DECODING MIDWEST JUNE PM2.5 EVENTS:
A SELF-ORGANIZING MAP APPROACH TO METEOROLOGICAL ANALYSIS

Victor Geiser
Advisor: Tsengel Nergui
LADCO Intern - Summer 2024
Prepared for Lake Michigan Air Directors Consortium (LADCO)

ABSTRACT: The Midwest is a land-locked mid-latitude geographical setting where complex atmospheric processes take place in conjunction with local emissions and transported air pollutants. Periodically, upwind wildland and prescribed fire smoke is transported into the region and results in unhealthy concentrations of fine particulate matter ($PM_{2.5}$) and at the surface. Comparisons of the meteorological conditions associated with typical high pollution days, versus those of fire smoke influenced days, are useful to forecasters and air quality planners.

To better understand the meteorological setting and pollutant transport pathways bringing fire smoke into the region, LADCO applied a spatial classification technique called a Self-Organizing Map (SOM) to daily average $PM_{2.5}$ concentrations using the 3-km resolution High Resolution Rapid Refresh (HRRRv4) reanalysis dataset for 2019-2023. The objective of the analysis is to identify the primary features of the physical and dynamical atmospheric conditions associated with air pollution episodes with and without the influence of smoke.

We will present the results of our SOM analysis of pollution episodes caused by wildland fires originating in the southwestern US and southwestern Canada. We used the SOM to identify the synoptic scale meteorological conditions and the anticipated increases in $PM_{2.5}$ during fire events. In addition, we investigated a key aspect of whether the long-range transported smoke aloft reached the surface. Vertical atmospheric characteristics such as wind shear, stability, and 24 changes in the geopotential height and temperature for fire-influence SOM nodes, highlight key upper-air features for vertical mixing and indicate whether air masses ascend or descend along the transport path between the fire smoke source and receptor monitors.

Our study offers two practical applications for air quality forecasters. First, using SOM to identify the weather patterns associated with typical high-pollution days provides historical data for similar-day analysis for exceptional event applications. Secondly, the identified synoptic weather patterns linked to fire smoke-influenced days provide insights into the expected increases in $PM_{2.5}$ concentrations due to fire smoke in the Midwest.

1. INTRODUCTION

Periodically, wildland fire smoke is transported into the Midwest and results in unhealthy concentrations of fine particulate matter with a diameter less than 2.5 micrometers ($PM_{2.5}$). Elevated $PM_{2.5}$ concentrations have been associated with a wide range of human health hazards and also affect many meteorological and chemical processes in the atmosphere. Moreover, the US Midwest's central placement within the North American continent and the Great Lakes makess it a common place for a diverse range of meteorological and chemical process to occur and converge. The identification of common meteorological conditions associated with significantly above normal $PM_{2.5}$ concentrations is applicable to both the fields of air quality and meteorology.

In this study, we examine the above idea using a Self-Organizing Map (SOM). Self-Organizing Maps were originally proposed in (Kohonen 1982) and are a type of artificial neural network that aim to find lower dimensional relationships in high dimensionality data whilst preserving the original structure (topology) of its input data. Unlike its more traditional counterparts, such as principal component analysis, it makes no underlying assumptions about relationships within the input data, such as linear relationships. SOMs have been applied within a wide variety of fields such as genomics (Törönen et al. 1999), astrophysics (Carrasco Kind and Brunner 2014), and economics (Deboeck and Kohonen 2013), and are commonly used tools in the fields of data mining (Vesanto and Alhoniemi 2000) and non-linear manifold learning and representation (Forest et al. 2021). More recently SOMs have been applied within the physical sciences as well and have been proven to be successful when applied to air quality and meteorology data (Hrust et al. 2009) (Hewitson and Crane 2002).

1.1 GOALS AND MOTIVATIONS

This project is motivated by persistent $PM_{2.5}$ episodes in the Great Lakes region and a need to further understand the nature of their origins and impacts. Additionally, given the importance of atmospheric aerosols on the earth's global radiation budget, insights gleaned from understanding where the pollutants occur and what effect they have on the global environment is relevant within a changing climate.

Leading to our overall research question:

How can Self Organizing Maps (SOMs) be used to identify meso-scale meteorological conditions associated with high PM2.5 and fire smoke impacted conditions in the LADCO region?

The goals laid out for this project are as follows:

1. To enrich the conceptual model regarding high concentrations of $PM_{2.5}$ in the Midwest by incorporating meteorological settings identified through the Self-Organizing Map (SOM) method.
2. To establish a basis for determining whether the overhead smoke observed by satellites descended to the surface and impacted concentrations of $PM_{2.5}$ at surface monitors.
3. Compare the synoptic weather conditions in the Midwest during air pollution episodes with and without the influence of wildfire smoke.

1.2 THE OBSERVATION OF RESULTS BY CONSIDERING A CASE STUDY

Affirmations complementing our SOM analysis can be observed clearly when following a $PM_{2.5}$ event that occurred over the LADCO region on June 25-30, 2023. The primary reason as to why this event was so anomalous was due to the impacts caused by wildfire smoke originating in Ontario and Quebec Canada. **Figure 1** illustrates the meteorological conditions during the transition between the first and second phases of the event that were dominated by an initial low-pressure system that aided in transporting polluted air into the US Midwest, followed by a high-pressure system event that led to stagnation conditions and greatly above average $PM_{2.5}$ and impacts.

Although not the primary topic of this study, throughout the remainder of this report we will provide extra visual elements considering this event. This is not only to display how the results of the SOM are observable when applied in a real-world context, but also as a quick reference to certain known conditions within our input data that will point to positive signatures regarding our SOM's functionality and performance.

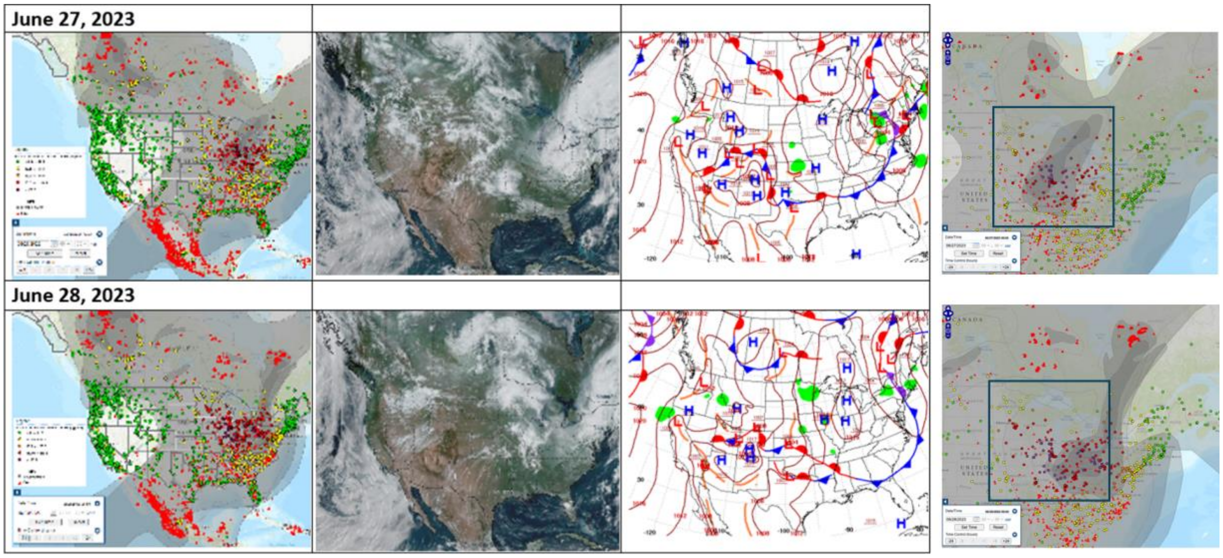


Figure 1 The meteorological conditions surrounding June 27th and 28th 2023 during which "very unhealthy" air quality was observed.

2. METHODOLOGY

The primary method contained within this study is the self-organizing maps algorithm itself. While there are a multitude of implementations for self-organizing maps written in many different programming languages, the implementation used in this study is the "MiniSOM" implementation. MiniSOM is an open-source and purely pythonic implementation of self-organizing maps that is available on GitHub. It gets its name from "minimalistic SOM" as its only dependency is the NumPy library, and it is generally used for small to medium sized datasets.

