CORRECTION OF BACK TRAJECTORIES UTILIZING MACHINE LEARNING

THESIS

Britta F. Gjermo Morrison, Captain, USAF

AFIT-ENP-MS-21-M-119

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENP-MS-21-M-119

CORRECTION OF BACK TRAJECTORIES UTILIZING MACHINE LEARNING

THESIS

Presented to the Faculty

Department of Engineering Physics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Atmospheric Science

Britta F. Gjermo Morrison, B.S.

Captain, USAF

March 2021

AFIT-ENP-MS-21-M-119

CORRECTION OF BACK TRAJECTORIES UTILIZING MACHINE LEARNING

Britta F. Gjermo Morrison, B.S.

Captain, USAF

Committee Membership:

Lt Col Robert C. Tournay, PhD
Chair

Maj Peter A. Saunders, PhD
Member

Dr. Astrid Suarez-Mullins
Member

AFIT-ENP-MS-21-M-119

Abstract

The goal of this work was to analyze 24-hour back trajectory performance from a global, low-resolution weather model compared to a high-resolution limited area weather model in particular meteorological regimes, or flow patterns using K-means clustering, an unsupervised machine learning technique. The global model tested was the National Centers for Environmental Prediction's (NCEP) Climate Forecast System version 2 (CFS), and the high-resolution model was NCEP's Rapid Refresh (RAP) numerical weather model. The duration of this study was from 2015-2019 for the contiguous United States (CONUS).

Three different machine learning algorithms were tested to study the utility of these methods improving the performance of the CFS relative to the performance of the RAP. The aforementioned machine learning techniques are linear regression, Bayesian ridge regression, and random forest regression. There was no statistical significance in the CFS improvement values between the three different algorithms test nor in the sum of the results comparing unclustered and clustered trajectories indicating the simpler machine learning methods could be used and using k-means clustering was not necessary. These results mean reducing computational time for the user. Additionally, the greatest improvement of CFS values occurred in July, August, and September when back trajectories did not travel as far in 24-hours compared to other seasonal atmospheric regimes.

Acknowledgments

First and foremost, I would like to thank my husband for being the ultimate AFIT hype man during my time here. While we were not stationed here together, I could not have completed this program without him and his never-ending encouragement and support. Second, I would like to give a big thank you to my advisor for all the assistance and guidance he provided during my time here. Additionally, thank you to my committee for their support and feedback during research and writing phases of my thesis. Finally, a huge shout-out and thank you to my classmates and friends here at AFIT for being a sounding board and sanity check throughout my entire program. I would have not gotten here without them.

Britta F. Gjermo Morrison

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMJ April, May, and June

CBRN chemical, biological, radiological, and nuclear

CFS Climate Forecast System numerical weather model

CONUS contiguous United States

CTBT Comprehensive Nuclear-Test-Ban Treaty

DoD Department of Defense

ELM extreme machine learning

EPA United States Environmental Protection Agency

ERA-40 European Centre for Medium-Range Weather Forecasts 40-year

HYSPLIT Hybrid Single-Particle Lagrangian Integrated Trajectory model

JAS July, August, and September

JFM January, February, and March

MDEQ Mississippi Department of Environmental Quality

METREX Metropolitan Tracer Experiment

MLP multi-layer perception

MM5 fifth generation Mesoscale Model

NCEP National Centers for Environmental Prediction

NOAA National Oceanic and Atmospheric Administration

NWP numerical weather prediction

OND October, November, and December

PFJ polar front jet

RAP Rapid Refresh numerical weather model

RUC Rapid Update Cycle numerical weather model

RF random forest

RNN recurrent neural networks

STE source term estimation

WRF Weather Research and Forecasting model

WRF-CMAQ WRF coupled with Community Multi-scale Air Quality

CORRECTION OF BACK TRAJECTORIES UTILIZING MACHINE LEARNING

I. Introduction

Having accurate weather data to calculate the origin of atmospheric pollutants or other materials is critical across many disciplines from the agricultural sector, environmental and public health impacts to national defense applications. An example of a national defense application is being able to detect and enhance situational awareness if there is an intentional or accidental release of chemical, biological, radiological and nuclear (CBRN) materials (Bieringer et al. 2017, 103). "The interpretation of dispersion model plumes, their interaction with the local environment, and the appropriateness of regional model guidance are starting to become part of the routine duties of the local meteorological forecaster" (Draxler 2006, 384). The method for calculating these plume trajectories is called source term estimation (STE) and it requires three variables: (1) the time of the pollutant release, (2) the location of the pollutant release, and (3) the amount of pollutant released (Zoellick 2019, 1).

STE calculation methods can be categorized into three groups: (1) direct forward simulations, (2) direct inverse modeling, and (3) nonlinear optimization based methods (Bieringer et al. 2017, 103). The method this study focuses on falls under the direct inverse modeling category. Specifically back trajectories in which the calculation starts at the observation location and uses backward parcel trajectories to determine the air parcel source (Bieringer et al. 2017, 105). However, STE calculations can be very difficult to accurately calculate because observations of material concentration and wind

data are usually sparse, and the chaotic turbulent wind field makes it difficult to recreate in models (Bieringer et al. 2017, 102).

Another factor that affects the accuracy of STE calculations is the access to computation resources to run mesoscale meteorological models (Draxler 2006, 384). The higher resolution, or mesoscale, meteorological models are required to resolve the local environment and terrain features. Dispersion models used to calculate the STE algorithms are highly dependent on the method used to feed in the meteorological data, whether it is derived from observations or model data (Draxler 2006, 384). The most common dispersion model used to calculate back trajectories is the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) (Stein et al. 2015, 2059).

The inspiration and scientific foundation for HYSPLIT started back in 1949 when United States Weather Bureau forecasters were charged with finding the source from radioactive debris originating from the first Soviet atomic test (Stein et al. 2015, 2061). At the time, back trajectories could only be calculated by hand using wind data from twice-daily radiosonde balloon measurements (Stein et al. 2015, 2061). The first version of HYSPLIT debuted in the 1980s and since then four more versions have been released with the latest HYSPLIT version going operational in 2020 (Stein et al. 2015, 2062; NOAA-ARL n.d.). This study used trajectory code adapted from HYSPLIT into Python.

The objective of this research was to analyze 24-hour back trajectory performance from a global (low-resolution) weather model compared to a regional (high-resolution) weather model in particular meteorological regimes, or flow patterns using K-means clustering, an unsupervised machine learning technique. The next step was to test three different supervised machine learning algorithms on the global weather model to study

the utility these algorithms have on improving the global model's performance relative to the regional model. Chapter II provides brief background descriptions on atmospheric transport and dispersion modeling, atmospheric regimes, the global and regional weather models, clustering and K-means, and the supervised machine learning correction methods. Chapter III explains the methodology and Chapter IV presents the results and analyses. Lastly Chapter V concludes the research by discussing the potential impacts of this research, limitations, and future research ideas.

II. Background

2.1 Atmospheric Transport and Dispersion Modeling

Having an accurate understanding a pollutant plume's dispersion is critical to identifying its source location and concentration. The method of obtaining this information is called source term estimation (STE) and it depends on three factors: (1) the time of the pollutant release, (2) the location of the pollutant release, and (3) the amount of pollutant released (Zoellick 2019, 1). "The applications for which this need arises span a broad set of technical disciplines and can range from agricultural, industrial, environmental, public health, transportation, to defense applications" (Bieringer et al. 2017, 103). When it comes to defense applications, STE is used for determining compliance with the Comprehensive Nuclear-Test-Ban Treaty (CTBT), which bans nuclear explosions and tests conducted in any environment (atmosphere, under water, and underground) (Bieringer et al. 2017, 104). To ensure compliance with this treaty with participating countries, atmospheric modeling and STE methods are used to identify sources and the characteristics of specific gases that are a by-product to nuclear fission (Bieringer et al. 2017, 104). These methods were applied to detect nuclear testing done by non-signing countries such as Democratic People's Republic of Korea (Bieringer et al. 2017, 104).

Bieringer et al. (2017) reviewed several different STE methods to address the difficulties surrounding picking an STE method, highlight commonalities in these different methods, and share lessons learned in order to help those pick an appropriate STE method when facing a new STE problem (Bieringer et al. 2017, 102). Bieringer and

his team point out "…existing surveys focus on the differences between the methods, but do not examine them in the context of their similarities and how this relates to their utilization for the end application" (Bieringer et al. 2017, 103). The paper divides the STE methods into three categories: (1) direct forward simulation methods, (2) inverse and back-tracking methods, and (3) nonlinear optimization methods (Bieringer et al. 2017, 102).

The first category is used when both the source and receptor (i.e. the observation) locations are known (Bieringer et al. 2017, 104). This STE method is used to determine the release amount or rate (Bieringer et al. 2017, 104). Strengths for using the direct forward simulation method includes its simplicity, relatively low computational requirements, and ease of implementation (Bieringer et al. 2017, 104). Additionally, this method is most effective in scenarios where the meteorological conditions are very well known (Bieringer et al. 2017, 104).

The second category, inverse and backtracking, is split into two subcategories. The first subcategory is back trajectories "…used in the meteorological community to study long-range pollution transport and determine likely source locations" (Bieringer et al. 2017, 105) (see Figure 1). Similar to the first category, the strengths of this method include its simplicity, low computational requirements, and ease of implementation (Bieringer et al. 2017, 105). However this back trajectory method is highly sensitive to the accuracy of the wind and "…not appropriate for estimation of release rates or precise source locations since this approach typically does not model atmospheric effects beyond parcel transport" (Bieringer et al. 2017, 105).

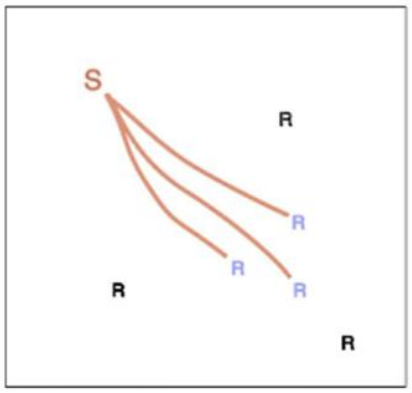

Figure 1. A visualization of the back trajectory subcategory. Receptors (R) that did not detect a contaminant are in black, light blue R's detected a contaminant, the trajectories of the contaminant are the orange lines, and the orange S is the source of the contaminant (Bieringer et al., 2017).

The second subcategory uses STE methods that give information on the evolution of the contaminant through transport and dispersion, as shown in Figure 2 (Bieringer et al. 2017, 105). "If the source location is known, this method can also be used to determine the release mass, timing, and/or release rate for that location" (Bieringer et al. 2017, 105). This method is adapted into many United States modeling applications such as Second order Closure Integrated Puff and HYSPLIT atmospheric transport and dispersion modeling systems (Bieringer et al. 2017, 106). These applications are used by the Department of Defense (DoD) and the National Oceanic and Atmospheric Administration. (NOAA) (Bieringer et al. 2017, 106). Even though this method is more capable than a simple back trajectory, it is still sensitive to meteorological inaccuracies and requires a high computational cost when there are a large number of concentrations observations (Bieringer et al. 2017, 106).

The third and final category is the nonlinear optimization, which uses a forward model of transport and dispersion (Bieringer et al. 2017, 111). "These iterative methods

implement a common set of steps: specify an initial population of trial solutions, compute the skill metric for each trial solution, modify the high-skill trial solutions, repeat until convergence" (Bieringer et al. 2017, 111). This method works by running the model forward in time repeatedly in order to maximize the skill of the model and minimizing the error (Bieringer et al. 2017, 107).

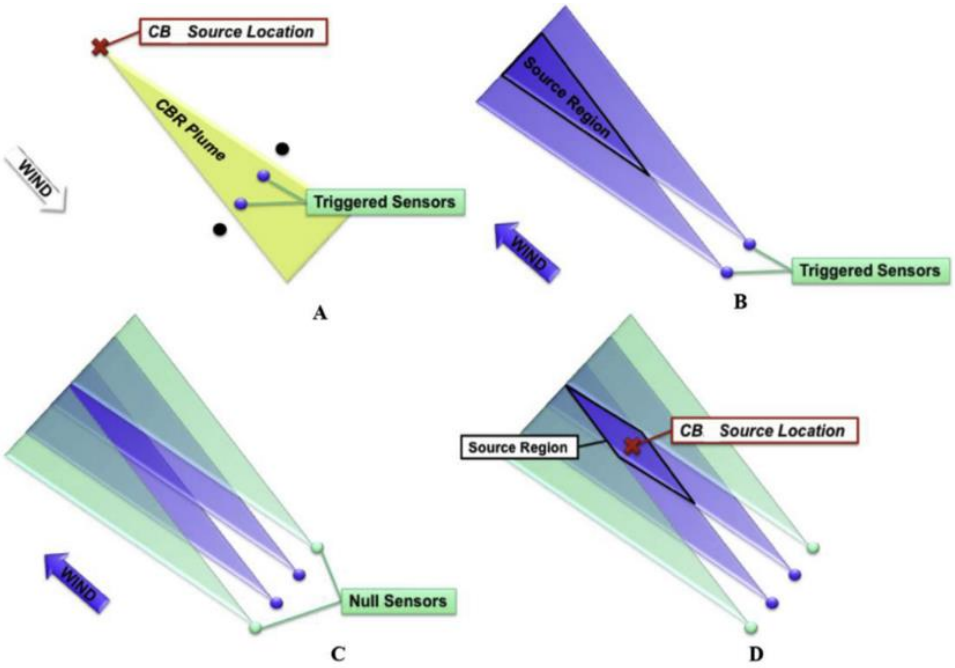

Figure 2. A visualization of the second STE subcategory. (A) a chemical is released in which a plume crosses a network of sensors; (B) shows the first step of the STE process where the winds are reversed and contaminant releases are made from the triggered sensor locations; (C) shows the STE process where again the winds are reversed and contaminant releases are made from the null sensor locations; and (D) illustrates the final step in the process where the inverse contaminant release information from the triggered and null sensors are combined to triangulate the source parameters (Bieringer et al., 2017).

One of the most effective and extensively used tools for modeling atmospheric transport and dispersion is HYSPLIT (Stein et al. 2015, 2059). The inspiration and foundation to create a trajectory program such as HYSPLIT started 1949 when the

Special Project Section of the U.S. Weather Bureau (now NOAA's National Weather Service) "…was charged with trying to find the source of radioactive debris originating from the first Soviet atomic test and detected by a reconnaissance aircraft near the Kamchatka Peninsula" (Stein et al. 2015, 2061). Over the last several decades, the HYSPLIT model has expanded from a simple single trajectory based on radiosonde observations to a model that can calculate how a parcel can be transport, dispersed, and deposited from local to global scales for several pollutants (Stein et al. 2015, 2059–61).

The HYSPLIT computation is made up of three elements: (1) particle transport by the mean wind, (2) a turbulent transport component, and (3) the computation of air concentration (Draxler 2006, 385). When calculating trajectories, the model uses a hybrid between the Lagrangian method, where your frame of reference (or grid) for calculating advection and diffusion moves with the fluid (or air parcel), and the Eulerian method, where your frame of reference (or grid) is stationary while the fluid (or air parcel) moves through it (Stein et al. 2015, 2059). Examples of how the HYSPLIT model is used are tracking and forecasting the release of radioactive material, wildfire smoke, wind-blown dust, pollutants from various stationary and mobile emission sources, allergens, and volcanic ash (Stein et al. 2015, 2059).

2.2 Atmospheric Regimes

Atmospheric regimes are generally defined as periodic large-scale spatial atmospheric structure and are a simple way to characterize atmospheric variability over a given region (Vrac and Yiou 2010, 1). A regime can persist over an area from one to several weeks (Michel and Rivière 2011, 1730). There are subjective and objective ways

to identifying different atmospheric regimes. Using a subjective method requires a significant understanding of meteorological knowledge over the area of interest, whereas an objective method can rely on "clustering methods to automatically group together large-scale atmospheric fields that are close to each other, and to create different clusters for fields that are very different from each other" (Vrac and Yiou 2010, 1).

2.3 Model Data

Since the purpose of this study was to the compare the output of trajectories from a global numerical weather prediction (NWP) model and a high-resolution, limited area NWP model, this section will give background on the two NWP models being used. The global model that will be tested is the National Centers for Environmental Prediction's (NCEP) Climate Forecast System version 2 (CFS). The archived datasets were retrieved from NOAA's National Centers for Environmental Information (formerly the National Climatic Data Center).

CFS went operational in March 2011 replacing the first version of CFS upgrading "nearly all aspects of the data assimilation and forecast model components of the system" (Saha et al. 2014, 2185). CFS operates at one-degree grid spacing, roughly the equivalent of 100-km grid resolution, and has 64 sigma-pressure vertical levels (Saha et al. 2014, 2186). For retrospective forecasts, CFS is able to run on four different time scales (Saha et al. 2014, 2186). The first time scale is a sub-seasonal time scale mainly used for the prediction of Madden–Julian oscillation and forecasts for the United States 2-6 weeks out (Saha et al. 2014, 2186). The second time scale forecasts nine months out and is referred to as the "long lead" seasonal prediction (Saha et al. 2014, 2186). The final two time

scales are decadal and centennial where the emphasis is less on forecast skill but more on climatology (Saha et al. 2014, 2186).

