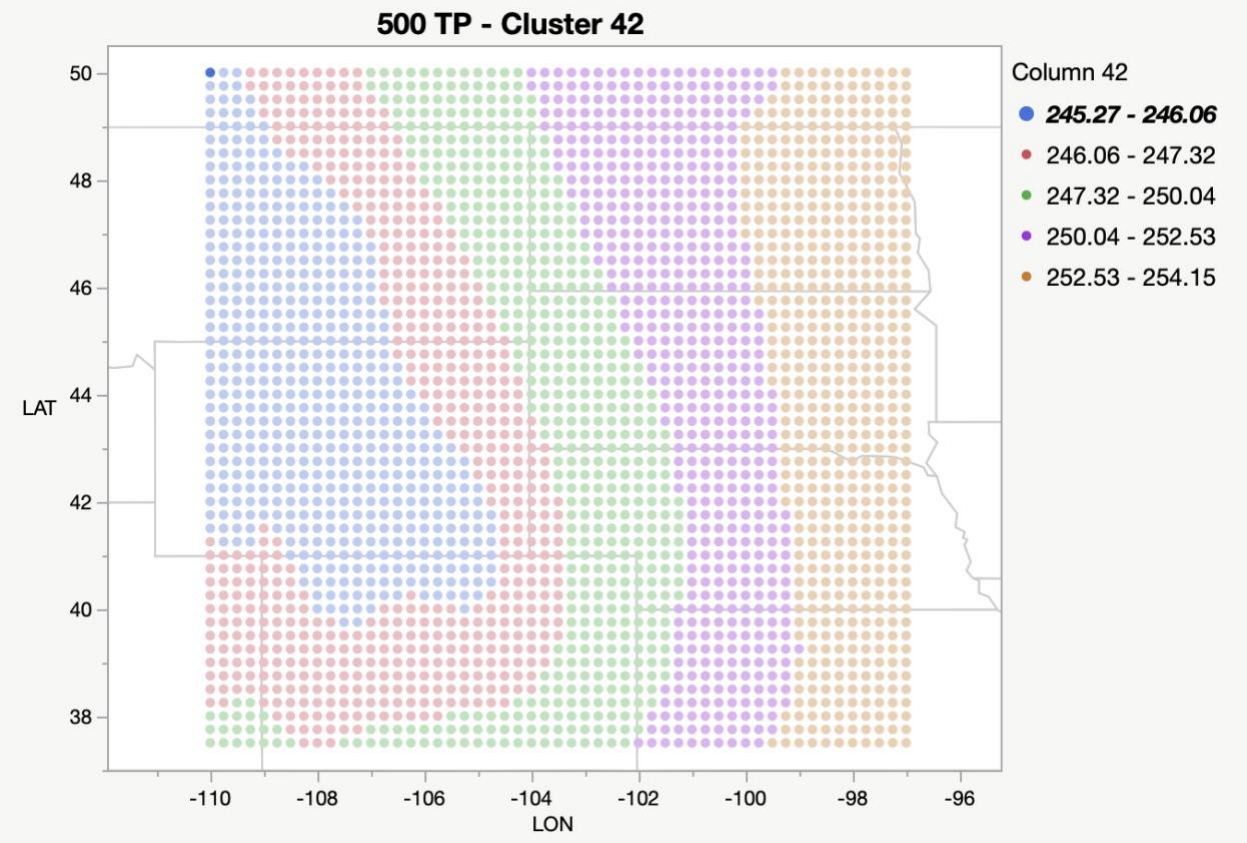

Figure 26. The composite image of 500hPa temperature fields associated with the cluster centroid 42. Temperature is measured in units kelvin (K).

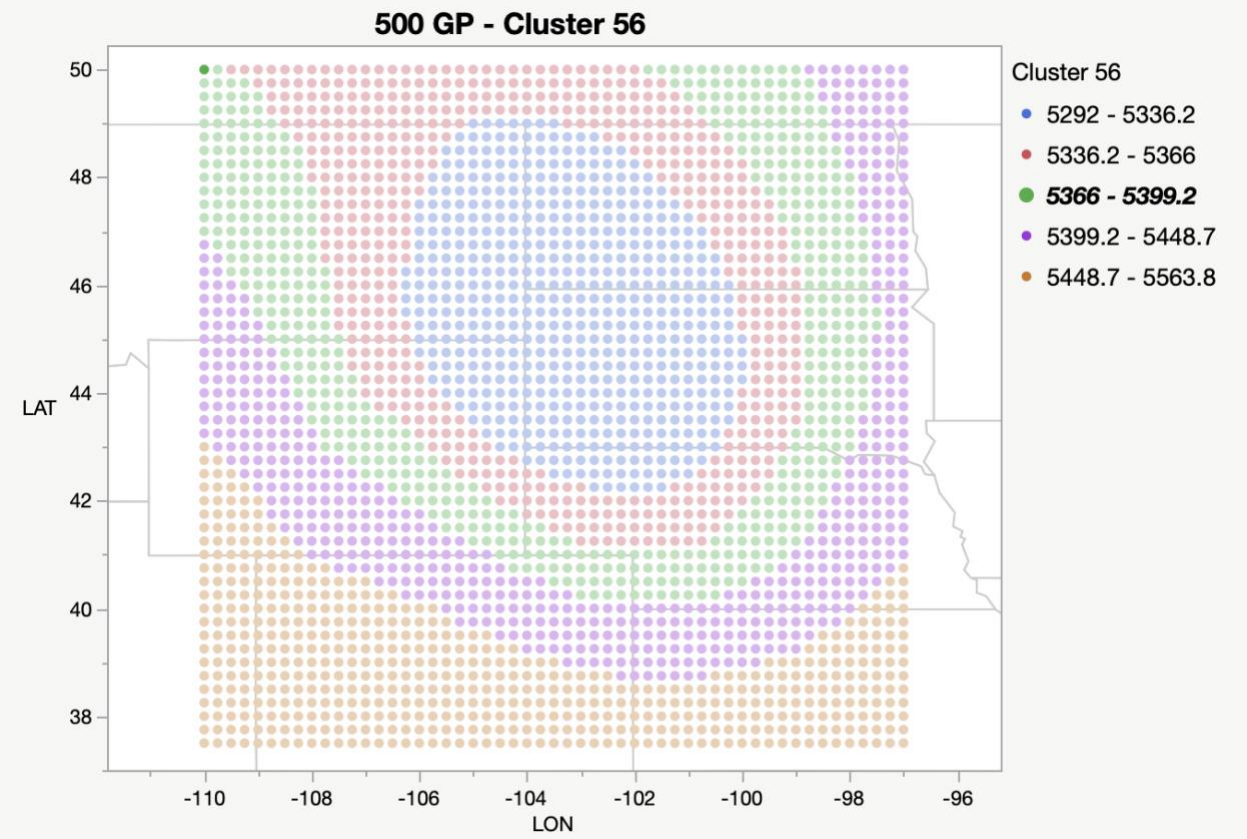

Figure 27. The composite image of 500hPa geopotential height fields associated with the cluster centroid 56. Height is measured in m.