The self-organizing maps algorithm seeks to produce a low-dimensional (usually two-dimensional) representation of the input space while preserving the topological properties of the original data.

2.2 INPUT DATA

Meteorological data: This study contains data from a variety of sources, however the primary data source that is used when training the SOM is daily meteorological reanalysis data, which is a blend of the 3-km resolution HRRRv4 (High Resolution Rapid Refresh) surface reanalysis and, 12-km resolution NAM (North American Model) reanalysis data. The dataset contains data for all June days between 2019 and 2023. The meteorological dataset has a spatial resolution (grid spacing) of 4km and is using the conditions at 18:00 UTC (12:00pm CST). The files were originally output in NetCDF format and are read into python through use of the Xarray package in python. Each meteorological file also contained the necessary projection information that allowed the data to be plotted on a 420 latitude by 444 longitude extent on a Lambert Conformal Conic projection. **Figure 2** displays the extent of our data ranges from a latitude and longitude of (34.163, -100.316) in the SW corner to (50.644, -78.027) in the NE corner. **Figure 3** provides an example visualization of one of our data variables, relative humidity at the 500hPa level.

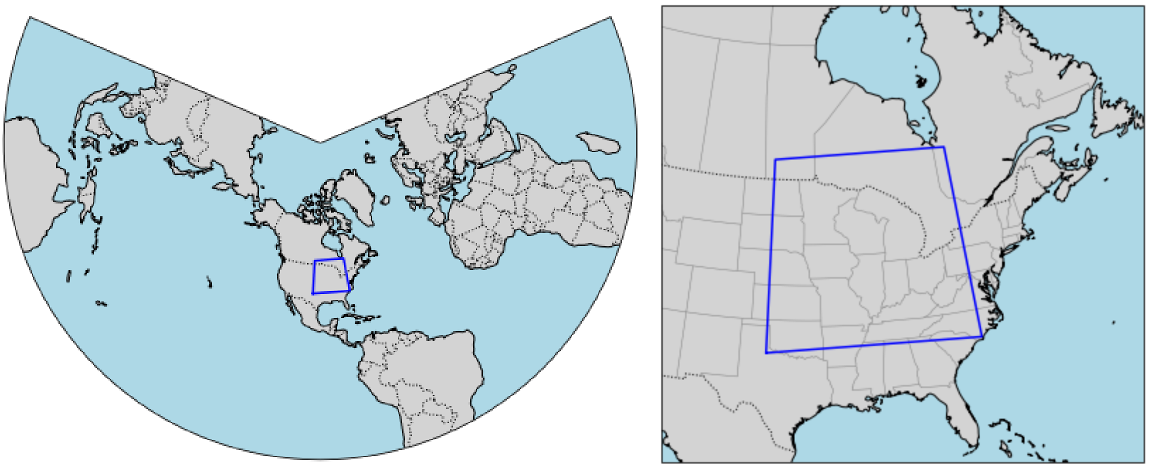


Figure 2 The (LADCO) region of interest for this study on a Lambert Conformal Conic projection.

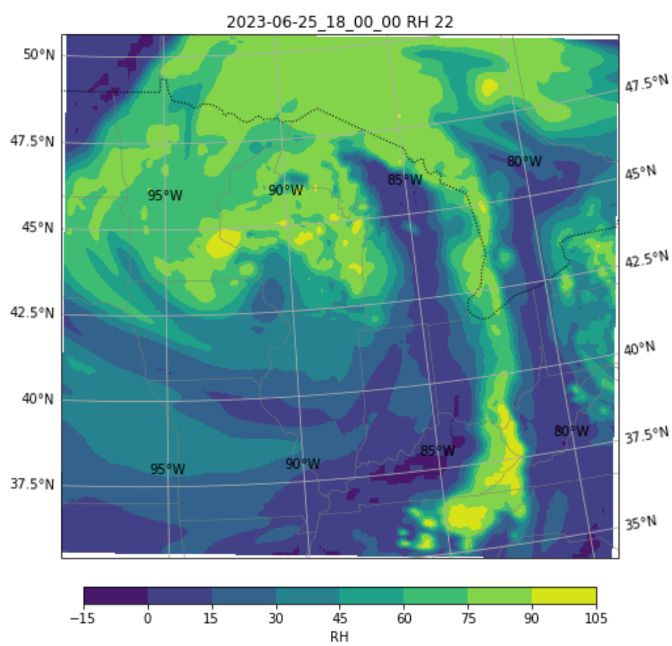


Figure 3 500hPa Relative Humidity (%) for 06-25-2023 at 18:00 UTC.

Although the meteorological data used for this study contains over 115 variables, only the variables that are used as inputs into the SOM will be discussed in this report. Mentioned here briefly are the names of these variables, their available vertical levels (model levels), their units, and associated abbreviations within the dataset. Motivations for why these variables were selected for SOM analysis can be found in section 4.

1. "PMSL" – Pressure at mean sea level (surface only). Units: Pascals (Pa)
2. "RH" – Relative humidity (model levels 1-40). Units: Percentage (%)
3. "TT" – Temperature (model levels 1-40). Units: Degrees Kelvin (K)
4. "UU" – Horizontal "U" wind component (model levels 1-40). Units: meters per second (m/s)
5. "VV" – Vertical "V" wind component (model levels 1-40). Units: meters per second (m/s)
6. "GHT" – Geopotential height (model levels 1-40). Units: meters (m)

Air quality data: In addition to the HRRR meteorological data, two other datasets were used for analysis purposes. The first is a tabular dataset containing observed $PM_{2.5}$ and data from the US EPA Air Quality System (AQS). We calculated additional SOM node metrics based off four columns contained within this dataset:

1. "value" – An observed $PM_{2.5}$ concentration in $\mu g/m^3$.
2. "std_log_value" (or $PM_{2.5}$ "anomaly") – A standardized value of the measured $PM_{2.5}$ concentration. Standardization (i.e., normalization) was done using the monthly mean and standard deviation of the log-transformed measured values at a monitor over the 2019-2022 period. This standardized value (i.e., anomaly) provides a measure for how much $PM_{2.5}$ concentration deviates from its typical mean.
3. "HMS_binary" – A binary flag variable (either 0 or 1) that determined if overhead smoke was identified at the location of a monitor through a satellite-driven product called the Hazard Mapping System (HMS).
4. "res_1sigma_std_log_value" (or "res1") – The residual value of $PM_{2.5}$ concentrations above and below 1 standard deviation, hich indicates how much the measure value was beyond the typically observed values a monitor.

Figure 4 shows a visualization of the above data variables for the $PM_{2.5}$ dataset: "value" (on the y axis) and "std_log_value" (on the x axis) with HMS_binary outlines. Values for "res_1sigma_std_log_value" would then be the data on the right side of the vertical dashed black line and with red outlines.

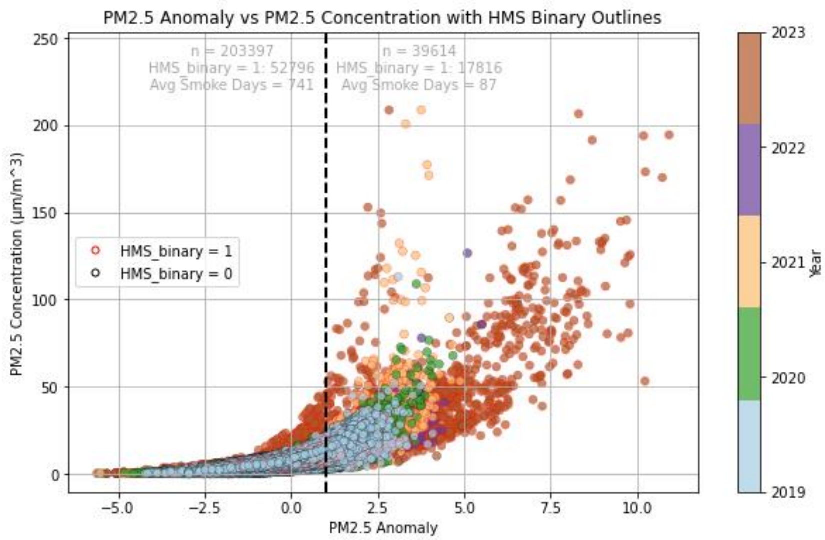


Figure 4 Scatter plot of $PM_{2.5}$ concentration vs. its standardized anomaly with outlines for overhead smoke.