The high-resolution model is NCEP's Rapid Refresh (RAP) numerical weather model. The RAP, "an hourly updated assimilation and model forecast system, replaced the Rapid Update Cycle (RUC)" in 2012 and version 2 of the RAP went operational in 2014 (Benjamin et al. 2016, 1669–71). The need for frequently updated, short-range NWP forecast products came from forecasters needing NWP models updating at or around the same time-scale they use to make decisions on issuing weather watches and warnings for rapidly changing weather events (Benjamin et al. 2016, 1669–70). RAP operates at a 13-km grid spacing and 51 sigma vertical levels (Benjamin et al. 2016, 1671). Its horizontal domain is about 3.5 times larger than the RUC covering virtually all of North America, including Alaska and the Caribbean Sea (Benjamin et al. 2016, 1671).

2.4 Clustering and K-Means

Clustering, a prevalent technique in any discipline that involves analysis of multivariate data, is used to objectively organize and group data into categories with measured or perceived inherent characteristics (Anil Jain 2010, 1). It is a type of unsupervised machine learning, which finds structure in a grouping of unlabeled data (Anwiti Jain, Rajavat, and Bhartiya 2012, 627). Jain (2010) explains clustering is used for three main purposes: (1) gain insight on the data to determine its underlying structure, (2) identify the degree of similarity through natural classification, and (3) a method to organize data and summarize it through cluster prototypes known as compression (Anil Jain 2010, 3). Clustering algorithms can be divided into two groups: hierarchical and

partitional (Anil Jain 2010, 3). The first category clusters data either bottom-up, where each data point is in its own cluster and similar data points/clusters are merged together, or top-down, where all data points start as one cluster and data is recursively divided into smaller, similar clusters (Celebi, Kingravi, and Vela 2013, 1). The partitional method "finds all the clusters simultaneously as a partition of the data" and is favored when looking for pattern recognition in data (Anil Jain 2010, 3).

One of the most popular and simple partitional clustering methods is the K-means method shown in Figure 3 (Anil Jain 2010, 1,3). The algorithm starts with *k* number of centroids, randomly chosen within the data points (Celebi, Kingravi, and Vela 2013, 2). Next each data point is assigned to the nearest centroid by calculating the distance between the data point and centroid (Figure 3b). The third step is to average the vector distance for all the data points in their respective clusters to find the new centroids by calculating the distance between the data point and centroid (Figure 3c). The last two steps are repeated until all the points converge and cluster centers stop moving (Figure 4e) (Celebi, Kingravi, and Vela 2013, 2). The goal of K-means clustering is to minimize the squared error sum over all *k* clusters (Anil Jain 2010, 4). In order to run the K-means algorithm, the user must define (1) number of *k* clusters, (2) cluster initialization, and (3) distance metric (Anil Jain 2010, 4).

Determining the number of clusters is not only the most important parameter, but also the most difficult one to define (Anil Jain 2010, 4,6). The method used in this study to determine the appropriate number of clusters is called the silhouette score. It evaluates the validity of the clustering to help determine the correct number of clusters (Rousseeuw 1987, 53). Two items are required to build silhouettes, (1) the partition obtained by some

clustering technique and (2) the collection of proximities (Rousseeuw 1987, 55). This will calculate how well an object matches the clustering it is assigned to (Rousseeuw 1987, 56). A silhouette score will have a range from -1 to 1 where the object is mismatched if the score is -1, unsure what cluster the object belongs to if its 0, and the object is well-clustered the closer the score is to 1 (Rousseeuw 1987, 56). This study used the "silhouette samples" and "silhouette score" functions from the Scikit-Learn library for Python.

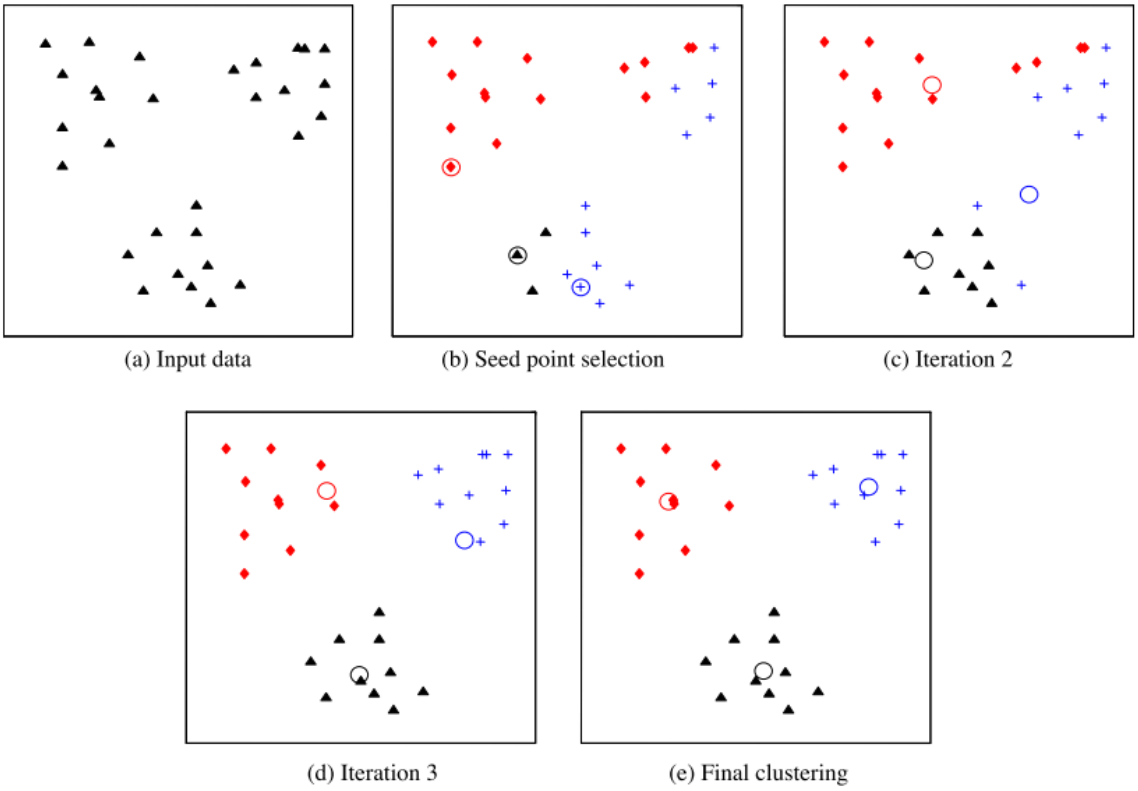

Figure 3. Illustration of the K-means method (Anil Jain, 2010).

2.5 Correction Methods

Once the trajectories for the RAP and CFS data were run, a series of post-processing techniques were run to study the usefulness of these methods in improving the

performance of the low resolution CFS relative to the performance of the RAP. This section provides background on the post-processing methods used in this study. Each method described below is a type of supervised machine learning method and can be imported as functions in Python using the Scikit-Learn library (Pedregosa et al. 2011).

The first post-processing correction method applied was linear regression. It is a simple regression that finds the relationship between one independent variable (x) and one dependent variable (y) (Gang Su and Yan 2009, 9). "The dependent variable is called the response variable, and the independent variable is called the explanatory or predictor variable" (Gang Su and Yan 2009, 10). The least squares estimation within the linear regression model finds the regression line that is "closest" to all the data points, known as a line of best fit (Gang Su and Yan 2009, 10–11).

The second post-processing technique utilized was Bayesian ridge regression. Since linear regression can have the tendency to overfit the model, ridge regression can be used to solve that problem (Bishop 2006, 9–10). Overfitting a model means it is too well trained on the training data and therefore has a high variance so the model is likely not to perform well on the testing data (Bishop 2006, 6). Ridge regression adds a bias constant which reduces variance estimates (Efendi and Effrihan 2017, 2). Bayesian ridge regression is when the parameter estimate for the ridge regression model is estimated using the Bayesian method (Efendi and Effrihan 2017, 2–3). "The method combines prior information about parameter (prior distribution of parameter) with the observed data (likelihood function) to obtain the information or distribution of parameter given data, posterior distribution" (Efendi and Effrihan 2017, 3).

The final post-processing technique applied was random forest regression. This method uses a training data set to build each decision tree and together those decision trees make the random forest (Boehmke and Greenwell 2019, 5). A decision tree is a series of vectored decisions or branches that lead to a classification; each branch of the tree represents a decision, occurrence, or reaction as in shown Figure 4 (Livingston 2005, 2; Boehmke and Greenwell 2019, 5). The random forest takes all the outputs from each decision tree and the classification with the most votes is the final solution (Livingston 2005, 1). Once the random forest has been built using a training dataset, the remaining data (or out-of-bag data) is used to test the performance of the individual trees as well as the entire forest (Livingston 2005, 1–2). "The average misclassification over all trees is known as the out-of-bag error estimate. This error estimate is useful for predicting the performance of the machine learner without involving the test set example" (Livingston 2005, 2).

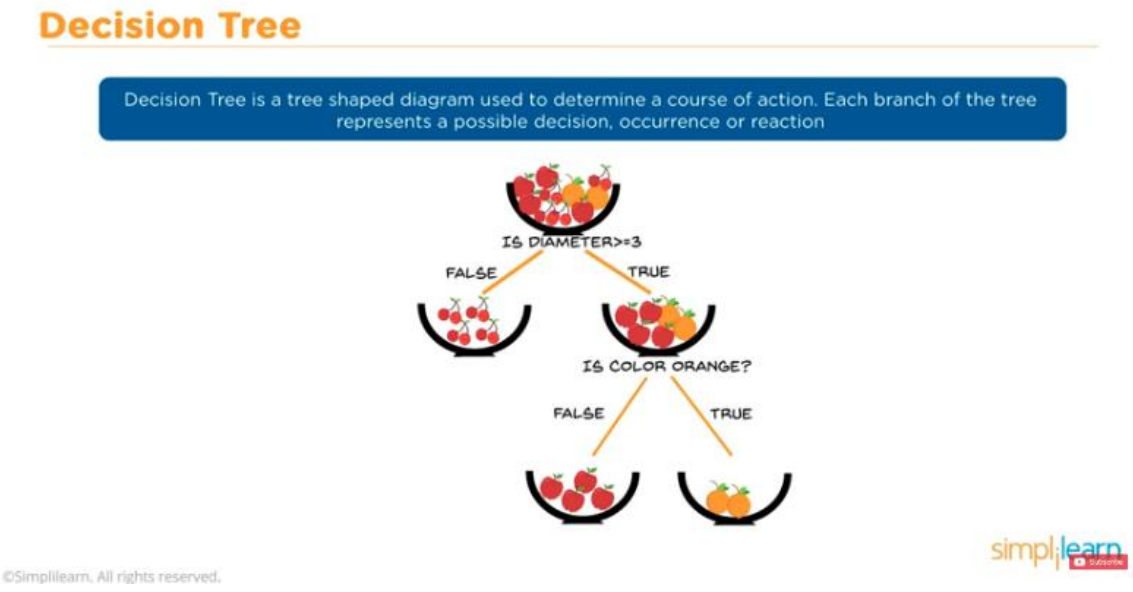

Figure 4. Visualization of how a decision tree works. Image courtesy of Simplilearn.

2.6 Previous Research

Draxler (2006) reexamined a previous analysis from 1984 to compare trajectory outputs of a global and mesoscale model data to predict the transport and dispersion of tracer plumes over Washington D.C. (Draxler 2006, 383–84). "The goal of this analysis is to use some historical experimental dispersion data appropriate over an urban area to evaluate how well routine meteorological model products in conjunction with a dispersion model could simulate pollutant releases" (Draxler 2006, 383). The historical experimental data used was from the Metropolitan Tracer Experiment (METREX); it was "conducted for more than [one year] in 1984 around Washington, D.C., and its suburbs and consisted of hundreds of tracer releases with air samples collected from a few kilometers to about 75 km from the tracer source locations" (Draxler 2006, 384).

The coarse-grid global meteorological data sets used in this study were the National Centers for Environmental Prediction (NCEP) reanalysis data and European Centre for Medium-Range Weather Forecasts 40-year (ERA-40) reanalysis data (Draxler 2006, 386). The high-resolution reanalysis data came from Pennsylvania State University–NCAR, fifth generation Mesoscale Model (MM5) (Draxler 2006, 386). These reanalysis data sets were then inputted into HYSPLIT and compared to measurements from the METREX experiment. Results showed that for shorter distances, dispersion calculated using MM5 data performed better. If only global data was available and needed to be used, ERA-40 out-performed NCEP-NCAR when it came to an 8-hour dispersion plume, but results were comparable for the 30-day sampling (Draxler 2006, 392). Draxler also found "that small changes in the wind direction with respect to the

tracer release location provided dramatic improvements in dispersion model performance" (Draxler 2006, 392).

Ivatt and Evans (2020) researched improving the prediction of tropospheric ozone transport using gradient-boosted regression trees (Ivatt and Evans 2020, 8063). They developed a predictor bias using GEOS-Chem, an atmospheric chemistry transport model, using outputs from the model and observations from both surface ozone measurements and ozone-sondes (Ivatt and Evans 2020, 8063). Surface ozone measurements came from European Monitoring and Evaluation Programme, the United States Environmental Protection Agency (EPA), and the Global Atmospheric Watch, and were compiled between 2010 and 2018 (Ivatt and Evans 2020, 8064). Ivatt and Evans developed the bias predictor by using hourly ozone observations from surface and ozone-sondes measurements for the training period of 1 January 2010 to 31 December 2015 (Ivatt and Evans 2020, 8065).

The machine learning technique they used to develop a function to predict the model bias was the gradient-boosting regression, which is a type of decision tree regression. "[G]radient-boosting regression relies on building a tree with a relatively shallow depth and then fitting a subsequent tree to the residuals" (Ivatt and Evans 2020, 8066). This process was repeated until the model was able to generalize the dataset without overfitting it (Ivatt and Evans 2020, 8066). The technique was selected because of its ability to identify non-linear relationships found in atmospheric chemistry, its relatively quick training time, and finally it is highly scalable (Ivatt and Evans 2020, 8066). Their results showed the bias-corrected models performed considerable better than the uncorrected models (Ivatt and Evans 2020, 8076–77). The methodology used in

this study could provide the starting point for a large-scale PM 2.5 and ozone case study described in section 5.3.

Cervone et al. (2008) studied the risk assessment of atmospheric emissions using machine learning. They referenced current atmospheric transport and dispersion models are only able to provide emergency responders the general risk of contamination over the entire time period (Cervone et al. 2008, 991). This paper focused on developing a methodology that would specify areas of risk under specific meteorological conditions using machine learning (Cervone et al. 2008, 991). The information required for their methodology was data on the source characteristics, an extensive database of concentration measurements at several sampling points, and ensuring atmospheric releases occurred under different meteorological conditions (Cervone et al. 2008, 991). This information was run through several transport and dispersion simulations and a clustering algorithm was used to group those results according to specific cloud characteristics (Cervone et al. 2008, 991–92). The results from the clustering algorithm were then inputted into a symbolic classification algorithm, which is a type of supervised machine learning "where the answer can be directly inspected and hence accepted or rejected" (Cervone et al. 2008, 995). To test their methodology, Cervone et al. did a case study over the Istanbul channel using data from 2004 and running weekly transport and dispersion simulations resulting in 53 runs (Cervone et al. 2008, 996–97). While they were able to identify areas that are at risk from a potential accident, they noted the paper focused on methodology rather than application and concluded that larger simulations would need to be run in order to increase statistical meaningfulness (Cervone et al. 2008, 998–99).

Yerramilli et al. (2009) examined three different atmospheric transport and dispersion simulations of air-borne releases from point elevated sources in the Mississippi Gulf coastal region (Yerramilli et al. 2009, 1055). The simulation took place from 1-3 June 2006 during a typical summer synoptic set-up where there was easterly synoptic flow associated with a high pressure center over the subtropical North Atlantic Ocean (Yerramilli et al. 2009, 1058). Each simulation used three different meteorological data sets that were then inputted into HYSPLIT (Yerramilli et al. 2009, 1055,1060). "The first two data [sets were] the NCEP global and regional analyses (NCEP's Final Analysis and Eta Data Assimilation System) while the third [was] a meso-scale simulation generated using the Weather Research and Forecasting [WRF] model with nested domains at a fine resolution of 4 km" (see their Figure 1) (Yerramilli et al. 2009, 1055). The emissions dataset was from the Mississippi Department of Environmental Quality (MDEQ) and the four major sources of pollution in the region that were considered were (1) Mississippi Power Company-Plant Jack Watson, (2) Chevron Products Company-Pascagoula Oil Refinery, (3) Mississippi Power Company-Plant Victor Daniels, and (4) Dupont Delisle Facility (Yerramilli et al. 2009, 1060). Their results showed that the WRF was more successful at capturing the diurnal patterns along the coastline, such as land/sea breezes, whereas the other two datasets showed a relatively uniform plume dispersion (Yerramilli et al. 2009, 1067–68). Additionally, "[m]odel concentrations with WRF data are closer to [MDEQ] observations up to 40 h and deteriorate thereafter" (Yerramilli et al. 2009, 1070).