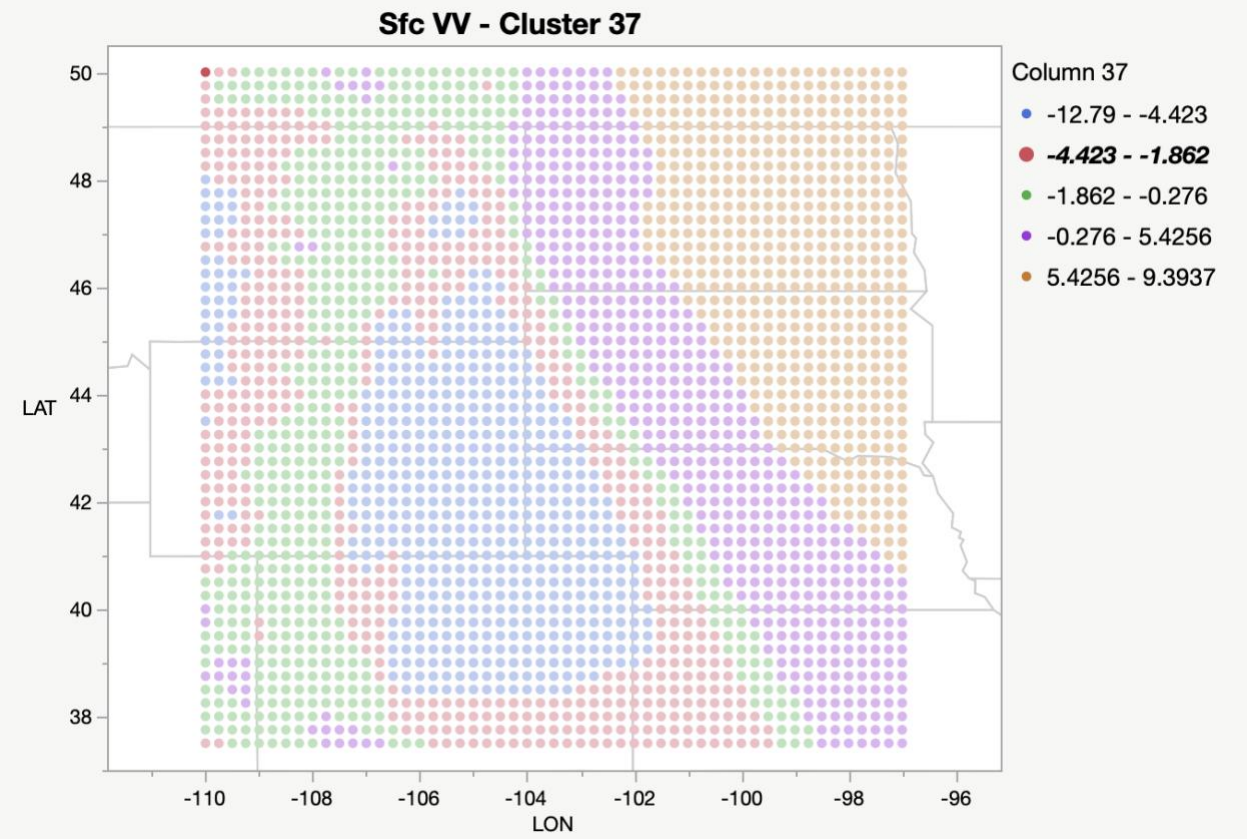

Figure 28. The composite image of surface v-wind fields associated with the cluster centroid 37. Positive values represent south to north flow and negative values represent north to south flow. Velocity is measured in ms^{-1}.

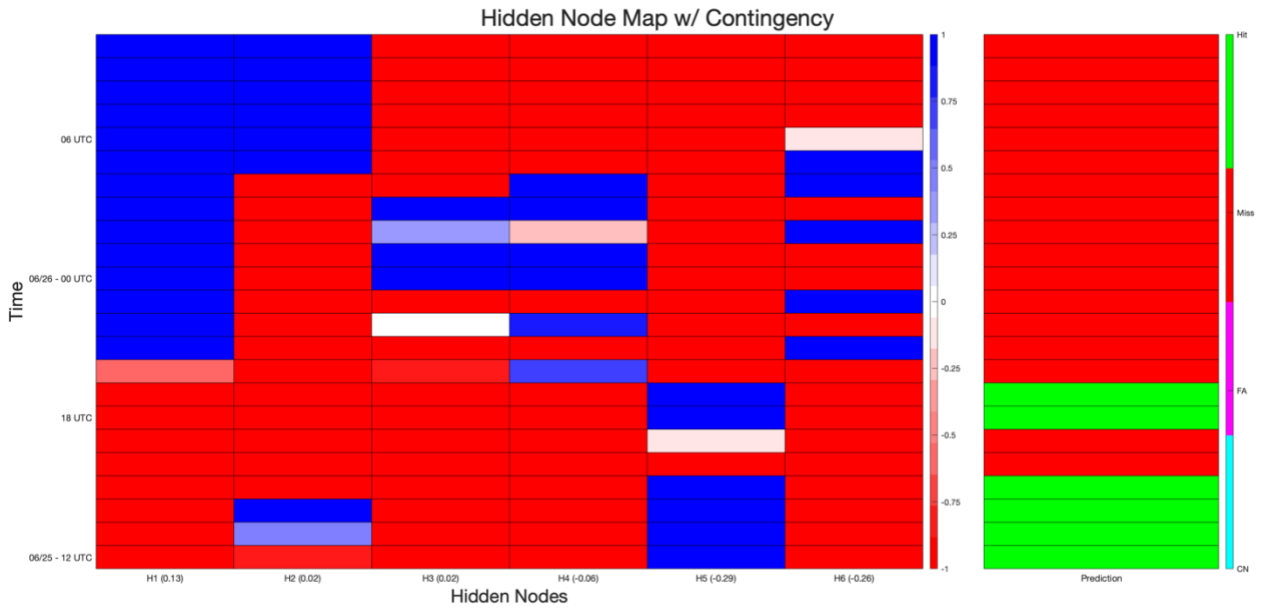

Figure 29. A map depicting the values of the hyperbolic tangent activation functions for each of the six hidden nodes (left) and the forecast contingency (right) from June 25th, 2011 at 12 UTC (bottom) to June 26th, 2011 at 12 UTC (top). Shades and values are as they are in Fig. 17.