The last dataset that we used during the final stages of this project is a "krigged" (spatially interpolated) $PM_{2.5}$ dataset. This dataset is an interpolated product based off the $PM_{2.5}$ ground sensor network. **Figure 5** is an example visualization of the krigged $PM_{2.5}$ dataset.

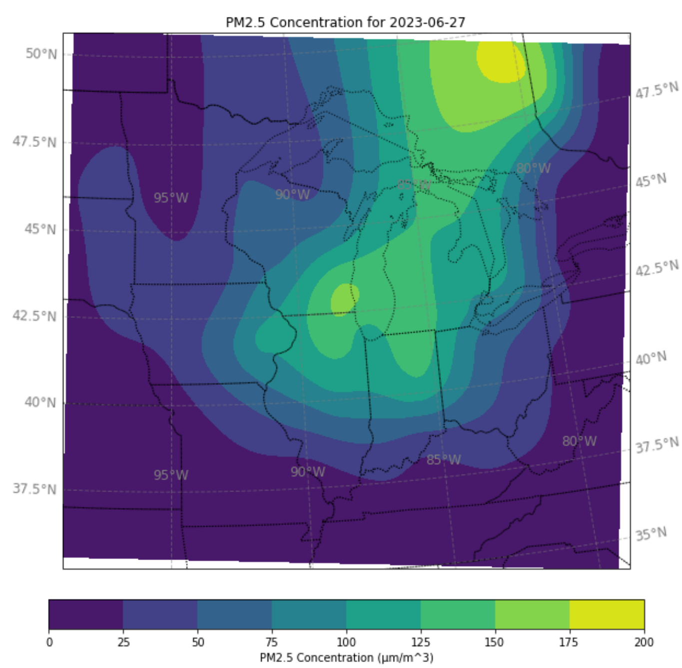


Figure 5 Krigged $PM_{2.5}$ field for 06-27-2023 displaying PM2.5 transport into the LADCO region.

Although mentioned here for completeness, the krigged $PM_{2.5}$ dataset is not used until **section 5**.

To ensure compatibility with MiniSOM two preprocessing steps needed to be applied to the data:

1. Vectorization – Due to the spatial nature of the data they needed to be vectorized in order to be input into MiniSOM.
2. Standard Scaling – Standard scaling (through Sci-kit learn) is a technique that scales the data between (-1 and 1) where each variable has a mean of "0" and a standard deviation of "1"

Vectorization adds to the dimensionality of our data substantially and, as we will see in our analysis, this will have lasting effects as far as our quantitative metrics and furthermore in our interpretive analysis. However, because no clear alternatives to vectorization currently exist, and vectorizing our data still allows for spatial patterns in our data to be represented, this is the standard approach. This raises the question: why not use an initial dimensionality reduction technique when applying preprocessing steps? The answer to this questions lies in the interpretive need to recreate and visualize our data. Further study could potentially look at applying techniques such as t-distributed Stochastic Neighbor Embedding or Uniform Manifold Approximation and Projection as a further preprocessing step, given the nature of the non-linear relationships we are attempting to explore, however this would make the final visualizations produced from the SOM less meaningful.

As a final data preprocessing step, a standardized scaler was applied to the meteorological data as preparation for input into the SOM. This study used the StandardScaler method built into Sci-Kit Learn, which scales the data to a range between -1 and 1 and a mean of 0. To account for the varying scales of the meteorological data, a standard scaler was applied individually to all variables.

2.2 DESCRIPTION OF THE SELF ORGANIZING MAPS ALGORITHM

Although descriptions of the Self Organizing Maps algorithm exist across many sources within the literature, a brief summary adapted from (Kohonen 1982) and (Hulle and Marc 2012) will be presented here.

Step 1: Initialization

The first step within the SOM algorithm happens when a grid of neurons (also called nodes) is initialized. Each neuron has a weight vector of the same dimensionality as the input data.

Step 2: Training algorithm

The self-organizing maps training algorithm has two main components:

1. The best matching unit (BMU)

The BMU is the neuron whose weight vector is closest to the input vector in terms of Euclidean distance. This can be mathematically expressed as:

$$c = argmin_i \ \parallel \boldsymbol{x}(t) - \boldsymbol{w_i}(t) \parallel$$

Where:

- c is the index of the BMU.

- $\boldsymbol{x}(t)$ is the input vector at time t.
- $\boldsymbol{w_i}(t)$ is the weight of the vector of the i-th neuron at time t.
- ‖·‖ denotes the Euclidean distance.

2. The weight update

The weight vectors of the BMU and its neighboring neurons are updated to move closer to the input vector. The update rule is:

$$\mathrm{w}_i(t+1) = \boldsymbol{w_i}(t) + \theta_{i,c}(t) \cdot \alpha(t) \cdot (\boldsymbol{x}(t) - \boldsymbol{w_i}(t))$$

Where:

- $\boldsymbol{w_i}(t)$ is the weight of the vector of the i-th neuron at time $\boldsymbol{t}$.
- $\boldsymbol{w_i}(t+1)$ is the updated weight vector
- $\alpha(t)$ is the learning rate, which decreases over time
- $\theta_{i,c}(t)$ is the neighborhood function centered on the BMU c, which determines the influence of the BMU on its neighbors. In our case this has a gaussian form:

$$\theta_{i,c}(t) = \exp\left(-\frac{\| r_i - r_c \|^2}{2\rho(t)^2}\right)$$

Where:

  - r_i and r_c are the positions of the i-th neuron and the BMU c int the grid, respectively
  - $\rho(t)$ is the neighborhood radius that also decreases over time
  - ‖·‖ denotes the Euclidean distance.

The neighborhood function and learning curve are time-dependent functions that decrease with time (the number of iterations) to ensure convergence. The neighborhood function is also controlled by a parameter σ (currently set to "1") however the way these functions behave while decreasing and the value of sigma are up to the user.

The current SOM implementation uses "linear_decay_to_zero" applied to the learning rate:

$$\alpha(t) = \alpha_0\left(1 - \frac{t}{T}\right)$$

Where:

- α_0 is the initial learning rate.
- t is the current iteration number
- T is the total number of iterations

And "asymptotic_decay" applied to the neighborhood function:

$$\rho(t) = \frac{\rho_0}{1 + \frac{t}{\sigma}}$$

Where:

- ρ_0 is the initial neighborhood radius
- t is the current iteration number
- σ is the time constraint that controls the rate of decay

2.3 SOM HYPERPARAMETERS AND CONFIGURATION

Within the primary Self-Organizing Maps algorithm established above, there also different configurations that can be achieved by tweaking a SOM’s hyperparameters. The hyperparameters for the LADCO SOM are as follows:

- `som_size = (3,5)`
- `sigma = 1`
- `learning_rate = .3`
- `ngb_function = 'gaussian’`
- `decay_function = 'linear_decay_to_zero'`
- `sigma_decay_function = 'asymptotic_decay’`
- `init = 'random'`
- `train = 'random'`
- `iterations = 200`
- `topology = 'hexagonal'`
- `activation_distance = 'euclidean'`
- `random_state = ‘64’`

Notable deviations from the default parameters include a hexagonal topology which allows our nodes to have more neighbors as opposed to the default rectangular topology. We have a slightly lower than normal learning rate that resulted in better performance via iterative experimentation and our learning rate decay function is set to decrease linearly as opposed to the standard asymptotic decay which resulted in better clustering via the clustering metrics as described in **section 2.4**. We set the number of iterations at 200 because more iterations did not result in significantly improved performance. It should be noted that, since our learning curve visualization in **Figure 8** uses quantization error, and since our data are highly dimensional, this curve appears in a slightly atypical fashion as opposed SOMs that may occur elsewhere within the literature. More about LADCO SOM’s learning curve will be discussed in **section 3**.