Finally, Feng et al. (2019) investigated methods for improving tropospheric ozone forecasts in Hangzhou, China since it has become the primary atmospheric pollution for

that region in recent years (Feng et al. 2019, 366). They first used a traditional atmospheric model, the WRF coupled with Community Multi-scale Air Quality (WRF-CMAQ) and compared those results to results from machine learning models (Feng et al. 2019, 366–67). The four machine learning models Feng et al. tested were extreme learning machine (ELM), multi-layer perceptron (MLP), random forest (RF) and recurrent neural networks (RNN) (Feng et al. 2019, 367). The hourly surface observations were measured at the Westbrook National Park, representing urban weather conditions and at Zhejiang Agriculture & Forestry University, representing rural weather conditions (Feng et al. 2019, 368). Additionally, hourly atmospheric pollutants were measured at Westbrook National Park (Feng et al. 2019, 368). Since Hangzhou experiences four distinct climates, Feng et al. made sure to run the WRF-CMAQ and machine learning models through daily simulations in January, April, July and October representing winter, spring, summer and autumn, respectively (Feng et al. 2019, 372).

According to their results, the WRF-CMAQ simulations captured the tropospheric ozone episodes in late July and early October, but only got the minimum value on April 7th and completely missed the episode from January 22nd to 26th (Feng et al. 2019, 372). Three out of the four machine learning methods out-performed the WRF-CMAQ, the exception being ELM (Feng et al. 2019, 372). Additionally, they were able to deduce that the formation of tropospheric ozone tends to be non-linear and non-smooth because RF performed better than ELM (Feng et al. 2019, 375). Feng et al. concluded that RNN, a method known for performing well at time series forecasting and characterization, performed the best out of all four of the machine learning models (Feng et al. 2019, 372).

III. Methodology

The purpose of this project is to compare performance of short-term (24-hour, from 18Z to 18Z) back trajectories using CFS model data (global, low-resolution) and RAP model data (regional, high-resolution). These regimes will be learned from the data using the unsupervised K-means clustering technique. Several methods will then be used to correct the CFS trajectory results to the RAP trajectory results. It is assumed that, due to the higher horizontal and vertical resolution, the RAP trajectory results are more accurate in the light of recent studies comparing global model performance to regional, high-resolution model performance with modern verification techniques (Mittermaier 2014, 202). Additionally, this study only focused on the trajectories' changes in latitude and longitude, and did not include changes in the vertical direction in the K-means clustering nor post-processing CFS correction techniques. For back trajectories, the RAP analysis was utilized and the CFS analyses was used for back trajectory analysis. The area of interest for this study was eastern CONUS covering the years 2015-2019. The trajectory code was adapted from HYSPLIT for Python.

This study plotted daily back trajectories using CFS and RAP only on days when both CFS and RAP had data archived during 2015-2019 to ensure each model had the same number of days analyzed. Similar to Zoellick (2019), a trajectory release grid was utilized, but for this study trajectories were released at half degree intervals over the CONUS from 34 to 42 degrees latitude and -100 to -85 degrees longitude leading to a release grid of size 17 by 31 points (Figure 5). To capture the uncertainty of the method, 50 trajectories were released from each grid point with random location error of +/- 5 km

and +/- 5 hPa resulting in over 1.5 million trajectory clouds released per model for this study.

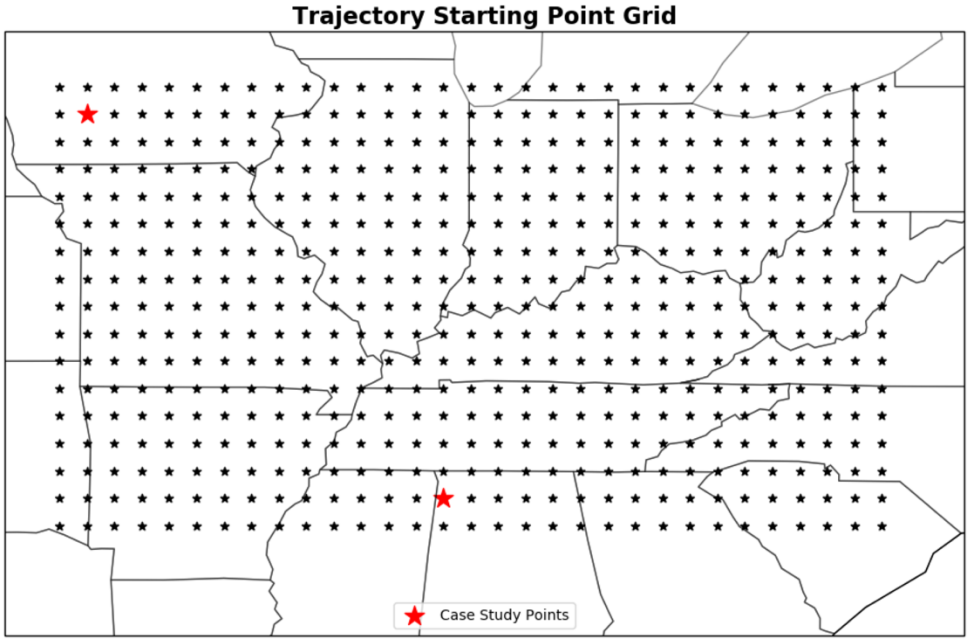

Figure 5. The trajectory release grid utilized in this study. Trajectories were released at half degree intervals over the CONUS from 34 to 42 degrees latitude and -100 to -85 degrees longitude leading to a release grid of size 17 by 31 points. The case study points were chosen to compare CFS and RAP performance.

After the trajectory data was compiled, the performance of the CFS relative to the RAP was analyzed by splitting up the 5 years of trajectory data into four groups, three months per group. A simple two-dimensional Euclidean distance metric was implemented to examine the performance between the RAP and CFS to compare the relative spread in the 50 back trajectories release from each grid point (Figure 5). Next the study utilized the K-means clustering technique, a type of unsupervised machine learning, to perform clustering of back trajectories across the entire domain. In order to determine the best number of *k* clusters, the silhouette scores for different *k* values were calculated.

With knowledge of the performance of the CFS relative to the RAP across different clusters in hand, the next step was to improve the performance of the low resolution CFS relative to the RAP using three different supervised machine learning techniques. These techniques were only applied to the endpoints of the CFS back trajectories to primarily focus on the origin of the trajectories and lowers computation time. Before each technique was tested, the five years of CFS back trajectory endpoint data was split into a training data set (three-fourths of the data) and a testing data set (one-fourth of the data). Both data sets were statistically similar and incorporated days from all four seasons for the midlatitudes. Additionally, the three correction methods were used on all data at once and the clustered data using the K-means clustering. The motivation behind breaking the data up by unclustered and clustered was to investigate if the CFS back trajectory would perform better or not post-correction methods since clustering the data before applying a correction method requires a longer run time for the Python script.

First simple methods were used, such as linear and Bayesian ridge regression then a more complex statistical technique, random forest regression was carried out. To determine the best values for the random forest regression parameters, Scikit-Learn's RandomizedSearchCV method, also referred to as hyperparameter tuning, was used (Bergstra and Bengio 2012). By defining a matrix of hyperparameter ranges, the function randomly samples from the grid to perform several iterations of random forest regression until all the values in the matrix have been used. Once complete, the method will return the best parameters to be used in random forest regression for the CFS back trajectory data.

IV. Results and Analysis

4.1 Overview

HYSPLIT-based back trajectory Python code was run to calculate 24-hour back trajectories (18Z-18Z) for RAP and CFS data from 2015-2019 using a starting point grid that was 17 by 31 points (see Figure 5). A total of 1,553 days were plotted during that five year window. Each starting point had 50 trajectories released from the surface with random location error of +/- 5 km and +/- 5 hPa. This chapter executes the methodology described in Chapter 3. Section 4.2 compares model performances for CFS and RAP back trajectories and section 4.3 compares the results from the three different correction methods tested on the CFS back trajectories. Section 4.4 analyzes a correction method results by cluster from K-means clustering and finally, section 4.5 investigates where in the CONUS domain a correction method was successful and unsuccessful.

4.2 Comparing CFS and RAP Back Trajectories

Figures 6 and 7 shows an example of the output of one day's worth of back trajectories for CFS and RAP, respectively. For the figures to be more readable, only 144 out of the 527 starting points were plotted overlaid with 50 released trajectories for each starting point. Overall both Figures 6 and 7 are showing anticyclonic flow which a result of a high-pressure system center over Maryland (Figure 8). The CFS back trajectories in Figure 6 show the highest amount of spread between trajectories originating at the same starting points in northern Mississippi and Alabama.

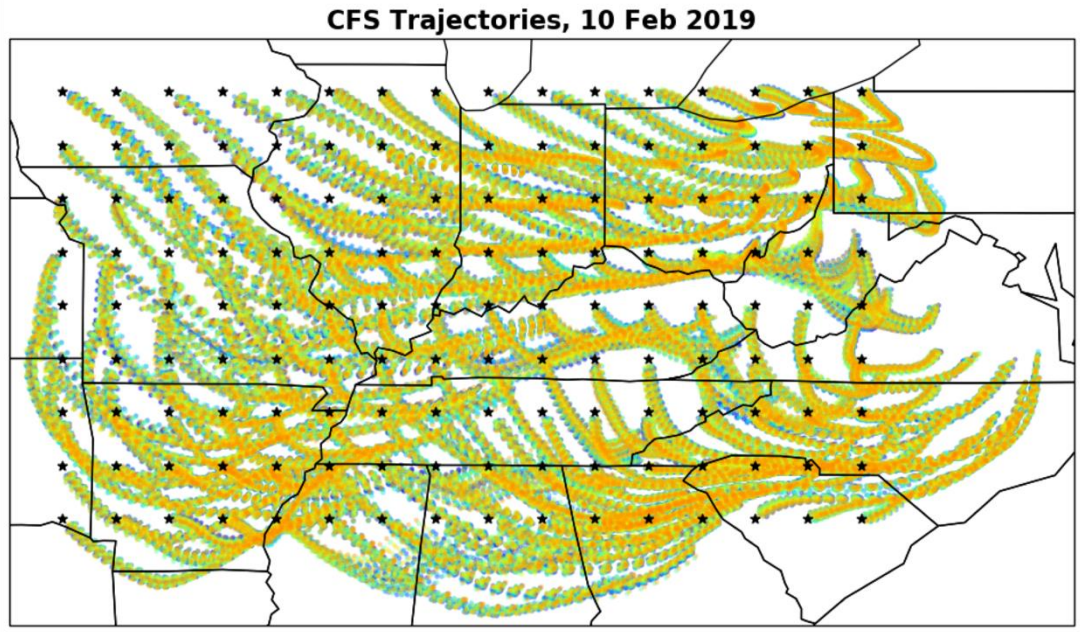

Figure 6. Example of a 24-hour back trajectory release for just one day of CFS data. Only 144 out of the 527 starting points are shown in this image and each starting point released 50 trajectories. A point is plotted every 30 mins in each trajectory.

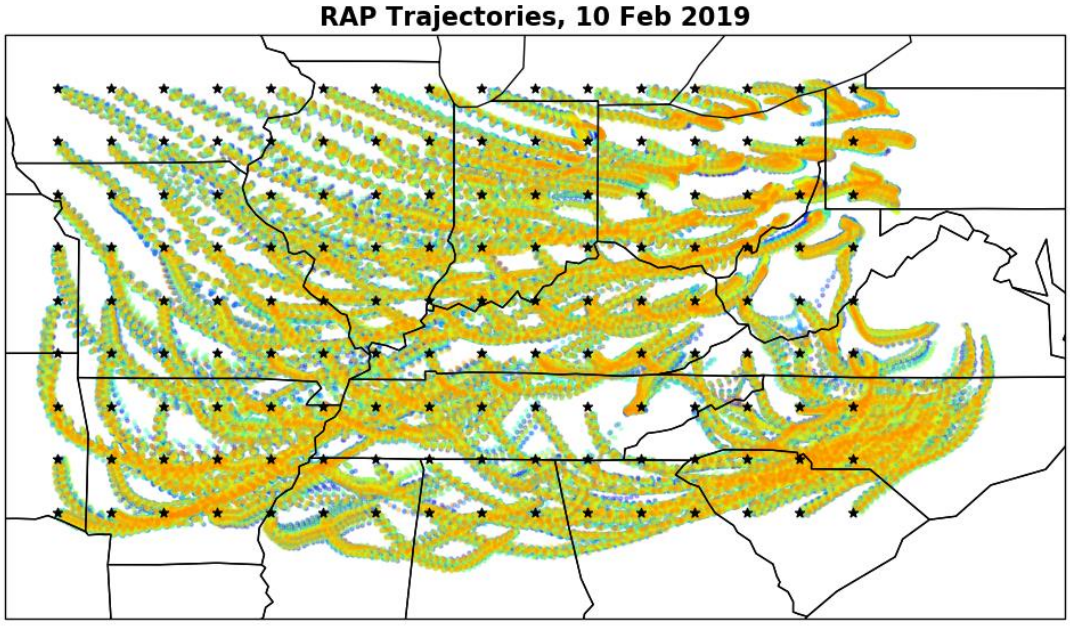

Figure 7. Example of a 24-hour back trajectory release for just one day of RAP data. Only 144 out of the 527 starting points are shown in this image and each starting point released 50 trajectories. A point is plotted every 30 mins in each trajectory.

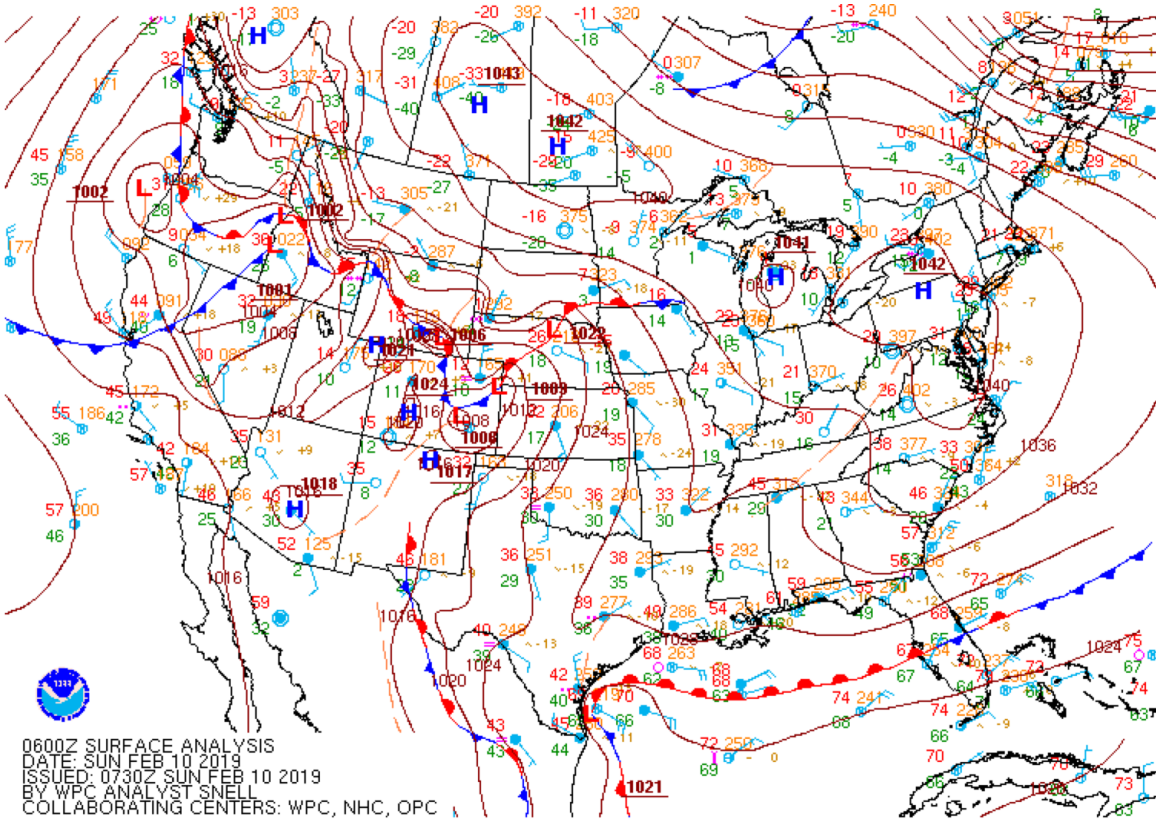

Figure 8. 06Z, 10 Feb 2019 surface analysis courtesy of NOAA.