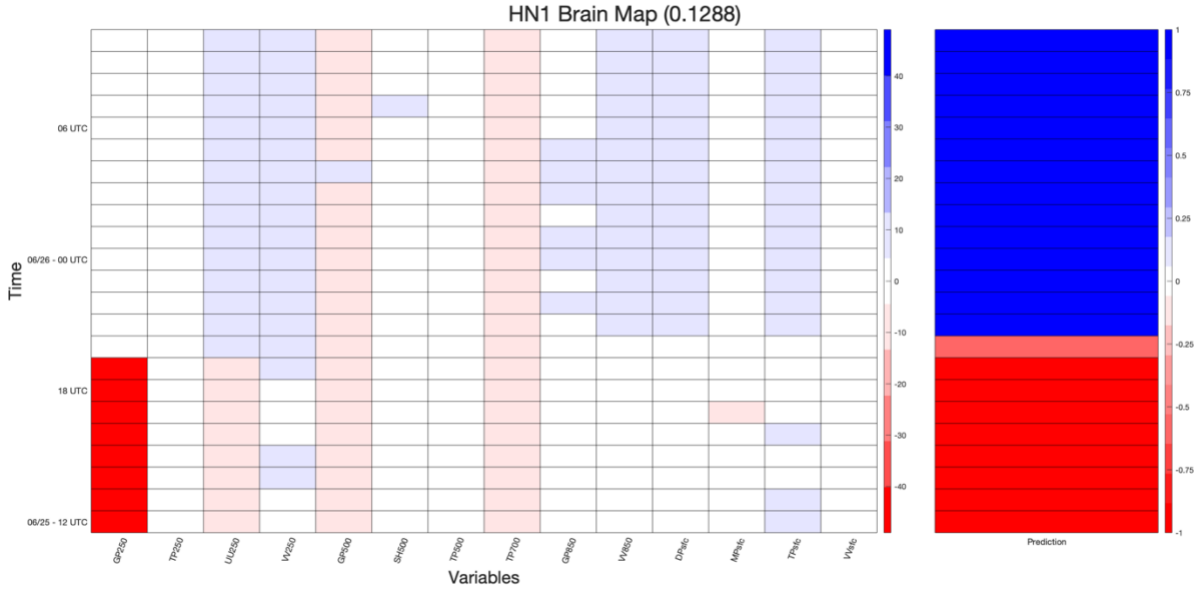

Figure 30. A map depicting the values of each variable within hidden node 1 (left) and value of the hyperbolic tangent activation function for hidden node 1 (right) from June 25th, 2011 at 12 UTC (bottom) to June 26th, 2011 at 12 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

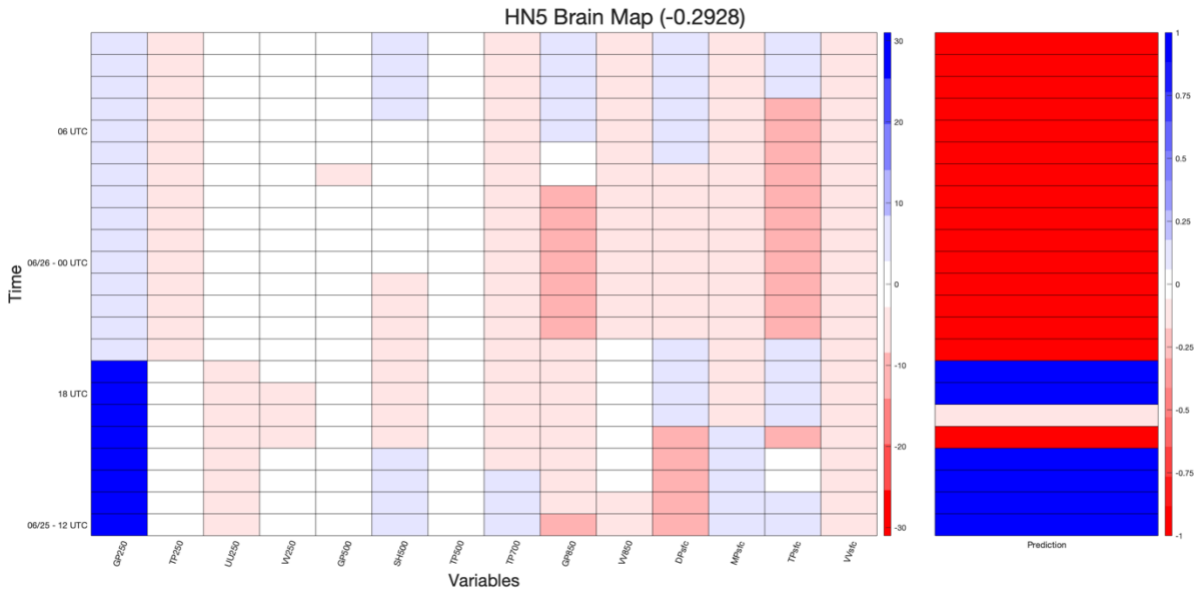

Figure 31. A map depicting the values of each variable within hidden node 5 (left) and value of the hyperbolic tangent activation function for hidden node 5 (right) from June 25th, 2011 at 12 UTC (bottom) to June 26th, 2011 at 12 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

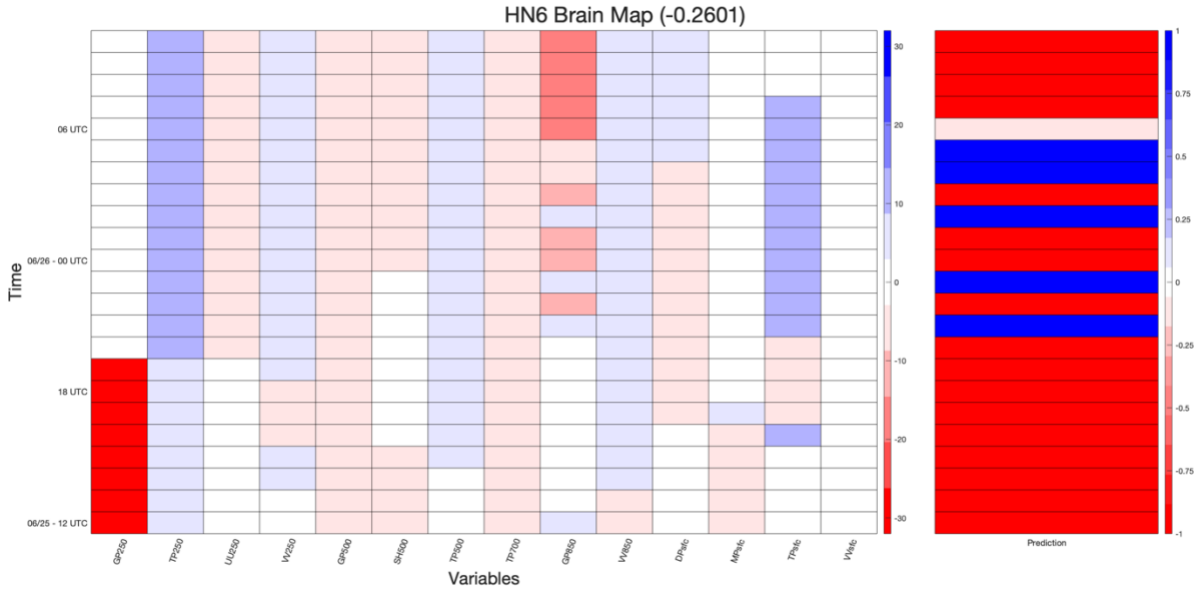

Figure 32. A map depicting the values of each variable within hidden node 6 (left) and value of the hyperbolic tangent activation function for hidden node 6 (right) from June 25th, 2011 at 12 UTC (bottom) to June 26th, 2011 at 12 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