Most importantly within the topic of SOM hyperparameters is the SOM size, which controls the number of output neurons within the SOM. The determination of this hyperparameter is often crucial to the functionality of the SOM and is often a trade-off involving capturing more general trends, sensitivity to outliers, and having enough nodes to capture the more nuanced and informative trends within the data (Hulle and Marc 2012). In the case of the LADCO SOM our SOM size appears to be limited primarily by the number of samples currently ingestible withing the workflow. We consider all June days between 2019 and 2023 leaving us with 149 samples (1 day of the reanalysis dataset is missing to generate due to an incomplete HRRR run for that time period). If SOM size increases in an attempt to capture harder to detect relationships within the data, we begin to observe nodes that have an activation response (or the number of samples from the input data that get classified as having that pattern, or activated that particular node during the training process) of 0. Due to this, determination the optimal SOM size for the LADCO SOM is an area for potential enhancement. However, if different climatological periods are considered or perhaps expanded upon in the future this may come naturally given the current implementation.

2.4 SUMMARY OF SOM AND NODE METRICS

In addition to the primary output of this study, which comes in the form of a visualization of the weights of LADCO SOM itself, there will also occur above or below each node, secondary node statistics calculated from averaged variables for all nodes that are included in the activation response for a particular node. By running each input vector through the SOM after the training period has completed, we are able to generate a map of which input vectors are considered to "match" that output node.

The secondary parameters visualized alongside the weights for each variable node are:

1. "Node (x,y)" – The node's position within the SOM hexagonal grid
2. "n = …" – The number of samples mapped to that particular node
3. "Smoke Days" – The number of identified days that met the condition mentioned in the "Res1" variable explanation
4. "Avg PM" – Average measured $PM_{2.5}$ concentrations over all monitors within the domain for days classified for a particular node
5. "Avg PM anom" – Similar to the Avg PM variable, but for standardized anomalies (std_log_value variable).
6. "Avg Res1PM" – A node average of the "res1" variable

2.5 VERTIAL PROFILE GENERATION PROCUDURE

Since our meteorological data are model derived and all vertical model levels are present within the meteorological data used for this study, we are able to produce an additional secondary (or tertiary) node analysis in the form of a visualization of a node's averaged vertical atmospheric profile. The profiles are point soundings in one location (although an extended explanation about developing the functionality further will occur in **section 5**) for each model level temperature, relative humidity, and u and v wind vector components. Visualization of the profile is handled by the MetPy library in Python.

With these visualizations of the vertical profile, we also display calculated environmental statistics based on the profiles generated for each node.

These statistics include:

1. "Node (x,y)" – Same as in **section 2.4**
2. "n = …" – Same as in **section 2.4**
3. "SH01" – The 0-1km environmental shear vector
4. "SH06 – The 0-6km environmental shear vector
5. "850hPa ω" – Vertical velocity at 850hPa
6. "CIN" – Convective Inhibition
7. "CAPE" – Convective Available Potential Energy
8. "#TI" – Number of temperature inversions (2°C / 100hPa)
9. "850-950hPa avg temp" – Average temperature difference between the 850 hPa and 950hPa levels
10. "Avg PM2.5 Anomaly" – Same as in **section 2.4**

3. RESULTS

This chapter includes the following three sections:

- **Section 3.1** includes SOM diagnostic plots that support the primary analysis.
- **Section 3.2** covers the primary results of the study, the visualization of the weights of LADCO SOM, and the conclusions that can be reached as a result.
- **Section 3.3** will present the vertical profile results described in **section 2.5**

3.1 SOM DIAGNOSTIC PLOTS

Presented in **Figure 6** and **Figure 7** are the LADCO SOM distance matrix (or u-matrix) with sum scaling and mean scaling respectively. The distance matrix is used to measure the distances between the nodes in a SOM grid. This distance can be scaled in various ways, two of which are a mean (average) scaled distance matrix and a sum scaled distance matrix. The sum scaled distance matrix, visualized in **Figure 6**, represents the total sum of distances between (the vector values of) a particular node and all other nodes in the SOM. The sum scaled distance matrix indicates the overall quality and separation of clusters in the SOM and it can also be used to inform where potential boundaries between clusters appear within the SOM. All weight visualizations in section 3.2 will use the mean scaled distance matrix as a background color, or "frame color" for reference.

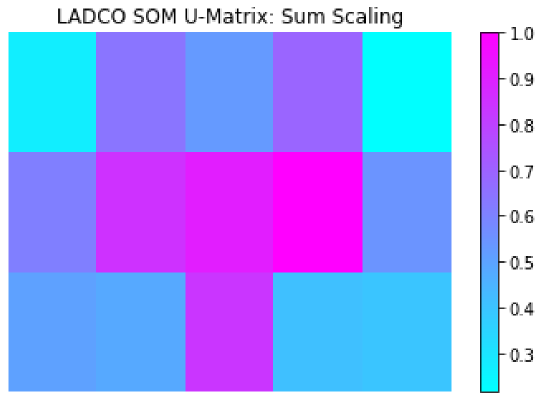


Figure 6 LADCO SOM distance matrix (sum scaled)

The mean scaled distance matrix, visualized in **Figure 7,** is a representation of the average distance between the prototype vectors of nodes in the SOM grid. This figure provides a measure of how smoothly the input space is represented by the SOM and it can visualize the average separation between clusters. Nodes with a high value within the mean scaled distance matrix represent nodes that are on average further apart from their surrounding neighbors. Higher values in this figure indicate nodes pattern that are significantly different in structure (in our case meteorological conditions) than its surrounding neighbors.

The middle sections of the LADCO SOM distance matrix have a higher total distance. Unfortunately, the high dimensionality of our data appears to affect the summed distance matrix quite a bit. The middle sections of the SOM (remember the hexagonal topology) with the more pinkish and purple values inform us that our central nodes appear to sit along a boundary between clusters (a result that is also apparent throughout the project). While we can read this result from the summed distance matrix, given the relatively small size of our SOM in general, it could have been inferred that nodes that are closer in weight values to one another would occur in a region of the SOM where we expect to see more intermediate meteorological regimes.

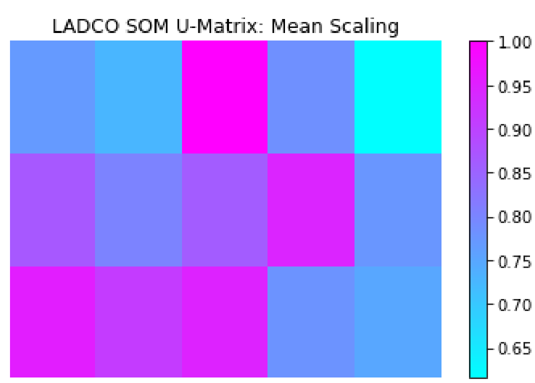


Figure 7 LADCO SOM distance matrix (mean scaled)

As much as the sum scaled distance matrix is informative (although predictable), the mean scaled matrix is even more so, especially in light of our highly dimensional data. **Figure 7**'s purpose is threefold, it displays nodes that are comparatively different in relation to their neighbors (higher average separation), it displays the relative smoothness of our SOM, and it illustrates how well the topology of the original data is being preserved within the SOM. Interpretable from **Figure 7** is a primary cluster of nodes with higher average separation in the lower left hand corner and throughout the middle of the SOM. With these nodes all having values in the upper ranges for mean scaled distance we can understand that (A) this local region of the SOM contains a wide range of types of cases (meteorological setups representing the placement and orientation of high pressure in this case) that are distinct from one another, and (B) looking globally at our entire mean scaled distance matrix, we can see that although topological relationships are being preserved, there still exist regions of the SOM that are better than others. Visualizations of the quantization error (QE) and topographic error (TE) learning curve for the LADCO SOM in **Figure 8** and **Figure 9** will examine possible reasons and justifications for this observation.