Figure 6 also show significantly less spread in back trajectories over Virginia, North Carolina and South Carolina closer to the center of the high-pressure compared to those same RAP trajectories in Figure 7. While conclusions on CFS and RAP performance cannot be drawn from just one day's worth of back trajectories, it is still important to note differences in daily CFS and RAP plots before looking at seasonal trends for each model over all 5 years' worth of data.

To get a better understanding of how each model performed throughout different times of the year, all five years of the trajectory data were split up into four parts: (1) January, February, and March (JFM), (2) April, May, and June (AMJ), (3) July, August,

and September (JAS), and (4) October, November, and December (OND). The majority of the figures to follow in this section will be showing the trajectory data analyzed in these quarterly segmented months.

Figures 9-12 compare the CFS (red) and RAP (cyan) trajectory mean and spread. Capturing the expected value of winds is problematic as the mean is subject to cancellation of winds from different directions. Other methods (median and mode) over capturing the average are even more problematic due to even more severe lack of representativeness of the distribution. All 50 trajectories were averaged for each starting point and the size of each point in the averaged trajectory is how much spread (or distance) there was between those 50 trajectories from that starting point. Overall, the RAP trajectories show more spread than the CFS trajectories, especially over areas with higher terrain such as the Appalachians and the Ozarks. This is likely due to the RAP trajectories experiencing more change in the vertical direction as a result of the terrain and getting above the planetary boundary layer and the RAP was more sensitive to the random location error applied to every trajectory released.

In Figure 9, both RAP and CFS trajectories overall are coming from the west/northwest, especially grid points located in the Midwest, which is consistent with the flow of mid-latitude cyclones moving west to east over CONUS. Winter months are known for having more mid-latitude cyclones pushing through this region as the polar front jet (PFJ) moves equatorward. Additionally, these northern tier RAP and CFS trajectory points exhibit the same about of spread based on trajectory point size, whereas RAP trajectories in Arkansas and Appalachian states show significantly more spread than the CFS trajectories as a result of complex topography in these regions.

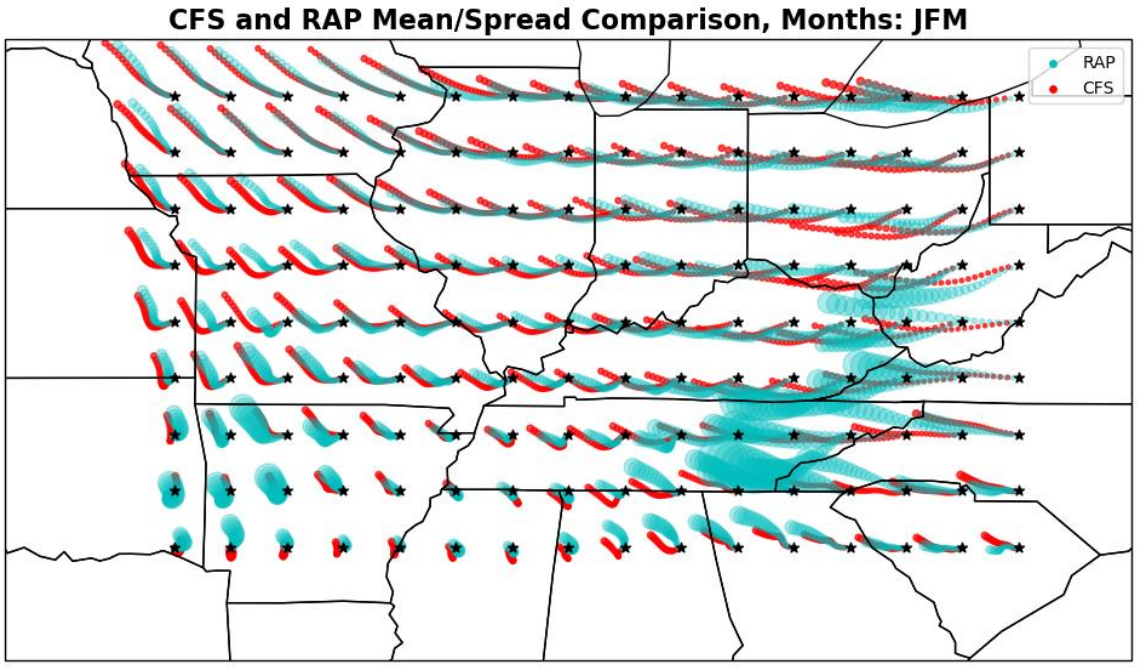

Figure 9. CFS (red) and RAP (cyan) trajectory mean and spread comparison plot for the months of January, February, and March. All 50 trajectories were averaged for each starting point and the size of each point in the averaged trajectory is how much spread there was between those 50 trajectories from that starting point. A point is plotted every 30 mins in the trajectory.

There is a shift in direction for back trajectories in Figure 10 compared to 9; the majority of back trajectories are now originating from the south/southwest as a result of seasonal weather patterns starting to shift during these months as the PFJ moves back poleward. The land heats faster than the water in the Gulf of Mexico causing a difference in air density and southern winds reflected in the back trajectories in the southern half of the grid points. Some of the most southern grid points through Alabama and Georgia show the RAP back trajectories with larger spread and not traveling as far as the CFS trajectories. In Iowa, however, there appears to be very little change in the trajectories' direction during these months. This is a consequence of averaging the trajectories and it indicates there is no dominant flow in this region in AMJ.

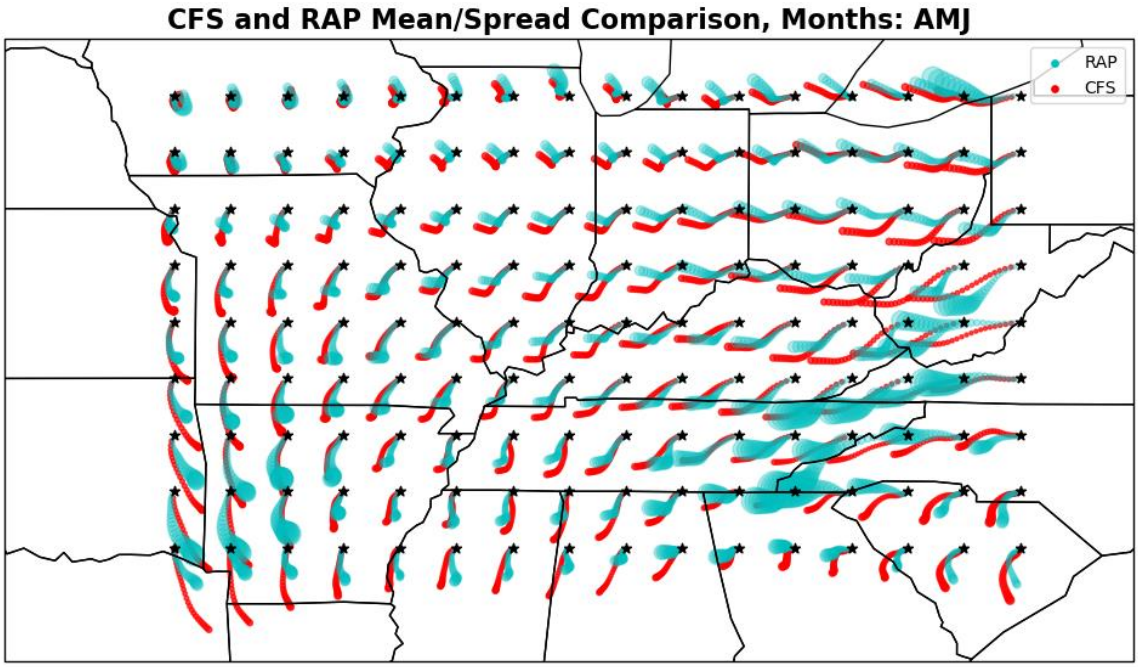

Figure 10. Same information as Figure 9 except this plot is months April, May, and June.

In Figure 11, the back trajectories are overall showing an anticyclonic flow suggesting high pressure dominates this region during JAS. Additionally, as the PFJ retreats poleward, it causes wind speeds to decrease, which explains why the back trajectories did not travel as far compared to Figures 9 and 10, specifically in Illinois, Indiana, and eastern Kentucky. This is also an effect of averaging the back trajectories when there is no dominant flow. In Arkansas, there is a significant disagreement between the RAP back trajectories and CFS. The directionality for CFS back trajectories is directly from the southeast whereas RAP back trajectories tend to curve clockwise from the southeast back to the starting grid point. Finally, over the Appalachians, the CFS back trajectories traveled further in 24 hours than the RAP trajectories, but RAP exhibited more spread.

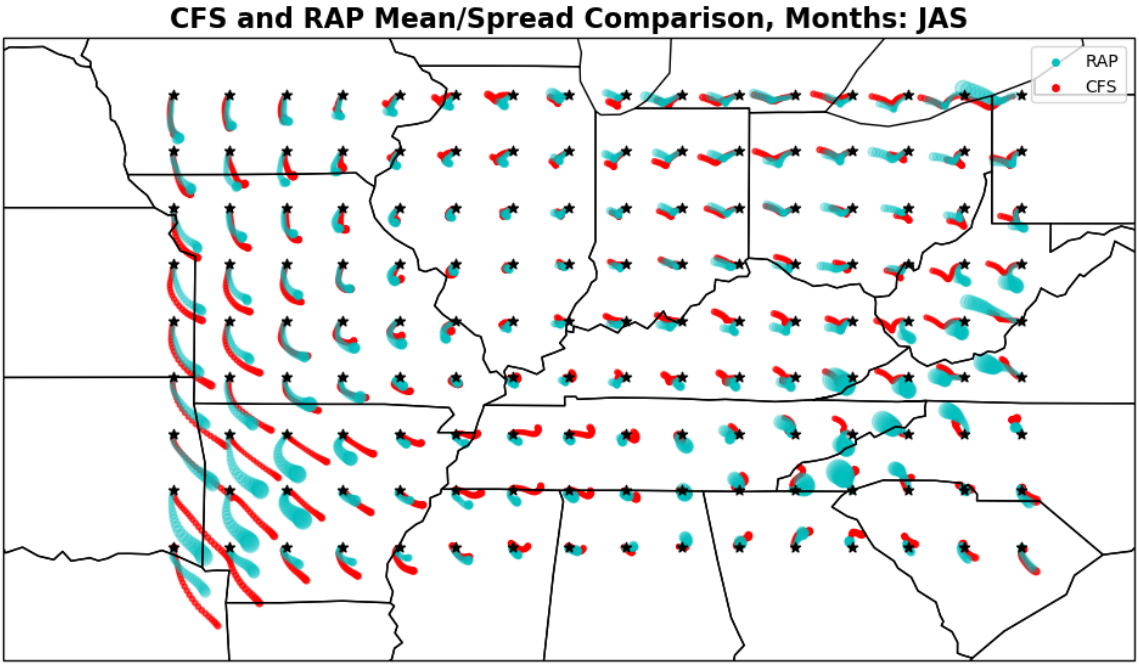

Figure 11. Same information as Figure 9 except this plot is months July, August, and September.

Back trajectories from both models in Figure 12 exhibit a very similar pattern shown in Figure 9 as seasonally the PFJ begins to move equatorward once again. The CFS back trajectories for the northern half of the grid points performed generally the same as the RAP back trajectories with the exception of the trajectories that traveled further than RAP in Iowa and Missouri. Once again, the areas with the largest difference in model performances was the Ozarks and the Appalachians where the RAP trajectory spread was the most significant. These differences, and previously mentioned ones, in the CFS model's performance compared to the RAP is what this study is striving to correct using the methods described in section 2.5.

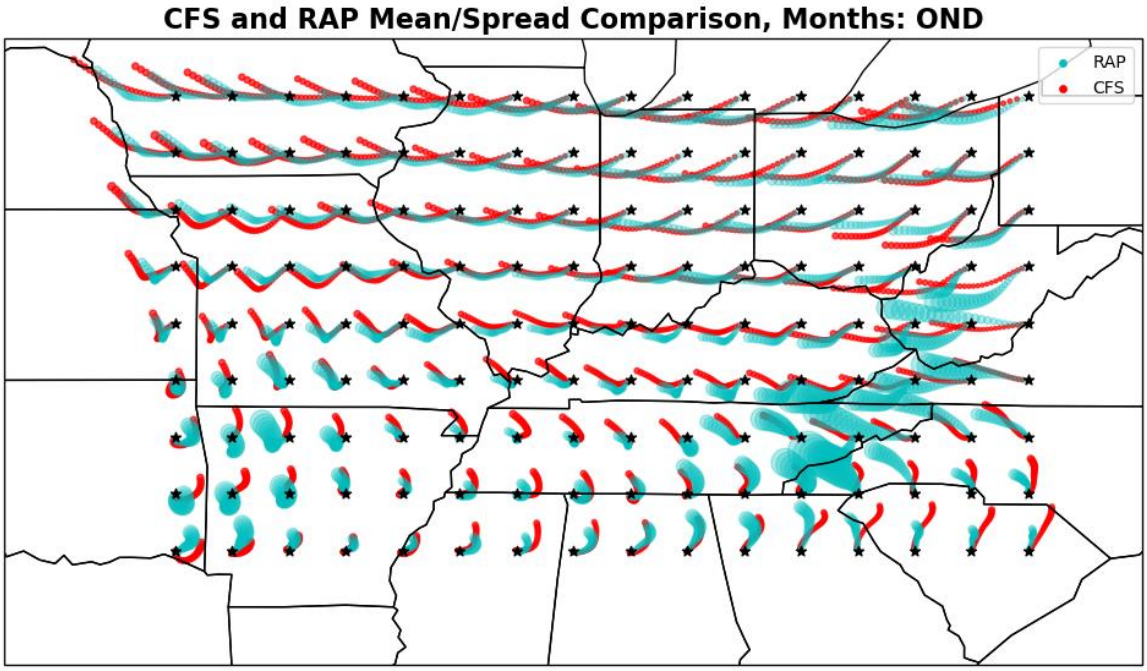

Figure 12. Same information as Figure 9 except this plot is months October, November December.

Figure 13-16 shows the analysis of all trajectories from every release grid point. The scatter plots on the left represent the distance between the first point and end point of the averaged 50 trajectories for each release grid point to compare the RAP results to the CFS. The black line is plotted where RAP equals CFS. The closer the points are to that line show when the models performed roughly the same. In other words, the RAP and CFS trajectories traveled roughly the same distance. The scatter plot on the right plots the average distance (or spread) between the end points for all 50 trajectories for every point on the release grid comparing the RAP to the CFS; and the black line represents the same information as the left scatter plots. The larger the value, the higher the spread, or uncertainty inherent in the trajectories.

Overall, there is much better agreement between the two models' performances for distances between the first point and end point of the trajectories (left side scatter plots) than there is for the scatter plots on the right. This indicates the CFS performed more closely to the RAP when it came to distances the back trajectories traveled, but was not as successful in displaying the same amount of endpoint spread as compared to the RAP. Points that are further away from the black line, especially in Figures 13 and 14, is where this study seeks to improve the CFS's performance using the correction methods described in section 2.5. Also Figure 13 right-side scatter plot shows there are no points on the black line indicating that zero CFS back trajectories had the same averaged endpoint spread as the RAP; meaning the CFS performed very poorly compared to the RAP.

Figure 15 is where the two scatter plot points are closest to the black, slope=1 line indicating JAS is when the CFS performed most closely to the RAP. Comparing the endpoint distance scatter plot points in Figure 15 to Figures 13, 14, and 16, back trajectories traveling the least amount of distance for both models. Additionally Figure 15 endpoint mean spread displays the lowest for both models, which indicates trajectory spread relates to the distance a trajectory travels.

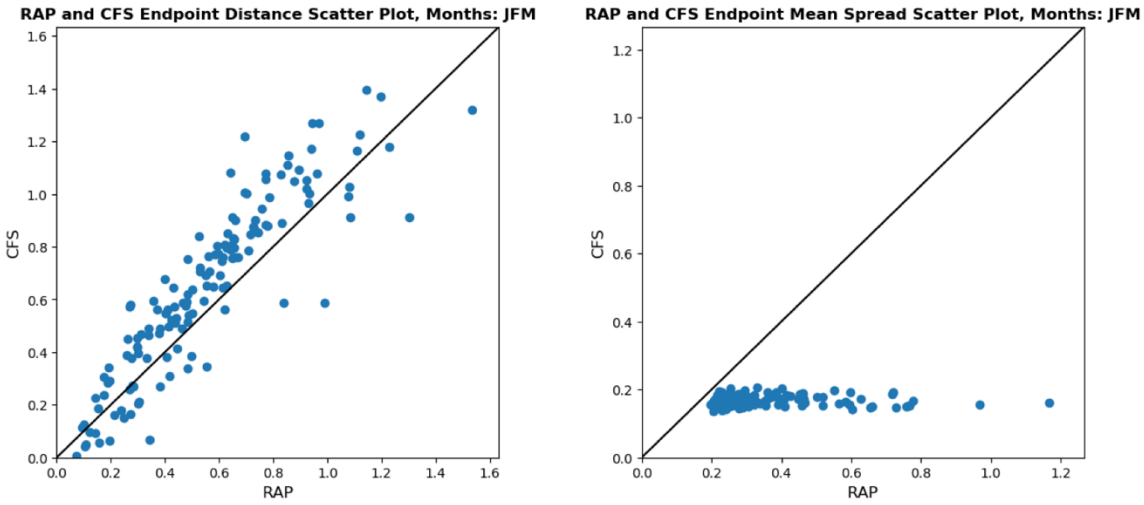

Figure 13. The scatter plot on the left is plotting the distance between the first point and end point of the averaged 50 trajectories for each release grid point to compare the RAP results to the CFS. The scatter plot on the right is plotting the average distance between the end points for all 50 trajectories for every point on the release grid comparing the RAP to the CFS. The black line plotted represents when the RAP equals the CFS or when the models performed the same. These plots are for January, February, and March.