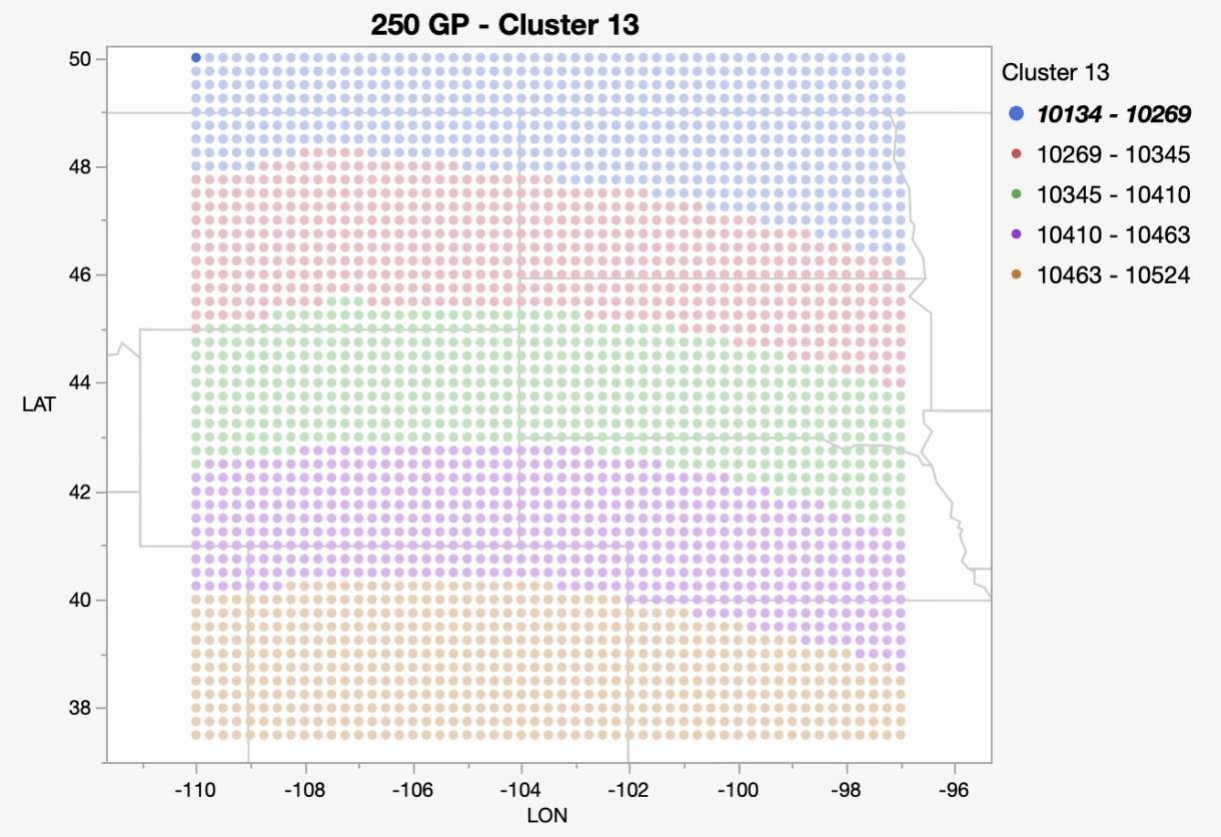

Figure 33. The composite image of 250hPa geopotential height fields associated with the cluster centroid 13. Height is measured in m.

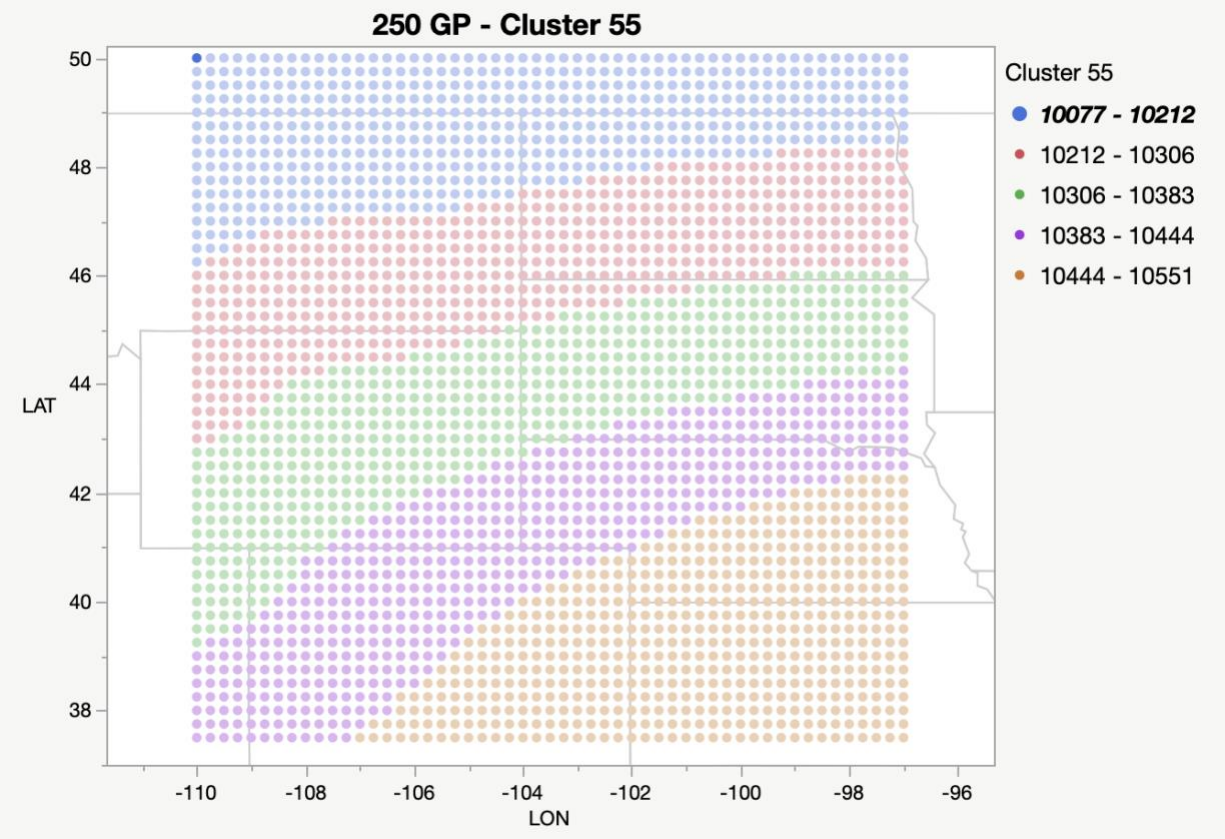

Figure 34. The composite image of 250hPa geopotential height fields associated with the cluster centroid 55. Height is measured in m.