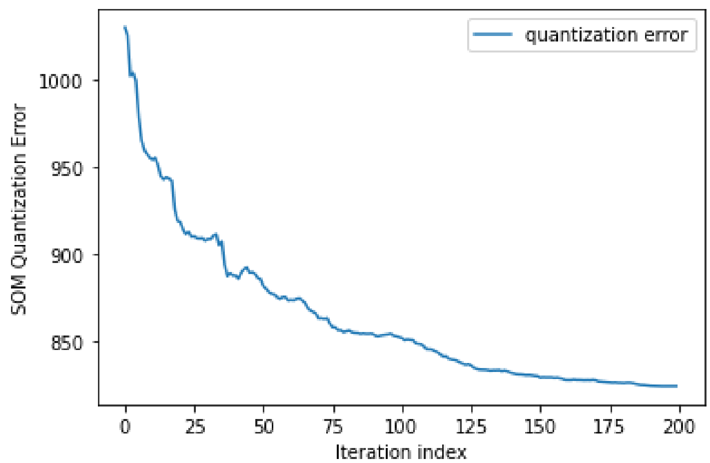


Figure 8 Learning curve showing quantization error decreasing with 200 iterations

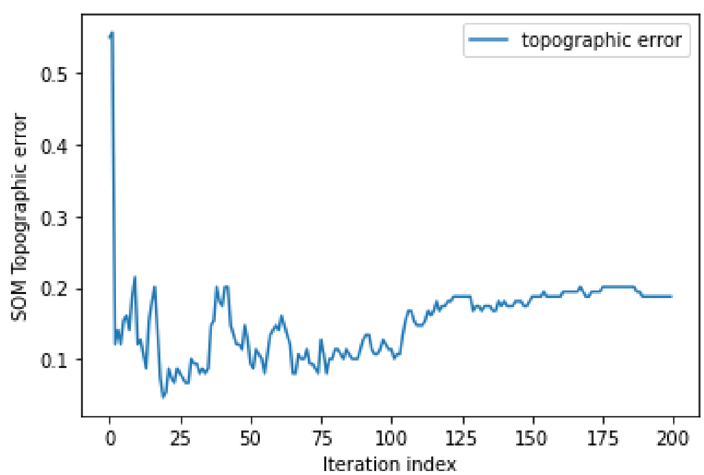


Figure 9 Learning curve showing topographic error decreasing but then increasing slightly with number of iterations

A very simple observation that can be reached from briefly examining the learning curves for the LADCO SOM is that both the QE and TE learning curves display atypical behavior. For the QE curve our final QE error is 824, which is much greater than the close to zero value normally observed for typical SOM applications. This is explainable keeping in mind how QE is calculated in the first place and considering the dimensionality of our data.

Quantization error is calculated by the formula:

$$QE = \frac{1}{N}\sum_{i=1}^{N} \parallel \boldsymbol{x}_i - \boldsymbol{w}_{BMU(i)} \parallel$$

Where:

- N is the number of samples.
- $\boldsymbol{x}_i$ is the $\boldsymbol{i}$-th sample in the dataset.
- $\boldsymbol{w}_{BMU(i)}$ is the weight vector of the Best Matching Unit (BMU) for the $\boldsymbol{i}$-th sample.
- $\|\cdot\|$ denotes the Euclidean distance.

As noted in the quantization error formula, QE is primarily a Euclidean distance measure. For the LADCO SOM the QE converges to 824 as a side effect of our data's dimensionality where our SOM is actually performing quite well, however because each of our input variables has a vectorized length of 186,480, combined with the face that one sample in the dataset has 6 variables, each sample has a vectorized length of 1,118,880. Considering our high dimensionality, this means that even small residuals between an input vector and its matching BMU are propagated in relation to the data's dimensionality, when in reality, a QE of 824 means that 824/1,118,880 = 0.00073… the average error of individual elements within the input array and its BMU is comparatively very small.

Turning our attention to the topographic error, we notice an unusual trend by iteration 25, in that our TE begins to increase and then level out with increasing iterations. Topographic error is calculated by finding the first BMU and the second BMU, and a sample for which these two nodes are not adjacent counts as an error. The topographic error is given by the total number of errors divided by the total number of samples. A similar trend was observed in Forest et al. (2021)where the phenomenon of increasing TE is correctly explained: "Topographic error shows the trade-off between self-organization … and the resulting clustering quality" who further went on to mention how "A practitioner could thus choose to use an early stopping strategy … but it would harm the quality of the clustering." Essentially, the increase in TE of the LADCO SOM is related to an increase in clustering quality, where, by to some extent ignoring the data's topological relationships, better clustering can be achieved.

While this result may initially be concerning, given the day-to-day variability in mesoscale meteorological patterns, it is expected that any given sample may not be similar enough to its second BMU to count as a topographic error, the outcome of which can be observed in both the mean scaled distance matrix in **Figure 7** and the explicit TE learning curve in **Figure 9**. Furthermore, the phenomenon of increasing TE may also be in part due to the high dimensionality of the input data diminishing the overall utility and representation of Euclidian distances in our data space, as seen by the QE learning curve in **Figure 8**.

3.2 LADCO SOM WEIGHT VISUALIZATIONS

The primary results from the LADCO SOM present what a classification of meteorological regimes looks like to a self-organizing map. The resulting clusters (nodes) are then compared using the metrics presented in **section 2.4**. This section will serve primarily to introduce the weight visualizations, where further discussion is prompted based on these results in **section 4**. Results will be presented in the order they were introduced in **section 2.2**. Sections 3.2.1 through 3.2.5 will cover the results of the purely meteorological LADCO SOM, in which other variations of this primary LADCO SOM being introduced later.

3.2.1 Variable: Mean Sea Level Pressure (surface level)

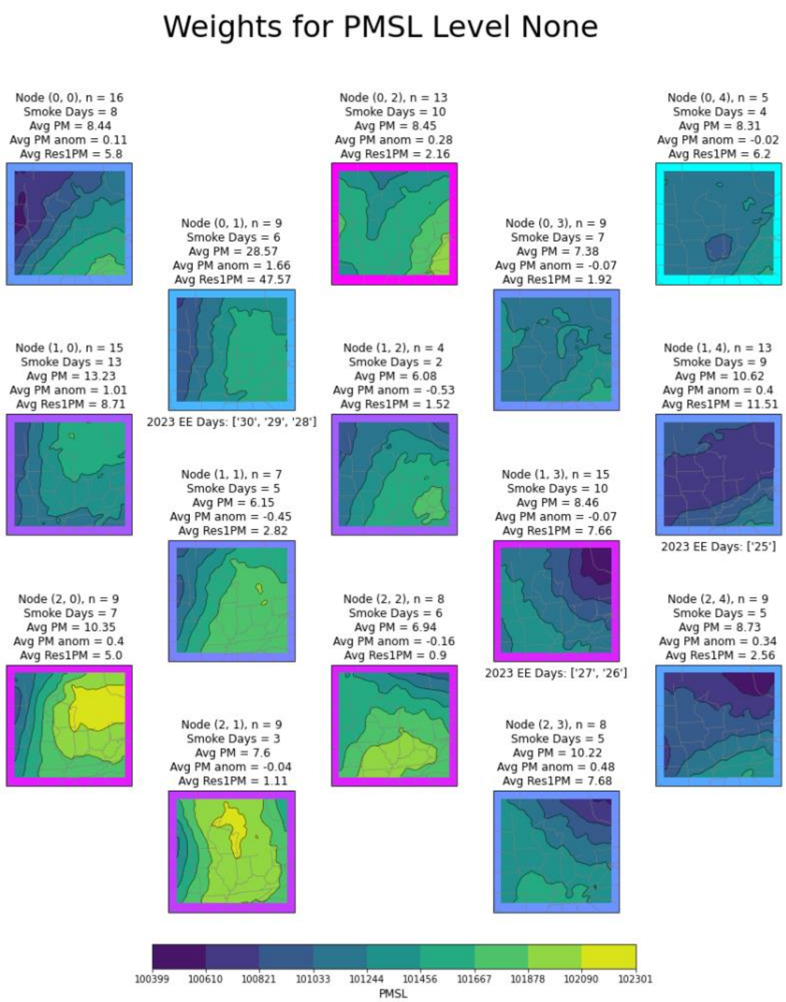


Figure 10 The weights for Mean Sea Level Pressure within LADCO SOM in Pascals.