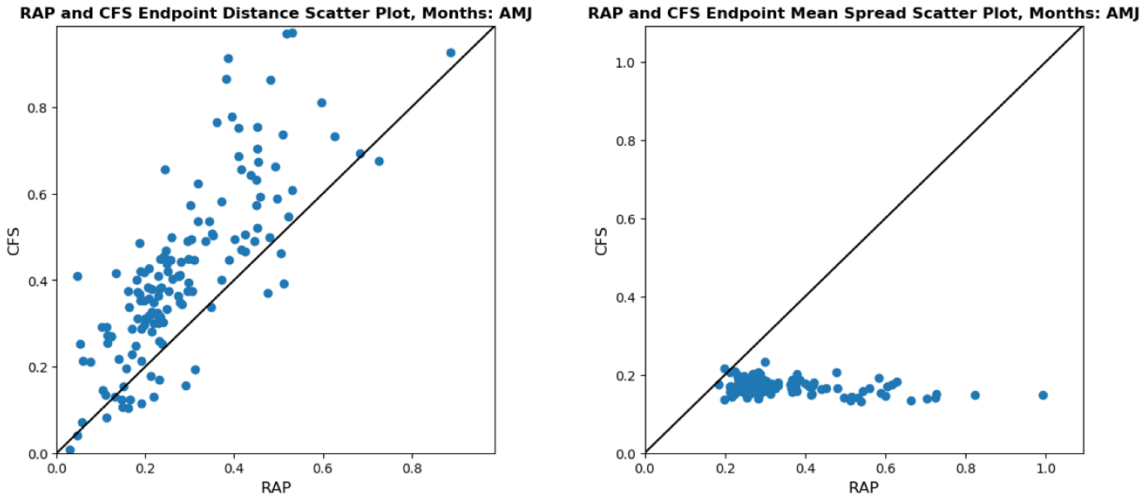

Figure 14. Same information as Figure 13 except this plot is months April, May, and June.

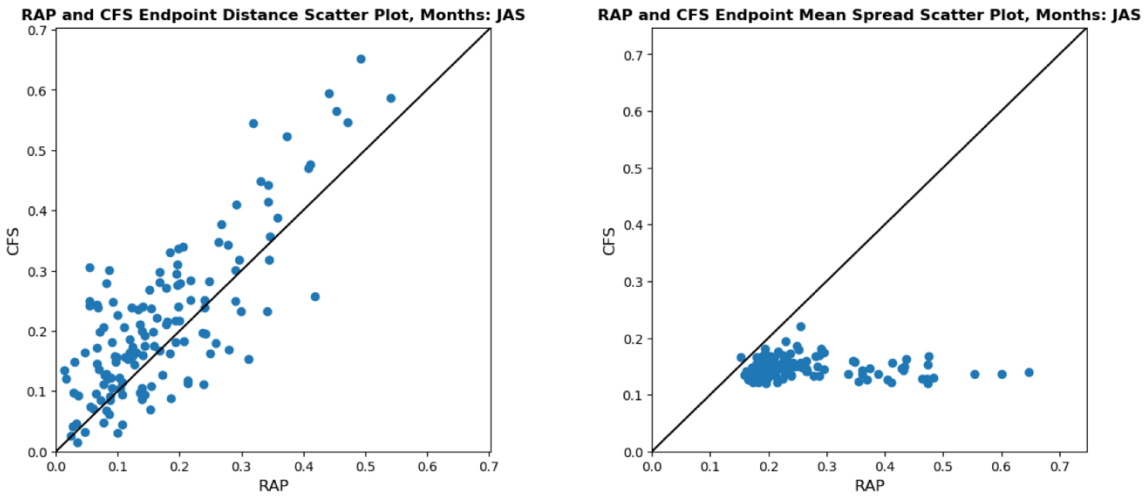

Figure 15. Same information as Figure 13 except this plot is months July, August, and September.

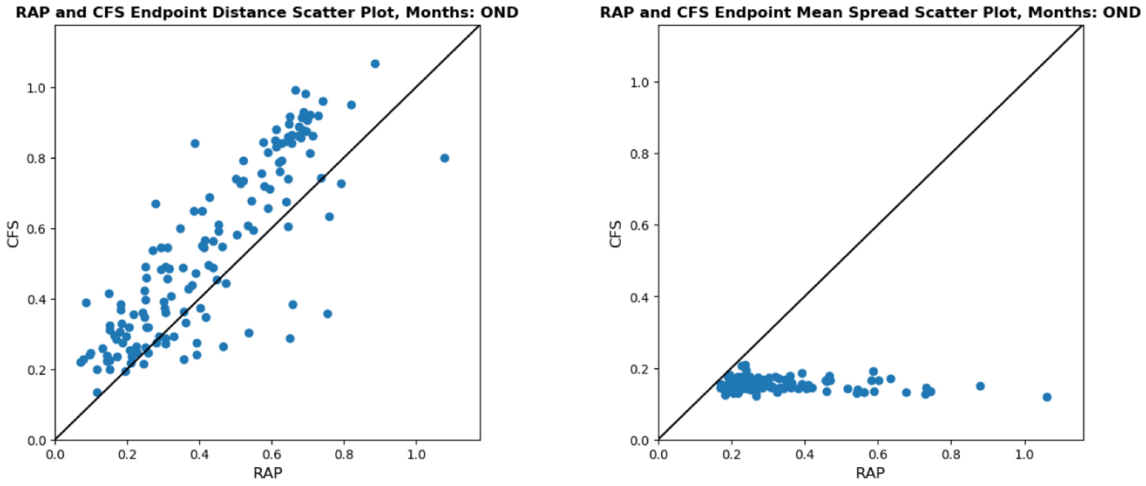

Figure 16. Same information as Figure 13 except this plot is months October, November, and December.

To analyze the RAP and CFS performance based on different atmospheric flow regimes, the back trajectories were inputted into a K-means clustering Python function. Based on the results of silhouette scores, the number of *k* clusters used were 4 and 8. The results of the K-means clustering is shown in Figure 17. Before the K-means clustering was calculated, all RAP and CFS trajectories' starting latitudes and longitudes were

adjusted so they all had the same origin in order for the K-means clustering to be run more effectively. This adjustment did not affect the relative distance from origin point of the rest of the points in the trajectories. For example, if the back trajectory starting point was 32 degrees north latitude and -85 degrees west longitude and the end point in the back trajectory was 33.3 north and -85 west, 32 and -85 were subtracted from all trajectory points, giving a starting point of (0,0) and an end point of (1.3, -1). The closer the CFS clustering is to the RAP clustering, the closer the CFS back trajectories performed to the RAP back trajectories. Additionally, the size of the X's and circles plotted in Figure 17 are a pictorial depiction of the spread in the clustered back trajectories for CFS and RAP, respectively. It highlights the RAP back trajectories had more spread than the CFS trajectories because the change in the size of the X's are more difficult to distinguish compared to the circles, representing the RAP's spread.

When comparing the four cluster results (left) to the eight cluster results (right) in Figure 17, the spread per cluster for both models is larger when there are 8 clusters. This figure also shows the further the clustered back trajectories traveled, the more spread there is for both CFS and RAP. Additionally, clustered back trajectories traveled furthest when originating from the northwest to the southeast of the origin. The quarterly segmented months for both cluster results generally show the clustered back trajectories originating from the same direction, but $k = 8$, the clustering shows trajectories that traveled two different distances from the same direction whereas the four clusters seemed to average those two distances.

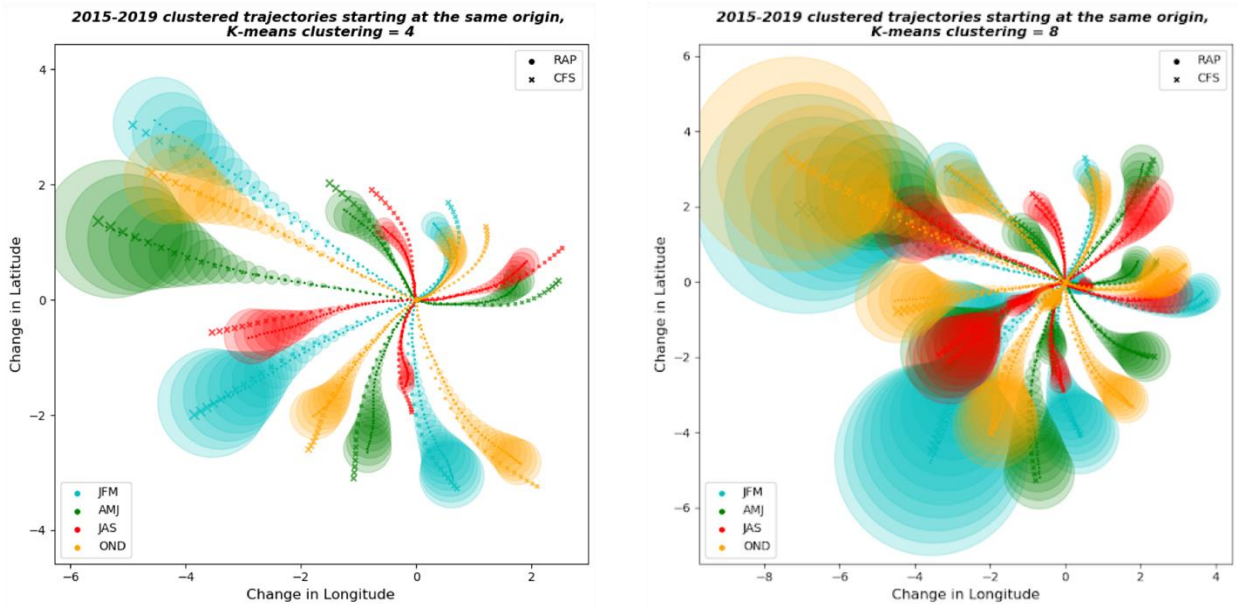

Figure 17. Results from the K-means clustering for the CFS and RAP back trajectories when $k = 4$ (right) and $k = 8$ (left). The starting latitude and longitudes for the trajectories were modified to have the same origin for the clustering could be done more effectively; it did not affect the rest of the points in the trajectories. The RAP back trajectories are the circle-shaped markers with size of the circle determined by mean spread per cluster. The CFS trajectories are the X-shaped markers with size of the circle determined by mean spread per cluster.

Figure 18 shows similar results as Figure 17 except it is just focused on $k = 4$ and each of the quarterly segmented months are their own figure for easier analysis on CFS and RAP performances. During JFM (cyan), CFS back trajectories in all four clusters travel further than the RAP trajectories but have very similar directionality except for back trajectories originating from the north-northeast relative to the origin. CFS trajectories originate more northeast relative to RAP trajectories which did not travel as far as CFS trajectories and originated more from the north relative to the origin.

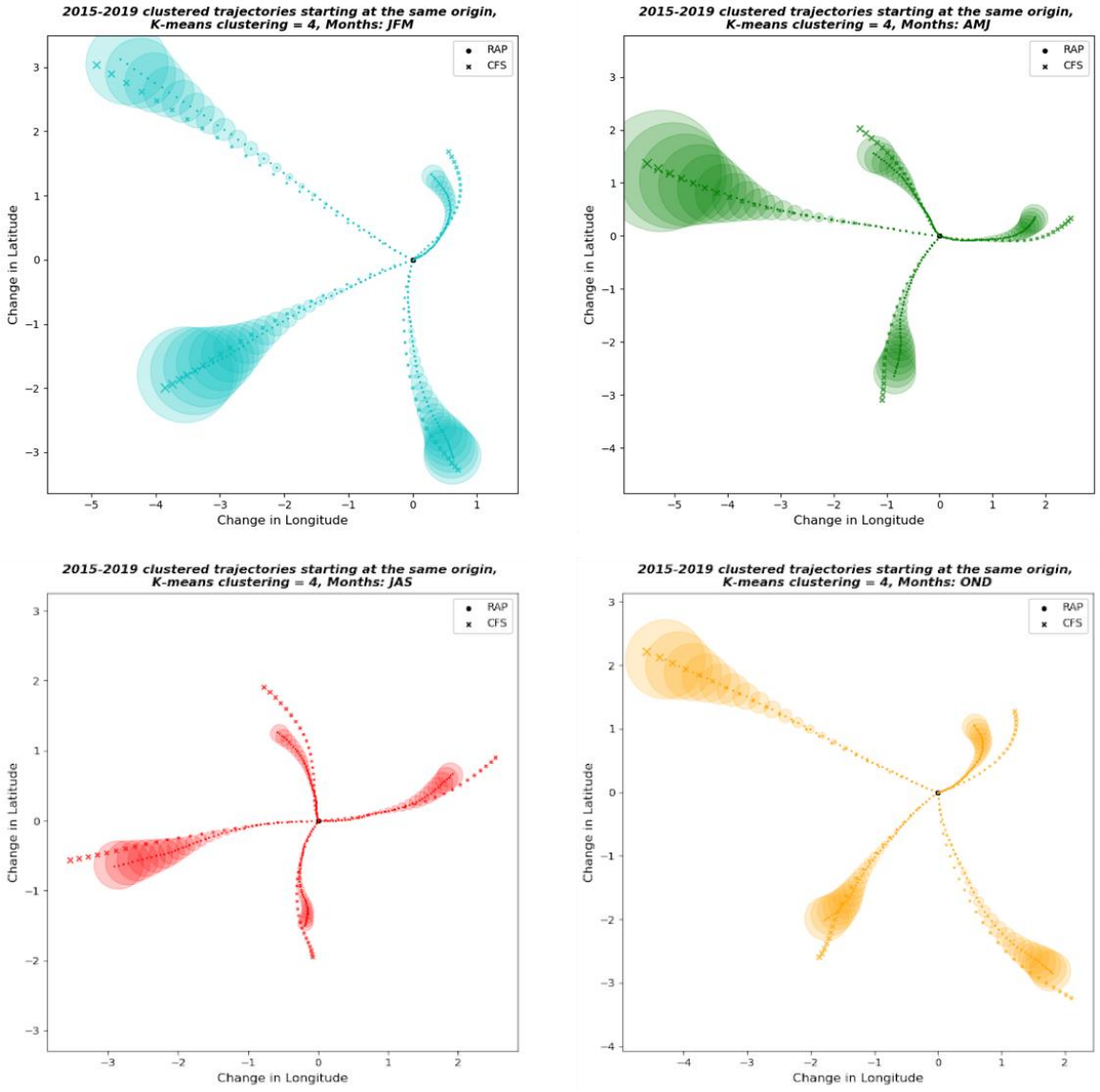

Figure 18. Results from the K-means clustering just for $k = 4$ where each of the quarterly segmented months are their own plot. The starting latitude and longitudes for the trajectories were modified to have the same origin for the clustering could be done more effectively; it did not affect the rest of the points in the trajectories. The RAP back trajectories are the circle-shaped markers with size of the circle determined by mean spread per cluster. The CFS trajectories are the X-shaped markers with size of the circle determined by mean spread per cluster.

AMJ (green) clusters show RAP back trajectories from the west having the biggest spread out of the four clusters whereas the CFS back trajectories showed very

little spread from this direction, but both models' clusters travel the same distance in 24 hours. The remaining three clusters' results for these months reveal the CFS back trajectories' directionality did not line up with the RAP's. Additionally, clustered CFS back trajectories originating to the northwest of the origin exhibited the most spread compared to the other three clustered back trajectories for AMJ.

For JAS (red), clustered back trajectories had the smallest change in latitude and longitude, in other words traveled the shortest distance in 24 hours, compared to the other plots in Figure 18. Additionally, all clustered CFS back trajectories had trajectories traveling further in 24 hours than the clustered RAP back trajectories. The clustered RAP back trajectories also showed the smallest amount of spread compared to the other plots in Figure 18 which again indicates that trajectory spread might be related to distance trajectories travel as discussed with Figure 15.

Finally, the OND (gold) clustered back trajectories showed similar results to the JFM plot where trajectories from the northwest to the origin traveled the furthest in 24 hours. Additionally, CFS back trajectories originating from the northeast had the most difficulty performing like the clustered RAP back trajectories for that same cluster, similar to the JFM plot. The clustered CFS back trajectories from the southwest and southeast traveled roughly half a degree further than the RAP trajectories.