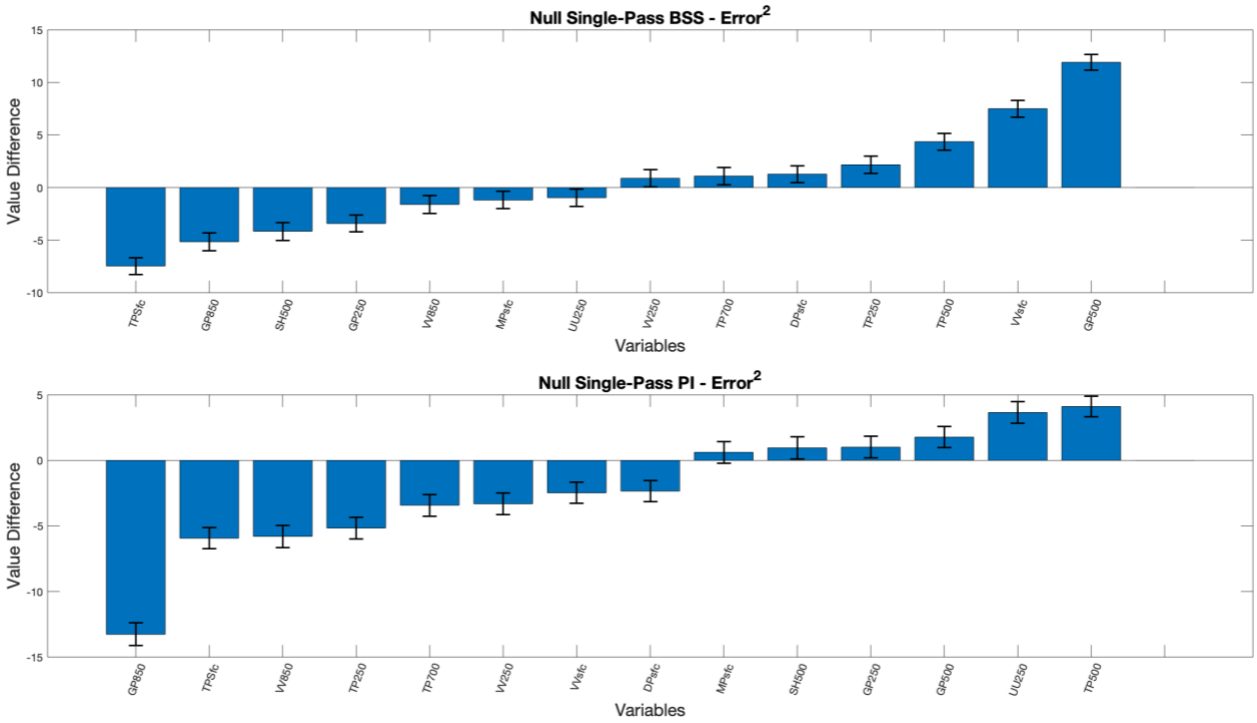

Figure 35. Comparison of single-pass rankings of squared error for both permutation importance (bottom) and backward sequential selection (top) for all null forecasts. Values and error bars are the same as they are for Fig. 3.

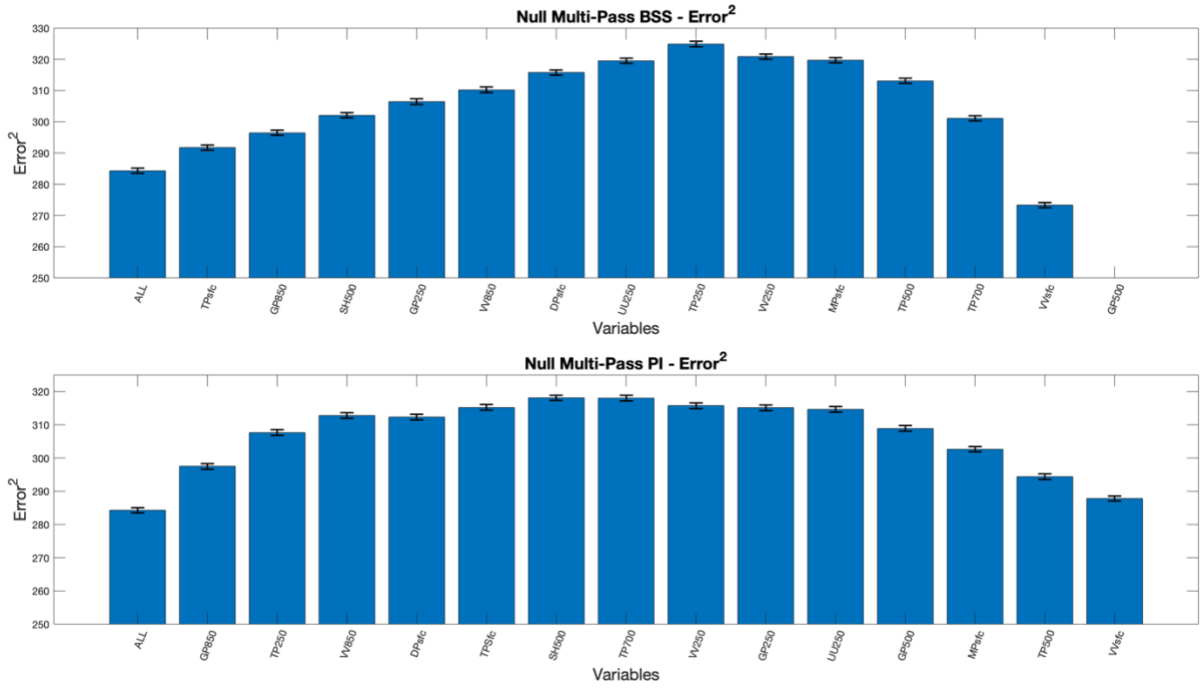

Figure 36. Comparison of multi-pass rankings of squared error for both permutation importance (bottom) and backward sequential selection (top) for all null forecast. Values and error bars are the same as they are for Fig. 5.