Mean Sea Level Pressure (MSLP) is primarily characterized by either high pressure (in the lower left corner) or low pressure (in the lower right corner). In between transition states with no dominant pressure pattern occurring within the middle of the SOM and in the upper right and left most corners. The MSLP weights present an overall view of conditions at the surface and will be referred to frequently through the remainder of this report. Node (0,1) is dominated by a weak high-pressure pattern and node (1,3) is dominated by a strong northeastern low-pressure system. As evident by the caption below, each node presents two very different meteorological pressure patterns that occurred within the 2023 Canadian wildfire event introduced in **section 1**, which will hereby be referred to as the 2023 EE (exceptional event).

The MSLP weights for within LADCO SOM suggest that higher $PM_{2.5}$ anomalies can be expected during high-pressure patterns less than 1016hPa. These nodes are associated with stagnation conditions and less dynamic motion within the atmosphere, and diminished advective processes transporting emissions or wildfire smoke out of the LADCO region. The MSLP field in nodes (2,3) and (2,4) describe a situation where the LADCO region is in between a high pressure system to the south (with presumably anticyclonic motion) and a low pressure system to the north (with presumably cyclonic motion). These two flows will enhance transport within the region which indicates that in low pressure dominated nodes, environments conducive to westerly transport are associated with stronger transport into the region and elevated $PM_{2.5}$ impacts occur as a result. MSLP was chosen as an input variable as MSLP is one of the most recognizable patterns in forecasting, and it is a surface variable that is normalized to account for the Appalachian Mountains.

3.2.2 Variable: 500 hPa level Relative Humidity

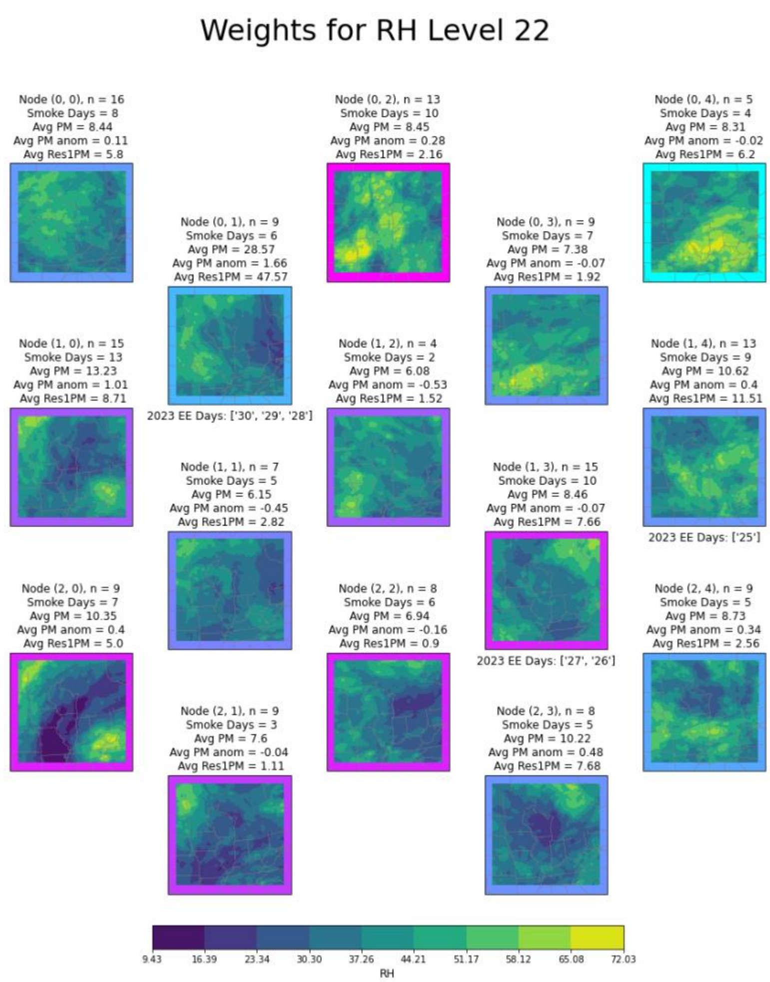


Figure 11 The weights for 500 hPa relative humidity within LADCO SOM in percentage (%).

The 500hPa relative humidity (RH) is a weak predictor for the LADCO SOM for June. Taking the average RH value for a node and comparing it to variation in the $PM_{2.5}$ field yields a Spearman Correlation of -0.067 and a P-value of 0.81. Despite RH being a weak predictor, some notable trends are still interpretable from the RH variable. Mainly within our high pressure dominated nodes we see definitive dry streaks at the mid-levels, and in the opposite corner we also see sharp moisture gradients within nodes that have near 0 average $PM_{2.5}$ anomaly. This may have not been noticeable from the MSLP weights as there is most likely some dilution of the field due to averaging within the pressure field, but these sharp RH gradients may be indicative of higher cloud cover over the LADCO region which would be associated in this case with frontal passage and storms, in turn leading to increase wet deposition and lower $PM_{2.5}$ anomaly. The choice to include relative humidity at the 500hPa level is motivated by variations in the mid-level moisture profile, variations that can become plainly visible in the vertical profile plots presented in **section 3.3**.

3.2.3 Variable: Surface Level Temperature

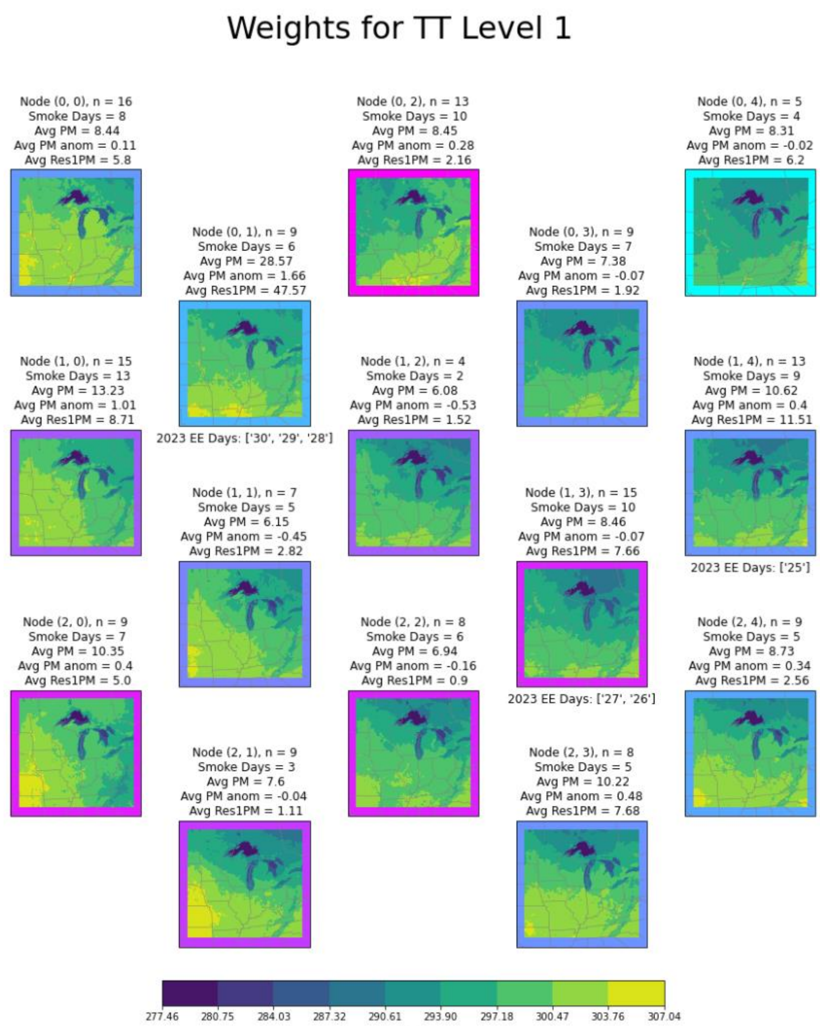
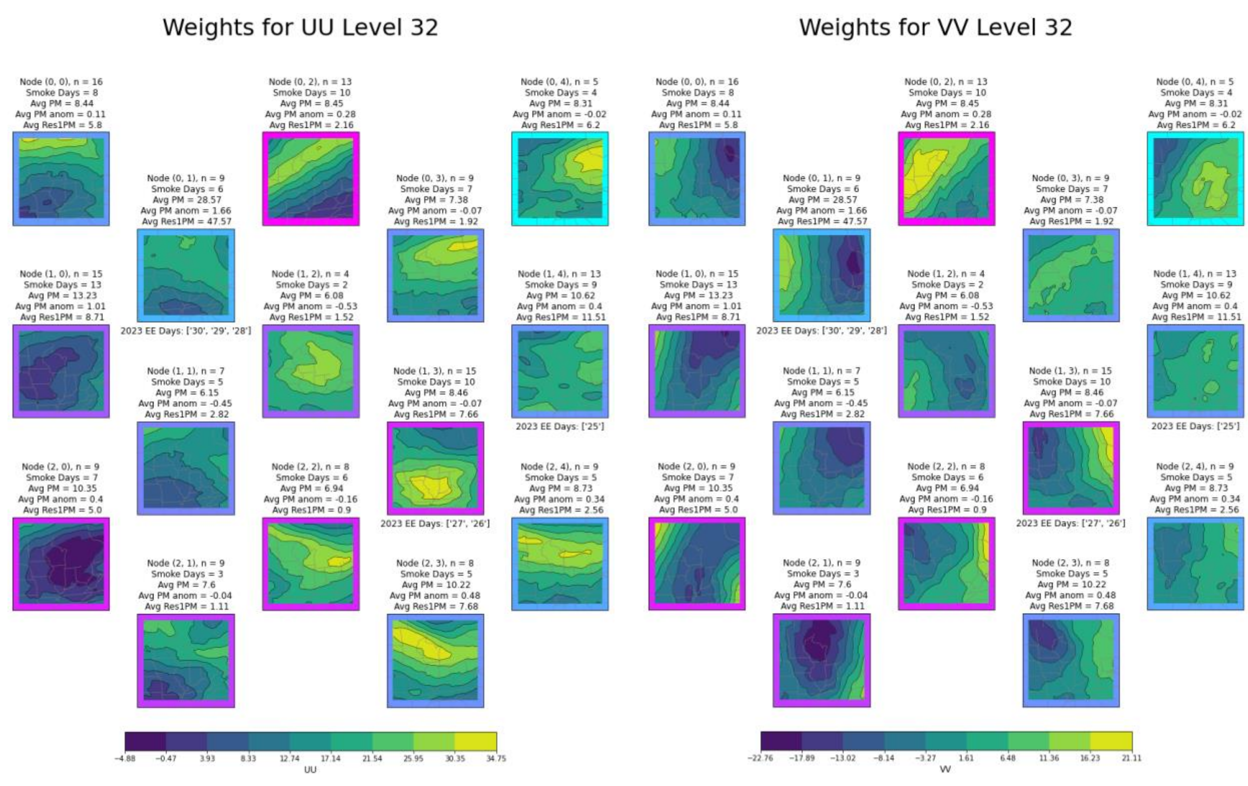