4.3 Correction Methods Applied to CFS trajectories

After the clustered CFS back trajectory results were analyzed and compared to the clustered RAP back trajectories, the correction methods were applied to the CFS clustered and unclustered data. Three different correction methods that were tested and

described in section 2.5 were (1) linear regression, (2) Bayesian ridge regression, and (3) random forest regression. The goal was to determine which correction method improved the low resolution CFS performance the best relative to the RAP.

The first metric used to measure how much the CFS improved was a two-dimensional Euclidean distance approach where the distance between the CFS and RAP endpoints were calculated and compared to the distance between the corrected CFS and RAP endpoints. If the distance decreased after a correction method was applied, that meant that correction method improved the CFS's performance relative to RAP for that trajectory.

The second metric used was a percentage metric created using the results from the first metric. It calculated out of all the CFS trajectories inputted into the correction method, how much did the CFS endpoint distance change (or improve) relative to the RAP endpoint. If the percentage was 0%, then the CFS trajectory endpoint stayed the same after a correction method was applied and therefore unsuccessful. If the percentage was 100%, then the CFS trajectory endpoint was successful corrected to the exact location of the RAP trajectory endpoint. The numbers presented in the following figures for all three correction methods are the total metric results for all the test trajectories combined.

Correction Method: Linear Regression

Figures 19-21 shows the metric results from the linear regression correction method for each of the quarterly segmented months for unclustered and clustered CFS test trajectory data. The first metric described previously is plotted as the bar graphs and

the second metric described is plotted as the black line graph. Additionally, the CFS trajectory percentage improvement values can be found in Table 1.

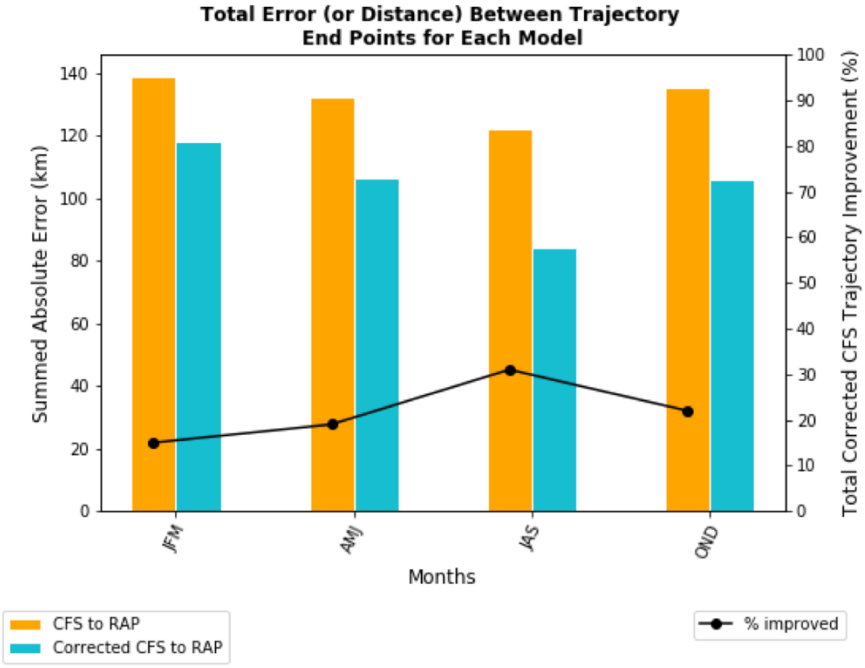

Figure 19. The results for unclustered CFS test trajectories from the linear regression correction method. The improvement metric for the endpoint distance comparison is plotted using bar graphs and its y-axis is on the left. The improvement metric for how much the CFS trajectories improved by is plotted in the black line graph and its y-axis is on the right.

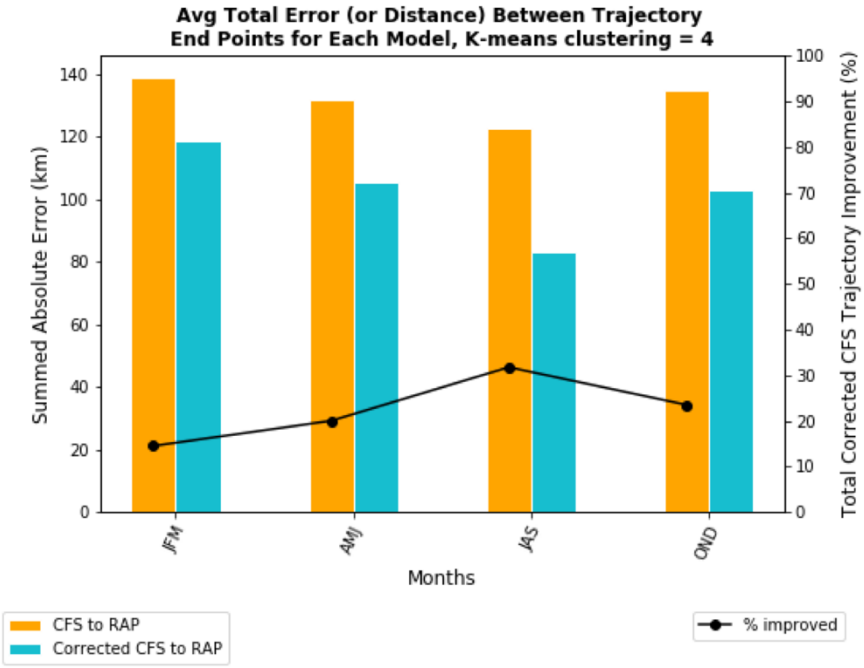

Figure 20. Same information as Figure 19 except the CFS test trajectories are clustered using K-means clustering when $k = 4$.

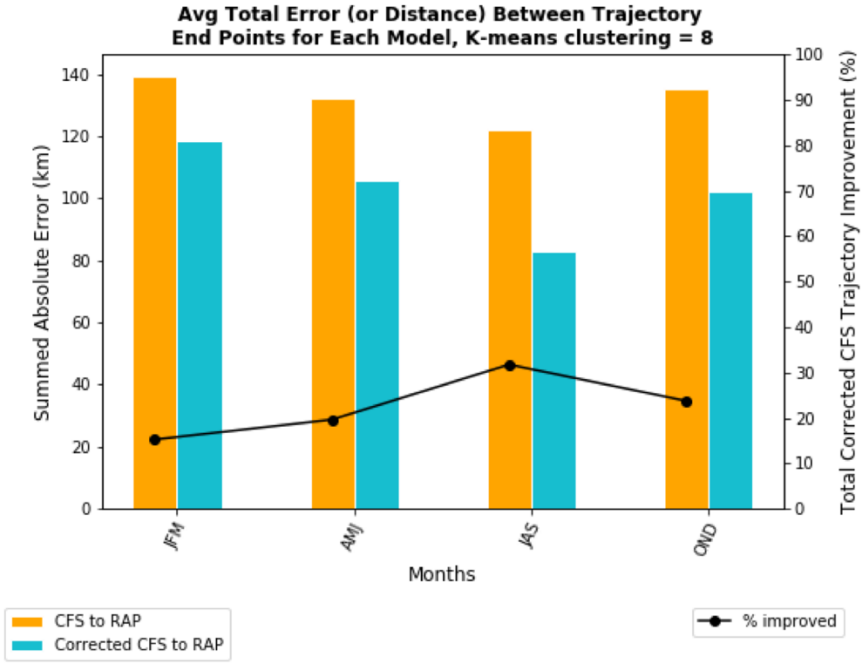

Figure 21. Same information as Figure 19 except the CFS test trajectories are clustered using K-means clustering when $k = 8$.

Table 1.

% Improved After Correction Method	**JFM**	**AMJ**	**JAS**	**OND**	***AVG %***
Unclustered CFS test trajectories	15%	19%	31%	22%	21.75%
Clustered CFS test trajectories, *k = 4*	14%	19%	31%	23%	21.75%
Clustered CFS test trajectories, *k = 8*	14%	20%	31%	23%	22%

Overall, based on results shown in Figures 19-21, linear regression demonstrated it can improve the low-resolution CFS back trajectory performance relative to high-resolution RAP back trajectories. As shown in all three plots and Table 1, the greatest improvement occurs in JAS and the smallest improvement in JFM. Looking back at Figures 13, 15, and 18, JAS is where the mean (or spread) for CFS and RAP endpoints were smallest, and the trajectories traveled the shortest distance. JFM is where the mean (or spread) for the endpoints were largest and trajectories traveled the furthest. Which suggests how effective a correction method might be at improving the CFS is related to how far the back trajectories travel and how much spread there is endpoints for the CFS and RAP back trajectories. While the percentages in Table 1 showed no statistically significant difference, the most improvement for CFS data was observed when it was clustered in eight groups shown in Figure 21. However, computational time could be reduced since clustering the data beforehand did not produce results that were statistically significant.

Correction Method: Bayesian Ridge Regression

Figures 22-24 shows the metric results from the Bayesian ridge regression correction method for each of the quarterly segmented months for unclustered and clustered CFS test trajectory data. Same improvement metrics were used and calculated as described in the previous method and shown in Figures 19-21; and CFS trajectory percentage improvement values can be found in Table 2.

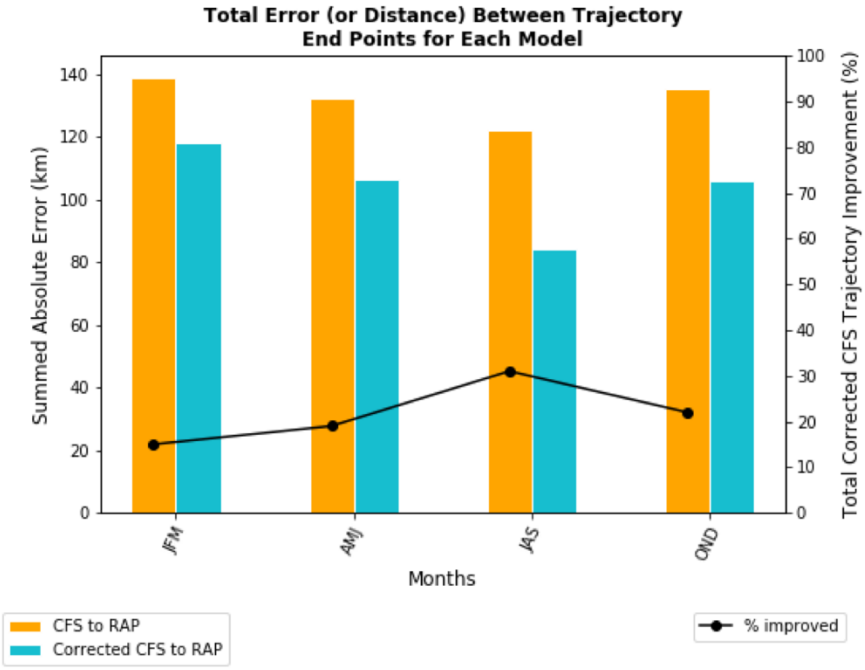

Figure 22. The results for unclustered CFS test trajectories from the Bayesian ridge regression correction method. The improvement metric for the endpoint distance comparison is plotted using bar graphs and its y-axis is on the left. The improvement metric for how much the CFS trajectories improved by is plotted in the black line graph and its y-axis is on the right.

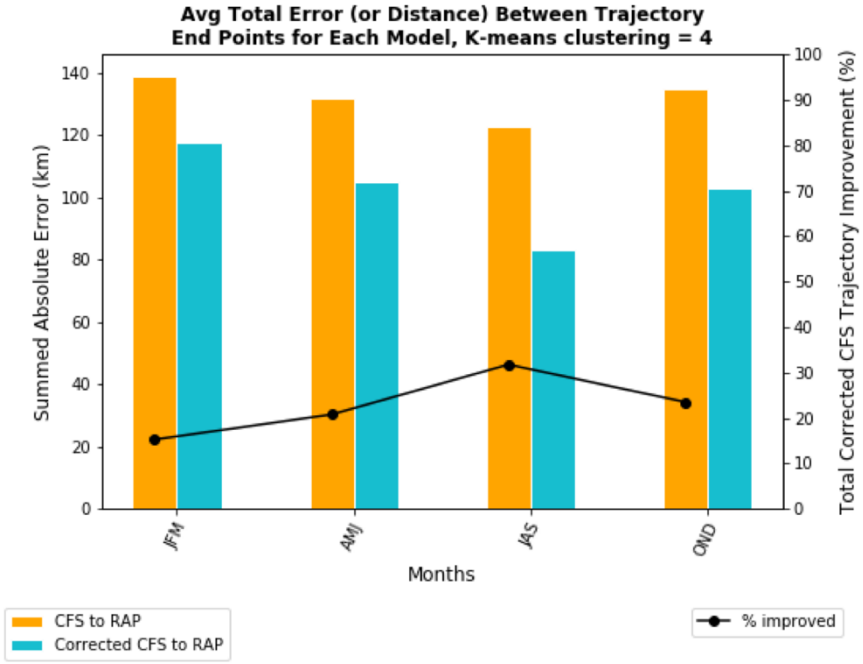

Figure 23. Same information as Figure 22 except the CFS test trajectories are clustered using K-means clustering when *k = 4*.

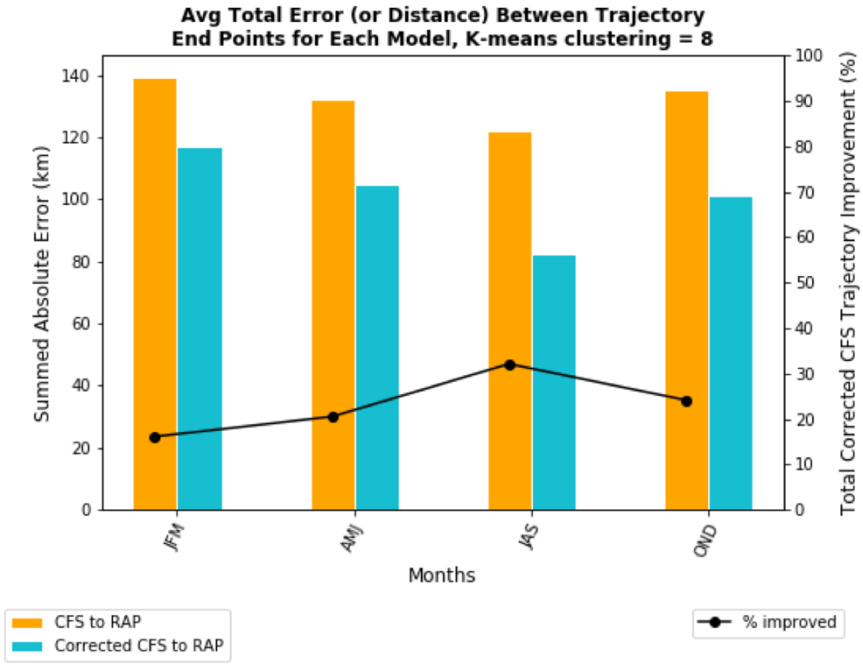

Figure 24. Same information as Figure 22 except the CFS test trajectories are clustered using K-means clustering when *k = 8*.

Table 2.

% Improved After Correction Method	**JFM**	**AMJ**	**JAS**	**OND**	***AVG %***
Unclustered CFS test trajectories	15%	19%	31%	22%	21.75%
Clustered CFS test trajectories, *k = 4*	15%	20%	31%	23%	22.25%
Clustered CFS test trajectories, *k = 8*	16%	20%	32%	23%	22.75%

Comparing Figures 22-24 results to Figures 19-21, Bayesian ridge correction method outperformed linear regression just slightly, although not statistically significant. Once again the biggest improvement in CFS back trajectories occurred in JAS and the smallest in JFM. Furthermore, Bayesian ridge resulted in the highest CFS improvement when the data was clustered into eight groups shown in Figure 24, but percentages results in Table 2 were close enough in value where clustering data before applying the correction method did not greatly impact the results.

Correction Method: Random Forest Regression

Figures 25-27 shows the metric results from the random forest regression correction method for each of the quarterly segmented months for unclustered and clustered CFS test trajectory data. Same improvement metrics were used and calculated as described in the previous linear regression method and shown in Figures 19-21; and CFS trajectory percentage improvement values can be found in Table 3.