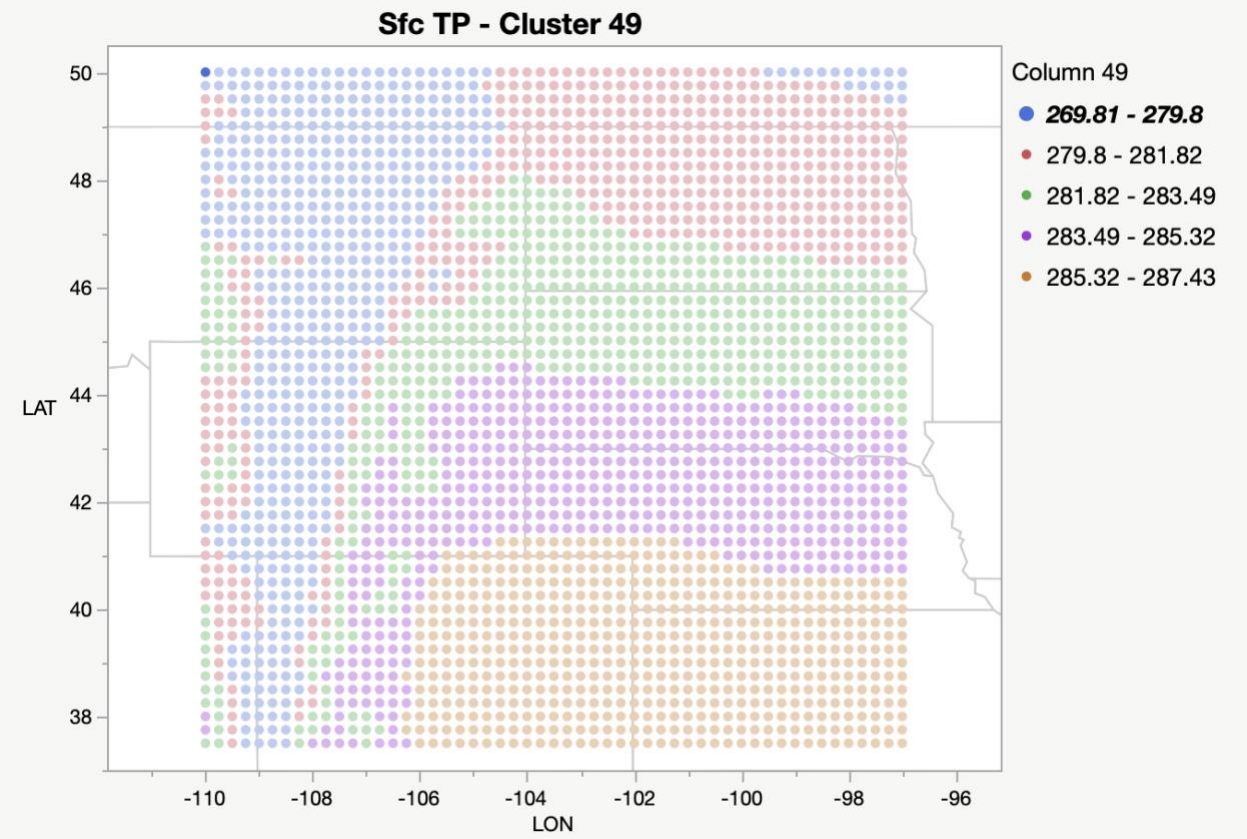

Figure 37. The composite image of surface temperature fields associated with the cluster centroid 49. Temperature is measured in units kelvin (K).

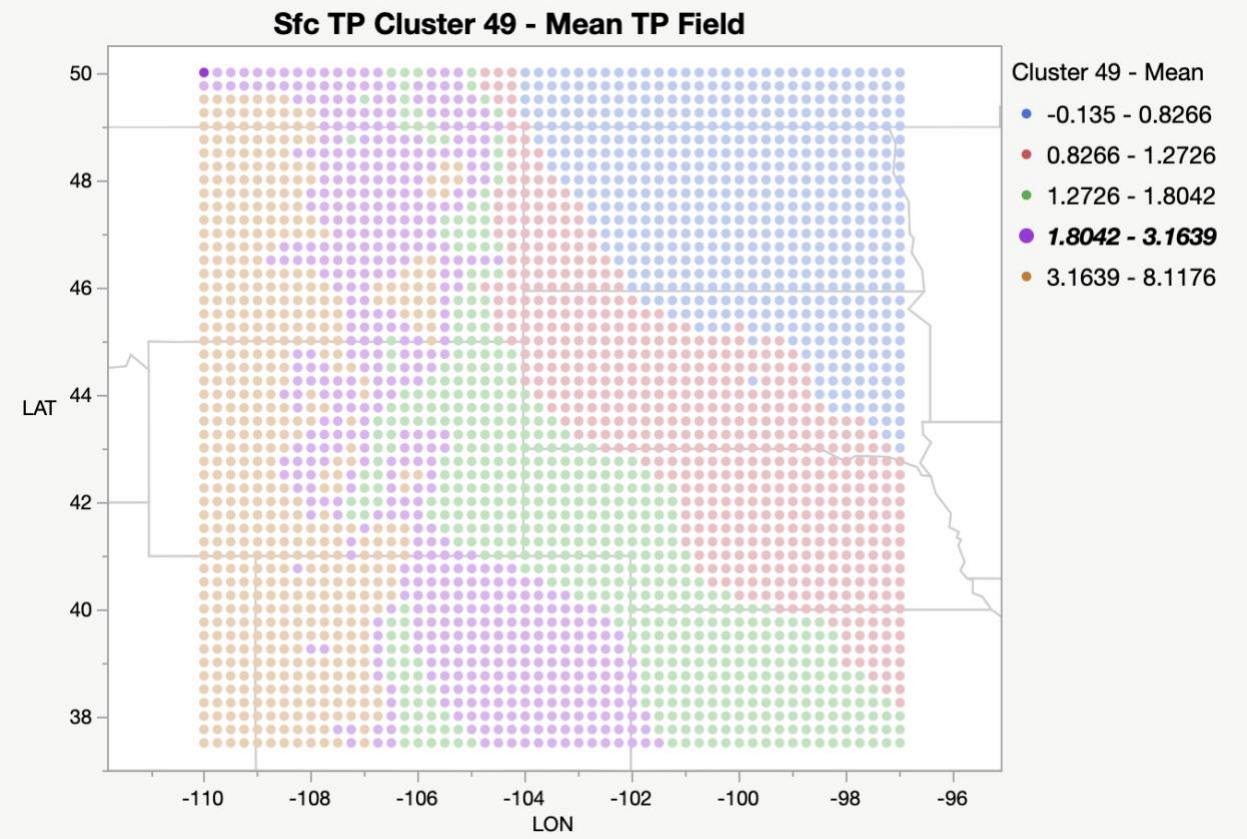

Figure 38. The difference between surface temperature fields associated with the cluster centroid 49 and the mean surface temperature field.