Figure 12 The weights for surface temperature within LADCO SOM in Kelvin (K).

Two primary surface temperature patterns emerge based on **Figure 12.** The left side of the SOM is characterized by warmer southern temperatures extending northward, and the right side is characterized by cooler northern temperatures extending southward. The statistical relationship between node averaged surface temperatures and $PM_{2.5}$ concentration is slightly stronger with a Spearman correlation 0.44 and a p-value of 0.099 indicating the relationship is slightly positive (higher temperatures correlate with higher $PM_{2.5}$ anomaly) and it is statistically significant at the 10% level. In addition, standard temperature tends where colder temperatures appear northward, and warmer temperatures occur southward. Nodes such as (0,2), which present a more unique temperature setup with a conveyor belt of warm air extending as far north as southern Michigan, associated with anticyclonic motion from the south, and node (0,4), which has a protrusion of colder air extending into Missouri, a trend consistent with previous associations of node (0,4) with the fronts passing over the LADCO region. Although low variability surface temperature makes trends within LADCO SOM harder to visualize, it was selected as input variable because surface temperature is one of the most commonly measured parameters in meteorology, and both $PM_{2.5}$ impacts, and human impacts can be better understood considering it.

3.2.4 Variable: 250hPa level U and V wind components



Figures 13a-b The weights for U and V wind components within LADCO SOM in meters per second (m/s).

The U **Figure 13a** and V **Figure13b** wind vectors are incorporated into the SOM as separate variables, however, to improve readability these are commonly combined into the total wind speed magnitude as seen in **Figure 16**. The 250hPa level is informative when diagnosing warm season jet streak patterns (as opposed to the more traditional 300hPa level in the cool season). **Figure 14** and **Figure 15** present a very abbreviated summary of some primary concepts of jet streak motion and dynamics from figures adapted from (Keyser and Shapiro 1986).

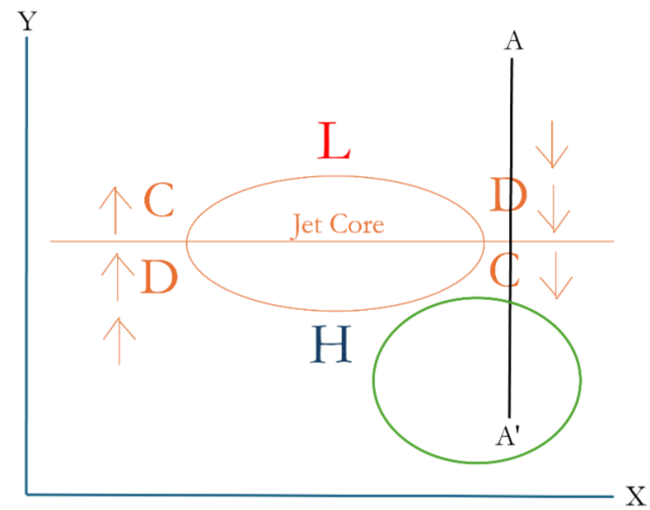


Figure 14 A schematic of jet core dynamics in horizontal plane.

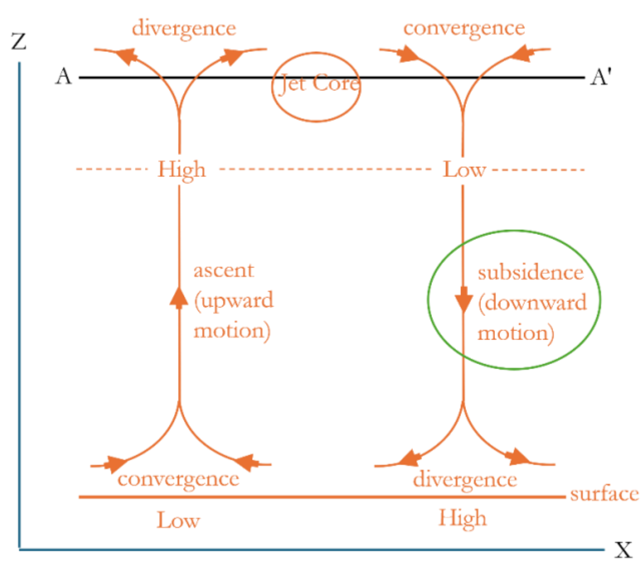


Figure 15 A schematic of jet core dynamics in a vertical cross-section view.

To summarize, quadrants of the jet core correspond with either upper-level convergence or upper-level divergence, which themselves are associated with vertical ascent or subsidence within the atmosphere. Our interpretation of **Figure 14** will rely on knowledge of these concepts.