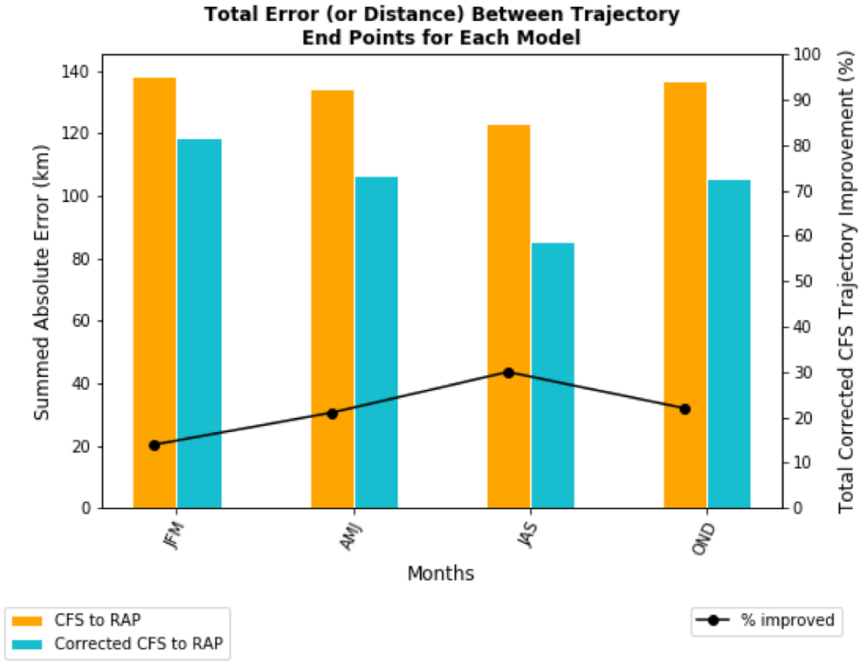

Figure 25. The results for unclustered CFS test trajectories from the random forest regression correction method. The improvement metric for the endpoint distance comparison is plotted using bar graphs and its y-axis is on the left. The improvement metric for how much the CFS trajectories improved by is plotted in the black line graph and its y-axis is on the right.

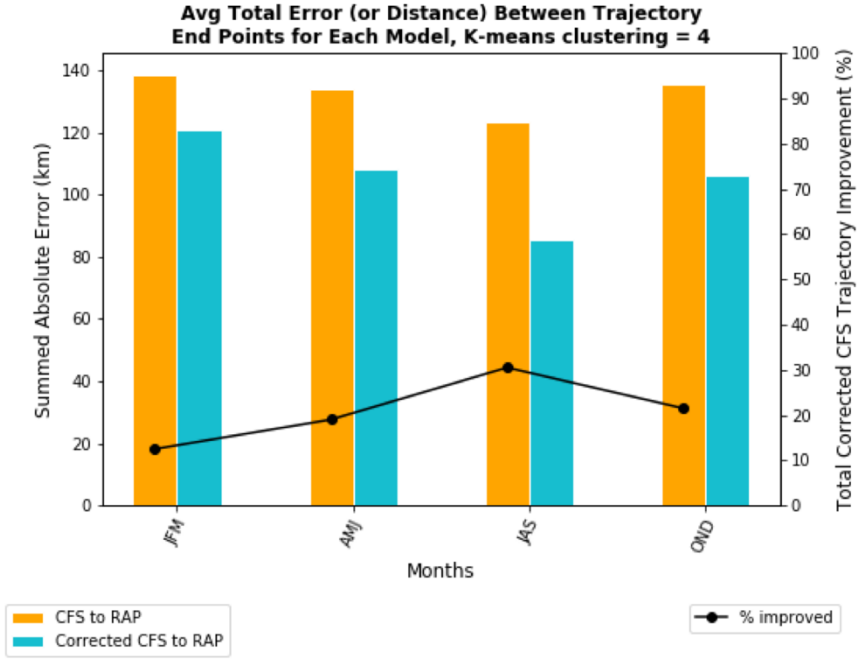

Figure 26. Same information as Figure 25 except the CFS test trajectories are clustered using K-means clustering when $k = 4$.

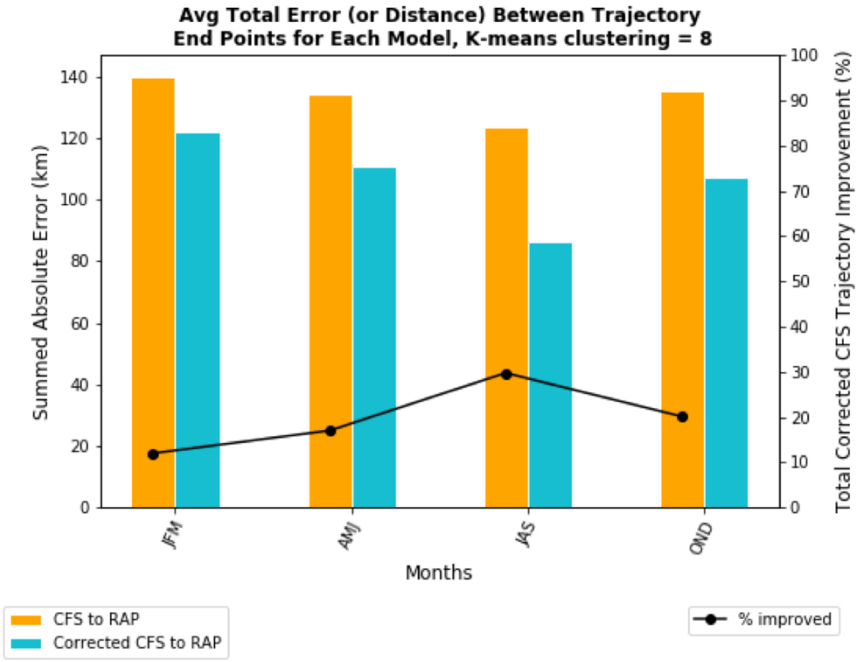

Figure 27. Same information as Figure 25 except the CFS test trajectories are clustered using K-means clustering when *k = 8.*

Table 3.

% Improved After Correction Method	**JFM**	**AMJ**	**JAS**	**OND**	***AVG %***
Unclustered CFS test trajectories	14%	21%	30%	22%	21.75%
Clustered CFS test trajectories, *k = 4*	12%	19%	30%	21%	20.5%
Clustered CFS test trajectories, *k = 8*	12%	18%	30%	20%	20%

Comparing Figures 25-27 results to Figures 19-21 and 22-24, random forest regression correction was the least effective compared to linear and Bayesian ridge regression, which was interesting considering it is the most complicated machine learning technique out of the three this study used. However, this is good news for the user because linear and Bayesian ridge regression requires less computational time to produce results than random forest regression. JAS was once again the biggest improvement for

CFS trajectories and the smallest improvement in JFM. Another difference between this correction method and linear and Bayesian ridge regression is that random forest was the most effective on improving the CFS trajectory performance when it was unclustered shown in Figure 25.

4.4 Correction Method Improvement of Each Cluster

This section covers the improvement of each cluster for Bayesian ridge regression when *k=4* during JAS. The results shown in Figure 28 used the same metrics that were discussed in section 4.3, but now the x-axis is for each cluster rather than the different quarterly segmented months. Figure 29 is the K-means clustering results for JAS from Figure 18 and each cluster is labeled to match the x-axis in Figure 28. CFS trajectory percentage improvement values by cluster can be found in Table 4.

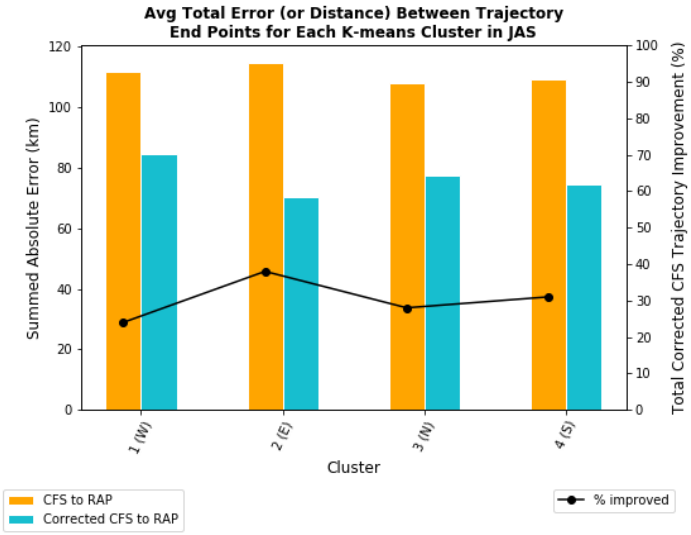

Figure 28. The results for clustered CFS test trajectories (*k=4*) from the Bayesian ridge regression correction method for each cluster. The improvement metric for the endpoint distance comparison is plotted using bar graphs and its y-axis is on the left. The improvement metric for how much the CFS trajectories improved by is plotted in the black line graph and its y-axis is on the right.

Table 4.

% Improved After Correction Method	**Cluster 1 - West**	**Cluster 2 - East**	**Cluster 3 - North**	**Cluster 4 - South**
JAS	24%	38%	28%	31%

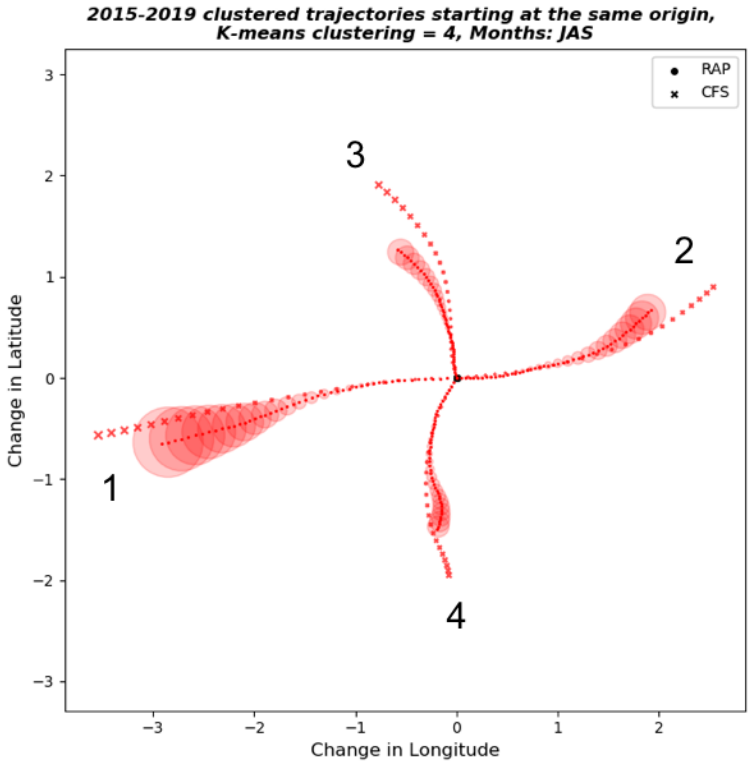

Figure 29. Results from the K-means clustering just for *k* = *4* in JAS, same data as shown in Figure 18. The clusters are numbered to correspond with the x-axis in Figure 28.

Clustered back trajectories that originated east of the origin had the highest percentage improvement. Whereas the worst CFS improvement are clustered back trajectories that originated from the west and traveled the furthest out of the four different clusters, as shown in Figure 29. These results suggest that CFS improvement percentages

are linked with how far the back trajectories traveled before the correction method, in this case Bayesian ridge regression, was applied, reaffirming what previous sections in this chapter indicated. Section 4.3 showed the highest improvement in CFS back trajectories occurred in JAS where trajectories traveled the least amount of distance and displayed the least amount of spread as discussed in section 4.2.

4.5 Correction Method Improvement by Spatial View

The data plotted in Figure 30 took the results from Figure 22 and put them on a map to show there the Bayesian ridge correction method was successful and where it was not for JAS. Areas that are red means the correction method successfully corrected the CFS endpoint to closer to the RAP endpoint and areas that are blue are where the correction method was unsuccessful in correcting the CFS endpoint closer to the RAP.

Overall, the correction method showed positive results over areas of flatter terrain and struggled over regions with changing terrain, such as the Appalachians and the Ozarks, which was expected considering the differences in model performances between the CFS and the RAP in these regions in Figure 11. However, the correction method was surprisingly successful in West Virginia which is still part of the Appalachians. This might be due to the averaged RAP back trajectories exhibiting less spread than the other trajectory points over the Appalachians shown in Figure 11, but this hypothesis will need to be studied further in future work.

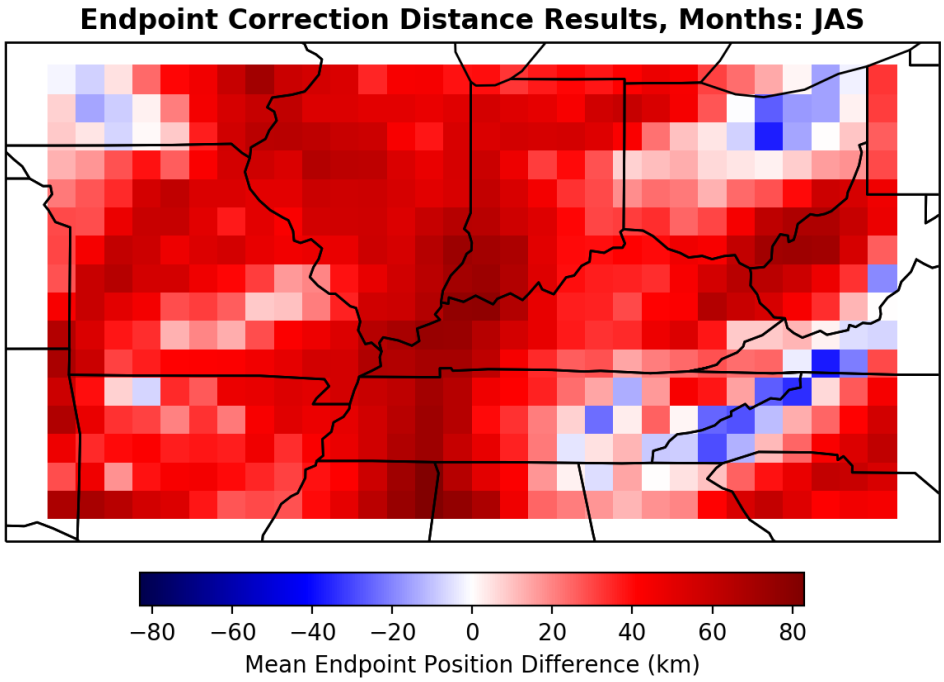

Figure 30. A map of where the Bayesian ridge regression correction method was successful (red) and unsuccessful (blue) averaged for all the testing days in JAS. Positive values mean the corrected CFS endpoints is closer to the RAP endpoints compared to the original CFS endpoint locations. Negative values mean the corrected CFS endpoints stayed in the same location as the original CFS endpoints or moved further from the RAP endpoints.

Areas where the correction method struggled that were unexpected were northern Ohio and southwest Iowa since these regions are relatively little change in terrain compared to the Appalachians and RAP spread in these regions compared to Arkansas or the Appalachians appears to be smaller in Figure 11. The current hypothesis is southwest Iowa and northern Ohio are impacted by more frontal systems increasing the RAP trajectory distance making it harder for the CFS to correct to, but this will need to be investigated more in a future study.

The next few images will compare the results from the case study points, referenced in Figure 5, to gain some insight as to possibly why the correction method was unsuccessful in southwest Iowa compared to northwest Alabama. It is important to note that these case study results are intended to be a proof of concept for this methodology and a more robust case study needs to be done in a future study.

The data plotted in Figure 31 shows the distance between the first point and end point of the averaged 50 trajectories for each day in JAS in the testing data set. The original distance between CFS starting and endpoints compared to the RAP are shown in orange, and distance between the CFS starting point and corrected CFS endpoint compared to the RAP is shown in blue. The black line is plotted where RAP equals CFS/corrected CFS. The closer the points are to that line show when the models performed roughly the same. Results for the northwest Alabama point is on the left and southwest Iowa on the right.

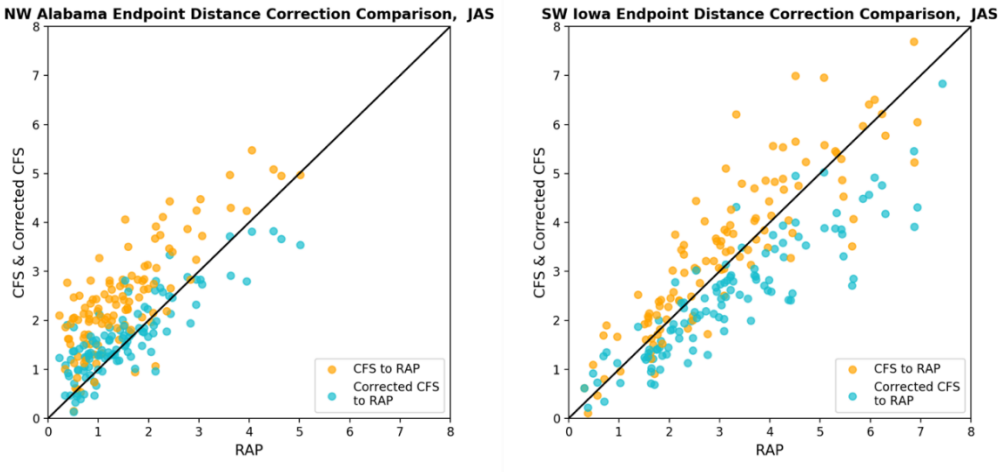

Figure 31. The scatter plots are plotting the distance between the first point and end point of the averaged 50 trajectories for each day in JAS in the testing data set. The black line plotted represents when the RAP equals the CFS/corrected CFS or when the models performed the same.