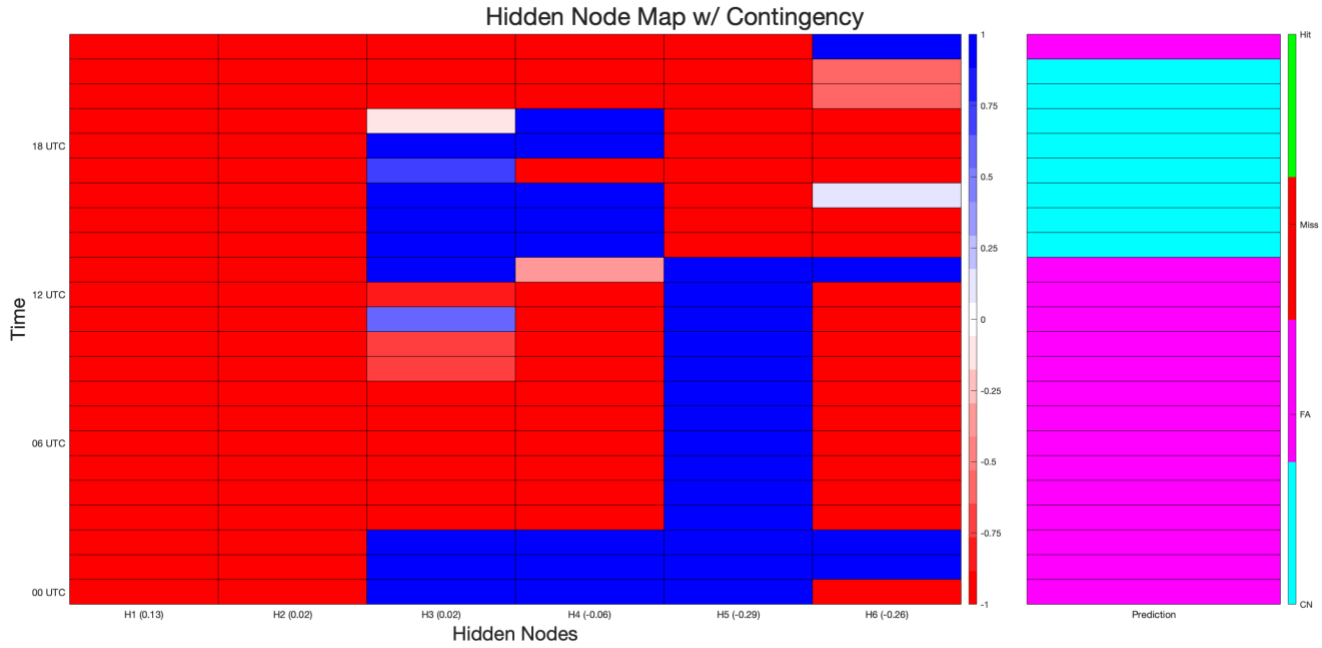

Figure 39. A map depicting the values of the hyperbolic tangent activation functions for each of the six hidden nodes (left) and the forecast contingency (right) for June 4, 2011, from 00 UTC (bottom) to 23 UTC (top). Shades and values are as they are in Fig. 17.

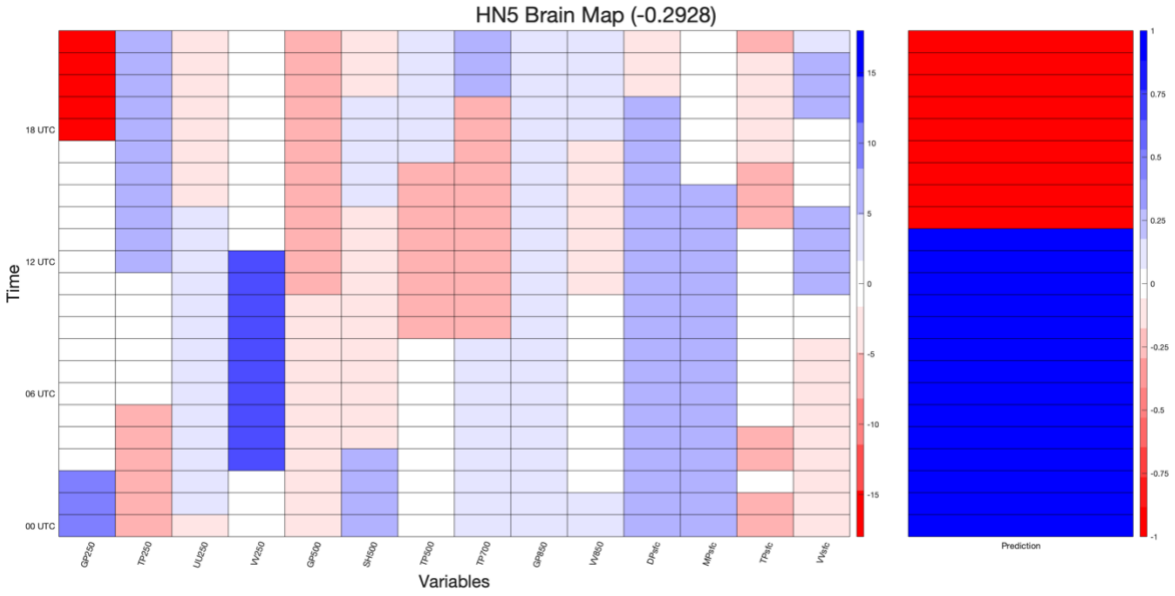

Figure 40. A map depicting the values of each variable within hidden node 5 (left) and value of the hyperbolic tangent activation function for hidden node 5 (right) for June 4, 2011, from 00 UTC (bottom) to 23 UTC (top). The colors of the left and right diagrams are as they are in Fig. 18, but the color scale is relative to this figure.

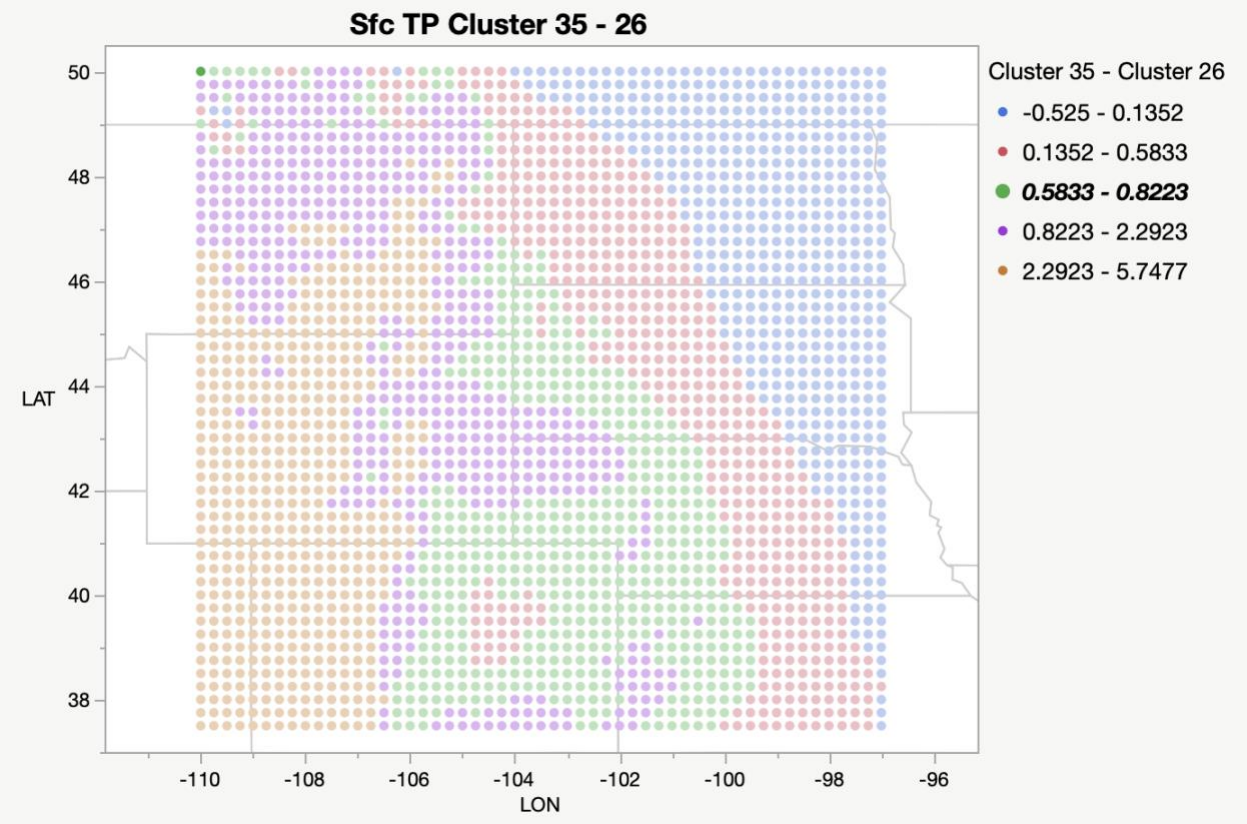

Figure 41. The difference between surface temperature fields associated with the cluster centroid 35 and cluster centroid 26.