When comparing the two scatter plots, they reveal the Iowa point back trajectories traveled further than the Alabama trajectories. This suggests that earlier hypothesis of more frontal systems impacting Iowa causing the back trajectories to travel further might be the reason as to why the Bayesian ridge correction method was not as successful for this region. In the Iowa scatter plot on the right, there appears to be a mirror-like flip between the orange and the blue points where the blue corrected CFS points are not any closer to the black line than the orange points. Meaning that after the correction method was applied, the RAP trajectories traveled than the corrected CFS endpoint resulting in worse results post-correction.

The Alabama scatter plot on the left on Figure 31, on the other hand, shows most of the orange points above the black line indicating before the correction method was tested, CFS back trajectories traveled further than the RAP. Post-correction method, the scatter plot shows a tighter clustering of blue points near and on the black line illustrating the corrected CFS endpoint performed more closely to the RAP. When comparing the blue points for the Alabama point back trajectories to Iowa's, it reaffirms the results plotted on Figure 30 where Bayesian ridge regression correction method was more successful in northwest Alabama than southwest Iowa.

The next figure shows an example of how this methodology might look operationally to the user trying to correct the low-resolution CFS back trajectories to the high-resolution RAP's output. These two plots again compare the results from northwest Alabama on the left, a successful example, to southwest Iowa on the right, an unsuccessful example. Shown in the plots are the 50 back trajectories released for each

model, just like Figures 6 and 7, except for just one starting point. RAP back trajectories are plotted in cyan and CFS in red, and CFS corrected endpoint, calculated based on the averaged CFS and RAP endpoint locations, is the black star. In the Alabama example, the plot shows the corrected CFS endpoint is closer to the RAP endpoints than the uncorrected CFS, whereas in the Iowa example, the corrected CFS endpoint is actually a worse solution than the uncorrected CFS back trajectories compared to the RAP.

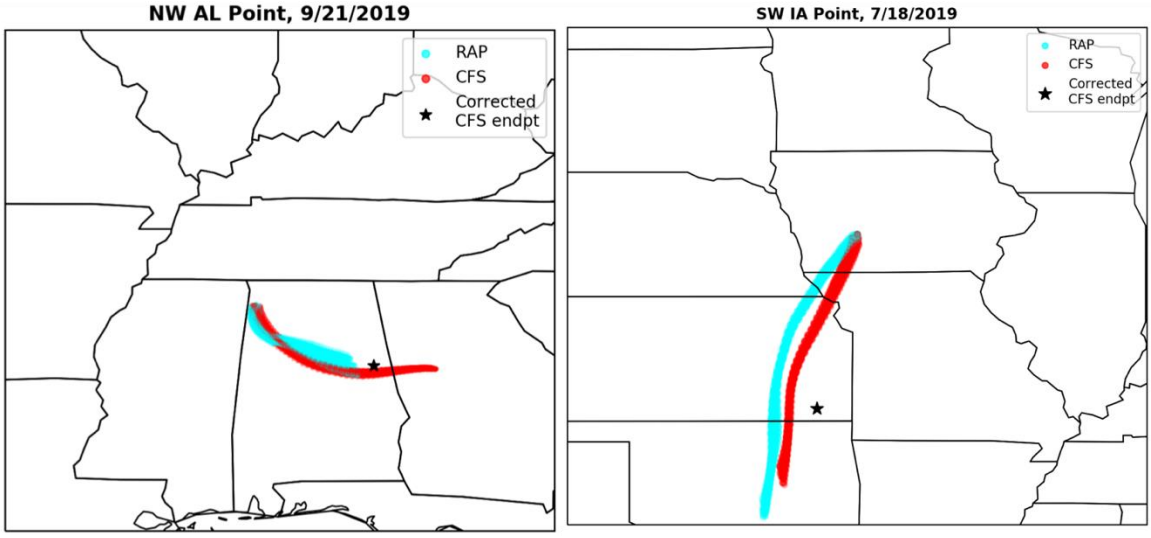

Figure 32. Example of how this methodology would be implemented operationally. Plotted are 24-hour back trajectories for northwest Alabama (left) and southwest Iowa (right). Each starting point shows 50 trajectories released for each model, RAP (cyan) and CFS (red); and CFS corrected endpoint, calculated based on the averaged CFS and RAP endpoint locations, is the black star.

V. Discussion and Conclusions

5.1 Summary of Results

Five years of short-term (24-hour) back trajectory data from 18Z to 18Z for CFS (global, low-resolution model) and RAP (regional, high-resolution model) were analyzed using K-means clustering, an unsupervised machine learning technique. The back-trajectories were released from a starting grid of 17 by 31 points (see Figure 5) from 34 to 42 degrees latitude and -100 to -85 degrees longitude. Each starting point had 50 trajectories released from the surface with random location error of +/- 5 km and +/- 5 hPa. It was assumed that, due to the higher horizontal and vertical resolution, the RAP trajectory results were more accurate based on recent studies such as Mittermaier (2014). To get a better understanding of how each model performed throughout different times of the year, all five years of the trajectory data were split up into four parts, three months each: (1) JFM, (2) AMJ, (3) JAS, and (4) OND.

Clustering was used to evaluate the performance for different meteorological regimes, or flow patterns. The number of *k* clusters used was 4 and 8, which was determined by calculating the data's silhouette score. The results from both clustering sets showed that the further the trajectories traveled in 24 hours, the larger the spread (or uncertainty) for those clusters (see Figure 11). Additionally, the RAP clustered trajectories displayed more spread than the CFS. This is likely due to the RAP model being more sensitive to terrain and the random location error applied since it is a higher resolution model than the CFS. Figures 9, 10, and 12 all confirmed RAP trajectories had the highest spread occurred in JFM and OND, whereas JAS exhibited the smallest spread.

Three different supervised correction methods were tested on the unclustered and clustered CFS data with the goal to see how effective the CFS model could be trained to perform similar to the RAP. These methods were linear regression, Bayesian ridge regression, and random forest regression. Overall all three methods resulted in the CFS's performance improving relative to the RAP and while not statistically significant, the Bayesian ridge regression just outperformed linear regression and random forest regression. Additionally, all three correction methods were most effective during JAS when the RAP and CFS trajectory spread was lowest and back trajectories traveled the shortest distance in 24-hours. The correction methods performed the worst during JFM when the RAP and CFS trajectory spread was higher due to back trajectories traveling further in 24-hours. Finally, when considering all the clusters together, the improvement percentages for unclustered vs clustered data were not statistically significant either, meaning these correction methods can be applied to data without needing to use K-means clustering. Meaning that, when considering all flow regimes together, CFS improvement results can be achieved by using unclustered data and simpler machine learning algorithms, such as linear and Bayesian ridge regression, which decreases computational time for the user.

Next the study analyzed Bayesian ridge regression improvements results for clustered back trajectories (*k=4)* in JAS by each individual cluster (Figure 28). The results showed the highest improvement in CFS clustered back trajectories were when they originated east of the origin, which meant they traveled the shortest distance and exhibited the smallest amount of spread. Whereas the worst improvement results

occurred in clustered CFS back trajectories originating to the west, and those traveled the furthest in 24 hours and had the largest spread.

Finally, the JAS results shown in Figure 22 were plotted on a map to examine what areas the Bayesian ridge correction was successful at improving the CFS endpoints locations relative to the RAP and where it was unsuccessful (Figure 30). Overall, the correction method was successful over areas of flatter terrain and less so over regions with changing terrain, which was not surprising considering how much spread the RAP back trajectories exhibited in those regions in Figure 11. An unexpected area where the correction was unsuccessful was southwest Iowa where RAP back trajectory spread was smaller compared to the Appalachians. When endpoint data from this region to a successful point in northwest Alabama, the Iowa back trajectories traveled further which suggests that might be the reason as to why the correction was not as successful. However, this last section of the results was meant to serve as a proof of concept for the methodology tested; a larger case study needs to be completed in a future study.

5.2 Limitations

One limitation of this research was during the process of retrieving the archived RAP data, more than one version of the model was downloaded in order to obtain five years' worth of data. This meant that the CFS version 2 model data was being compared to many versions of the RAP model data, rather than just one version of RAP. This could have impacted the percentage numbers in Tables 1-3. Another limitation with the CFS and RAP data was that each model did not have data archived for every day from 2015-

2019, and in order to accurately analyze both model's performances, back trajectories were only plotted on days when both RAP and CFS had data archived.

5.3 Future Work

This study just focused on correcting the mean endpoint latitude and longitude. The next step would be to investigate correcting the difference in the amount of spread between the CFS and RAP as well. Another area of future work would be to examine advanced machine learning techniques using neural networks, such as long-short term memory (LSTM) and convolutional LSTM neural networks to examine how those machine learning techniques would improve the CFS relative to the RAP. Additionally, incorporating the back trajectories change in vertical direction into the clustering and machine learning algorithms to provide insight if this variable improves how well the CFS data can be improved when running machine learning algorithms. Another focus for future work is to re-run the RAP back trajectories at one-degree grid spacing rather than 13 km horizontal grid spacing, which is the same grid spacing as CFS, to examine a change grid spacing as it relates to back trajectory performance.

Finally, this study focused mostly on testing different methodologies rather than application in an expansive case study such as testing the results on a STE case study using daily EPA PM 2.5 and ozone station data from the western plains area. This is being proposed for two reasons: to have a clear distinction between "clean" and polluted areas that is more difficult to make to the east but also to bring the terrain of the Rocky Mountains to the west into the examination.

5.4 Conclusion

Overall the results from this study demonstrated that the CFS's back trajectory performance can be improved using machine learning. This study displayed that, in certain situations, computational costs can be reduced for operational forecasters by utilizing global weather model data for trajectory calculations versus running a custom, high resolution limited area model. This method worked especially well during the summer months, when winds are weaker and trajectories travel a shorter distance. The research accomplished in this study provides the framework methodology to apply to different machine learning algorithms or broader case studies to better understand relationships between different atmospheric regimes and implementing post-processing techniques to global models.

Bibliography

Benjamin, Stanley G., Stephen S. Weygandt, John M. Brown, Ming Hu, Curtis R. Alexander, Tatiana G. Smirnova, Joseph B. Olson, et al. 2016. "A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh." *Monthly Weather Review* 144 (4): 1669–94. https://doi.org/10.1175/MWR-D-15-0242.1.

Bergstra, James, and Yoshua Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research* 13: 281–305.

Bieringer, Paul E., George S. Young, Luna M. Rodriguez, Andrew J. Annunzio, Francois Vandenberghe, and Sue Ellen Haupt. 2017. "Paradigms and Commonalities in Atmospheric Source Term Estimation Methods." *Atmospheric Environment* 156: 102–12. https://doi.org/10.1016/j.atmosenv.2017.02.011.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC.

Boehmke, Brad, and Brandon Greenwell. 2019. "Random Forests." In *Hands-On Machine Learning with R*, 203–19. Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377-11.

Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. 2013. "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm." *Expert Systems with Applications* 40 (1): 200–210. https://doi.org/10.1016/j.eswa.2012.07.021.

Cervone, G., P. Franzese, Y. Ezber, and Z. Boybeyi. 2008. "Risk Assessment of Atmospheric Emissions Using Machine Learning." *Natural Hazards and Earth System Science* 8 (5): 991–1000. https://doi.org/10.5194/nhess-8-991-2008.

Draxler, Roland R. 2006. "The Use of Global and Mesoscale Meteorological Model Data to Predict the Transport and Dispersion of Tracer Plumes over Washington, D.C." *Weather and Forecasting* 21 (3): 383–94. https://doi.org/10.1175/WAF926.1.

Efendi, Achmad, and Effrihan. 2017. *A Simulation Study on Bayesian Ridge Regression Models for Several Collinearity Levels*. *AIP Conference Proceedings*. Vol. 1913. https://doi.org/10.1063/1.5016665.

Feng, Rui, Hui jun Zheng, An ran Zhang, Chong Huang, Han Gao, and Yu cheng Ma. 2019. "Unveiling Tropospheric Ozone by the Traditional Atmospheric Model and Machine Learning, and Their Comparison:A Case Study in Hangzhou, China." *Environmental Pollution* 252: 366–78. https://doi.org/10.1016/j.envpol.2019.05.101.

Gang Su, Xiao, and Xin Yan. 2009. *Linear Regression Analysis: Theory and Computing*. *Linear Regression Analysis: Theory and Computing*. https://doi.org/10.1142/6986.

Ivatt, Peter D., and Mathew J. Evans. 2020. "Improving the Prediction of an Atmospheric Chemistry Transport Model Using Gradient-Boosted Regression Trees." *Atmospheric Chemistry and Physics* 20 (13): 8063–82. https://doi.org/10.5194/acp-20-8063-2020.

Jain, Anil. 2010. "Data Clustering: 50 Years beyond K-Means." *Pattern Recognition Letters* 31 (8): 651–66. https://doi.org/10.1016/j.patrec.2009.09.011.

Jain, Anwiti, Anand Rajavat, and Rupali Bhartiya. 2012. "Design, Analysis and

Implementation of Modified K-Mean Algorithm for Large Data-Set to Increase Scalability and Efficiency." *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*, 627–31. https://doi.org/10.1109/CICN.2012.95.

Livingston, Frederick. 2005. "Implementation of Breiman's Random Forest Machine Learning Algorithm." *Machine Learning Journal Paper*, 1–13.

Michel, Clio, and Gwendal Rivière. 2011. "The Link between Rossby Wave Breakings and Weather Regime Transitions." *Journal of the Atmospheric Sciences* 68 (8): 1730–48. https://doi.org/10.1175/2011JAS3635.1.

Mittermaier, Marion P. 2014. "A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites." *Weather and Forecasting* 29 (2): 185–204. https://doi.org/10.1175/WAF-D-12-00075.1.

NOAA-ARL. n.d. "Air Resources Laboratory - HYSPLIT." Accessed February 24, 2021. https://www.arl.noaa.gov/hysplit/hysplit/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Thirion Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, et al. 2011. "Scikit-Learn: Machine Learning in Python Fabian." *Journal of Machine Learning Research* 12: 2825–30.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (C): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Saha, Suranjana, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, et al. 2014. "The NCEP Climate Forecast System Version 2." *Journal of Climate* 27 (6): 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1.

Stein, A. F., R. R. Draxler, G. D. Rolph, B. J.B. Stunder, M. D. Cohen, and F. Ngan. 2015. "NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System." *Bulletin of the American Meteorological Society* 96 (12): 2059–77. https://doi.org/10.1175/BAMS-D-14-00110.1.

Vrac, Mathieu, and Pascal Yiou. 2010. "Weather Regimes Designed for Local Precipitation Modeling: Application to the Mediterranean Basin." *Journal of Geophysical Research Atmospheres* 115 (12): 1–23. https://doi.org/10.1029/2009JD012871.

Yerramilli, Anjaneyulu, Challa Venkata Srinivas, Hari Prasad Dasari, Francis Tuluri, Loren D. White, Julius M. Baham, John H. Young, et al. 2009. "Simulation of Atmospheric Dispersion of Elevated Releases from Point Sources in Mississippi Gulf Coast with Different Meteorological Data." *International Journal of Environmental Research and Public Health* 6 (3): 1055–74. https://doi.org/10.3390/ijerph6031055.

Zoellick, Casey L. 2019. "Source Term Estimation of Atmospheric Pollutants Using an Ensemble of HYSPLIT Concentration Simulations." *Theses and Dissertations*, no. 2210: 111. https://scholar.afit.edu/etd/2210.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
22-03-2021	Master's Thesis	August 2019 – March 2021

TITLE AND SUBTITLE

CORRECTION OF BACK TRAJECTORIES UTILIZING MACHINE LEARNING

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Gjermo Morrison, Britta F., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENP-MS-21-M-119

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Technical Applications Center
1020 South Patrick Dr.
Patrick AFB, FL 32935
321-854-8931 astrid.suarez-mullins@us.af.mil
ATTN: Dr. Astrid Suarez-Mullins

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RHIQ (example)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this work was to analyze 24-hour back trajectory performance from a global, low-resolution weather model compared to a high-resolution limited area weather model in particular meteorological regimes, or flow patterns using K-means clustering, an unsupervised machine learning technique. The duration of this study was from 2015-2019 for the contiguous United States (CONUS). Three different machine learning algorithms were tested to study the utility of these methods improving the performance of the CFS relative to the performance of the RAP. The aforementioned machine learning techniques are linear regression, Bayesian ridge regression, and random forest regression. These results mean reducing computational time for the user. Additionally, the greatest improvement of CFS values occurred in July, August, and September.

15. SUBJECT TERMS

Machine learning, K-Means, clustering, atmospheric regimes, atmospheric transport and dispersion modeling, source term estimation, trajectories

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Lt Col Robert C. Tournay, AFIT/ENP
U	**U**	**U**	**UU**	19	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4743 robert.tournay@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18