VI. TABLES

Short Form ID	Long Form ID	Units
250GP	250hPa Geopotential Height	m
250TP	250hPa Temperature	K
250UU	250hPa Zonal Wind	m s^{-1}
250VV	250hPa Meridional Wind	m s^{-1}
500GP	500hPa Geopotential Height	m
500SH	500hPa Specific Humidity	kg kg^{-1}
500TP	500hPa Temperature	K
700TP	700hPa Temperature	K
850GP	850hPa Geopotential Height	m
850VV	850hPa Meridional Wind	m s^{-1}
SfcMP	Mean Sea Level Pressure	Pa
SfcTP	Surface Temperature	K
SfcDP	Surface Dew Point Temperature	K
SfcVV	Surface Meridional Wind	m s^{-1}

Table 1: Short- and long-hand variable identifiers and units of measurements. ERA5 data collected from https://cds.climate.copernicus.eu/cdsapp#!/home.

Miller Classification	Main Identifier
Type A	Along Dry Line
Type B	Ahead of Cold Front
Type C	Along Stationary Front
Type D	500hPa “Cold Core” Temperatures
Type E	Ahead of Warm Front

Table 2: Key features used to identify and classify Miller synoptic convective weather patterns. More details can be collected from https://www.weather.gov/media/zhu/ZHU_Training_Page/thunderstorm_stuff/thunderstorms_tutorial/Thunderstorms.pdf.

Hidden Node	Weight
Node 1	0.1288
Node 2	0.0218
Node 3	0.0204
Node 4	-0.0590
Node 5	-0.2928
Node 6	-0.2601

Table 3: Weights assigned to each hidden node.

VII. REFERENCES

Selvaraju, R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017: Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int. Conf. on Computer Vision, Venice, Italy, IEEE, https://doi.org/10.1109/ICCV.2017.74.
[https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf]

McGovern, A., R. Lagerquist, D. Gagne, G. Jergensen, K. Elmore, C. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. Bulletin of the American Meteorological Society, early online release, https://doi.org/10.1175/BAMS-D-18-0195.1

Molnar, C., 2018: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub, URL https://christophm.github.io/interpretable-ml-book/

Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell, 1, 206–215, https://doi.org/10.1038/s42256-019-0048-x

Ribeiro, M., S. Singh, and C. Guestrin, 2016: "Why should I trust you?": Explaining the predictions of any classifier. Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, ACM, https://doi.org/10.1145/2939672.2939778.

Krause, J., A. Perer, and N. Kenney, 2016: Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. Conf. on Human Factors in Comp. Sci., Chicago, IL, ACM, http://dx.doi.org/10.1145/2858036.2858529.

Craven, M. and J. Shavlik, 1995: Extracting tree-structured representations of trained networks. Proceedings of the Conf. on Adv. In Neural Infor. Proc. Sys., 24-30, http://dl.acm.org/doi/10.5555/2998828.2998832.

Breiman, L., 2001: Random forests. Mach. Learn., 45, 5-32, https://doi.org/10.1023/A:1010933404324.

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts, 2013: Understanding variable importances in forests of randomized trees. Conf. on Neural Information Processing Systems, Lake Tahoe, CA, Neural Information Processing Systems Foundation, http://dl.acm.org/doi/10.5555/2999611.2999660.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? J. Atmos Oceanic Technol., 32, 1209-1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, 2016: Learning deep features for discriminative localizations. Conf. on Computer Vision and Pattern Recognition, Las Vegas, Nevada, IEEE, https://doi.org/10.1109/CVPR.2016.319.

Bera, K., K. Schalper, D. Rimm, V. Velcheti, and A. Madabhushi, 2019: Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. Nat. Rev. Clin. Oncol., 16, 703-715, https://doi.org/10.1038/s41571-019-0252-y.

Hamilton, M., P. Hoang, L. Layne, J. Murray, D. Padgett, C. Stafford, and H. Tran, 2014: Applying machine learning techniques to baseball pitch prediction. Proceedings of the 3rd Inter. Conf. on Pat. Recog. Applications and Methods, Angers, France, ICPRRAM, 520-527, https://doi.org/10.5220/0004763905200527.

Camacho-Urriolagoitia, O., I. Lopez-Yanez, Y. Villuendas-Rey, O. Camacho-Nieto, and C. Yanez-Marquez, 2021: Dynamic Nearest Neighbor: An Improved Machine Learning Classifier and Its Application in Finances. Applied Sciences, 11, 8884, https://doi.org/10.3390/app11198884.

Wang, H., C. Ma, and L. Zhou, 2009: A brief review of machine learning and its applications. 2009 Inter. Conf. on Info. Engin. and Comp. Sci., Wuhan, China, IEEE, 1-4, https://doi.org/10.1109/ICIECS.2009.5362936.

Barredo, A., and Coauthors, 2020: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 58, 82-115, https://doi.org/10.1016/j.inffus.2019.12.012.

Miller, T., 2019: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence., 267, 1-28, https://doi.org/10.1016/j.artint.2018.07.007.

Likas, A., N. Vlassis, J. Verbeek, 2001: The global k-means clustering algorithm. Pattern recognition, 36, 451-461, https://doi.org/10.1016/S0031-3203(02)00060-2.

Gotwalt, C. 2011: "JMP Neural Network Methodology". SAS Institute, 11 pp.

Stracuzzi, D., and P. Utgoff, 2004: Randomized variable elimination. J. Mach. Learn. Res., 5, 1331-1362

Roebber, P. 2022: A review of artificial intelligence and machine learning activity across the United States National Weather Service. NOAA Technical Memorandum. NWS MDL 86, 25 pp.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. Q. J. R. Meteo. Soc. 146, 1999-2049, https://doi.org/10.1002/qj.3803.

Sarle, W. 1983: Cubic Clustering Criterion. SAS Institute. SAS Technical Report. 108, 56, https://support.sas.com/kb/22/addl/fusion_22540_1_a108_5903.pdf.

Mamalakis, A., I. Ebert-Uphoff, and E. Barnes, 2022: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. Environmental Data Science, 1, E8, https://doi.org/10.1017/eds.2022.7