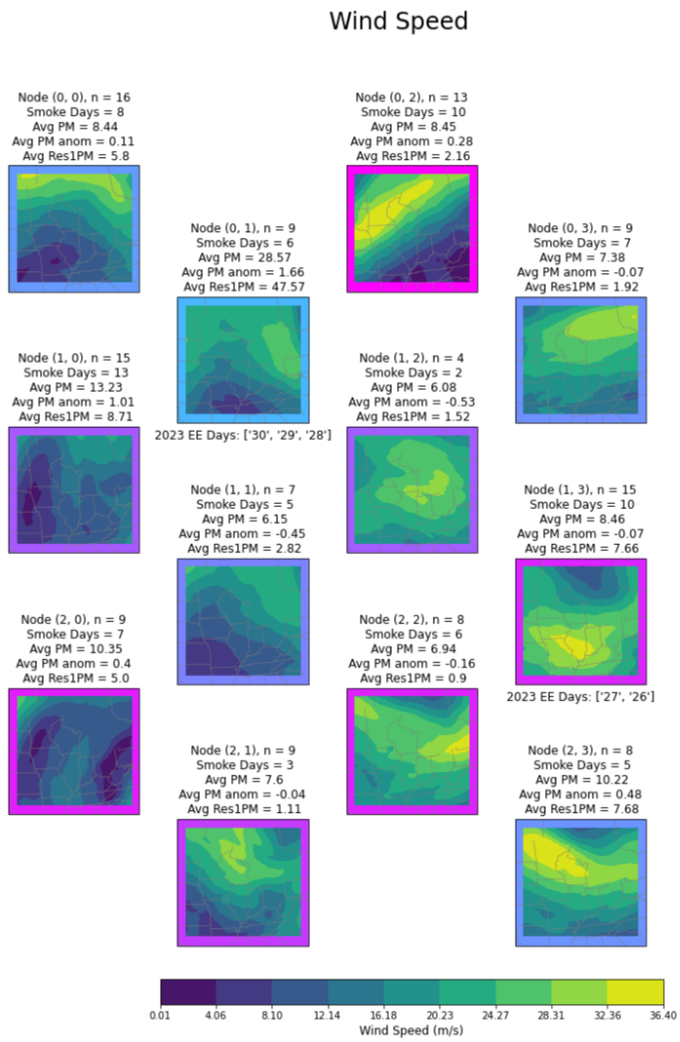


Figure 16 250hPa level wind speed field derived from LADCO SOM in meters per second (m/s).

Analyzing the wind speed field from the LADCO SOM yields several conclusions. Among them, is evidence that upper-level convergence (atmospheric subsidence) and downward motion within the atmosphere is associated with higher $PM_{2.5}$ anomaly; however, the inverse is not explicitly true. Upper-level divergence is generally a feature found within the LADCO region in conjunction with severe storms and deepening mid-latitude low pressure systems. While not necessarily for severe storms, these general quadrants of the jet streak are often used as forecasting tools that can inform where areas of more severe weather are expected. LADCO SOM adds an additional layer of understanding and importance to analyzing the jet streak layer as areas where upper-level convergence is expected can be seen as an indicator for higher $PM_{2.5}$ concentrations at the surface. Given the conclusions interpretable by these fields, as well as the valuable information the jet steak layer presents to forecasters in terms of synoptic level transport, the U and V wind fields presented a natural choice in terms of inclusion into LADCO SOM.

3.2.5 Variable: 850hPa Geopotential Height

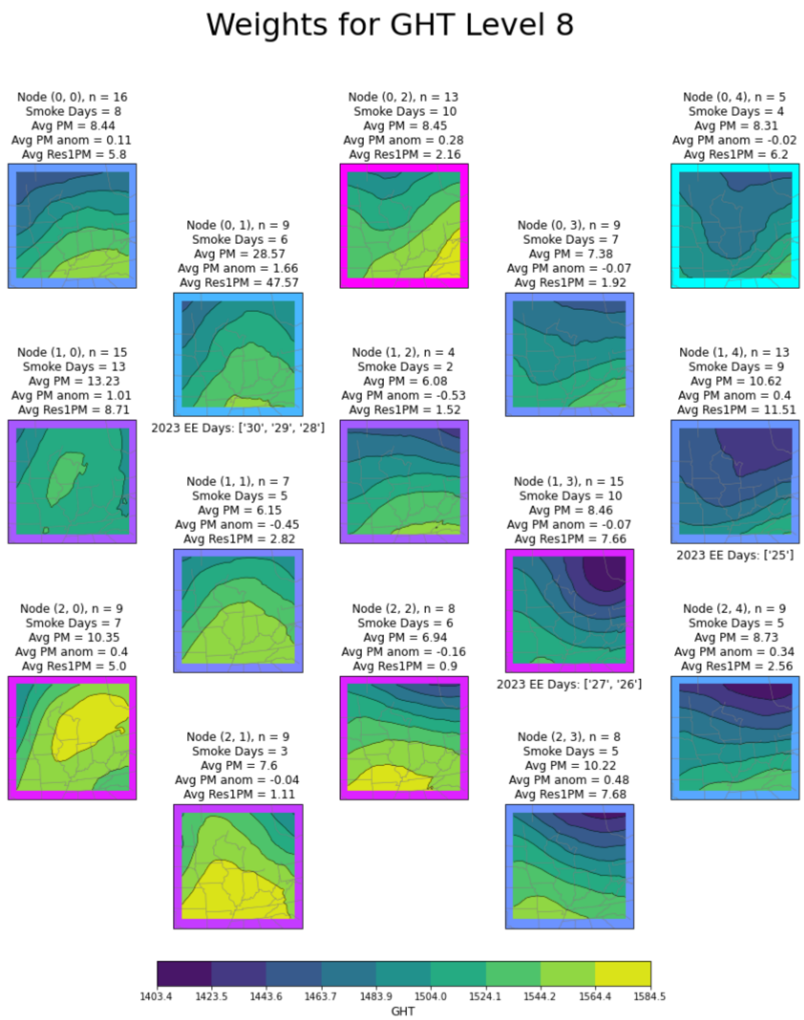
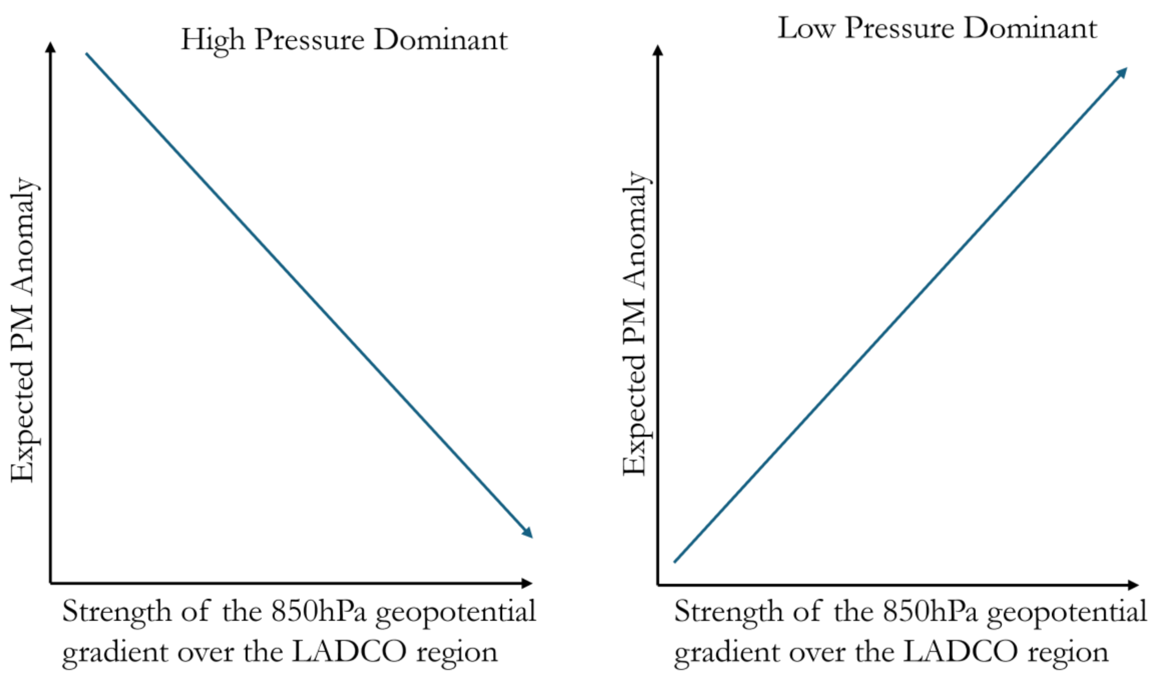


Figure 17 The weights for geopotential height within LADCO SOM in meters (m).

The geopotential height field was chosen as an input into the SOM in an attempt to give LADCO SOM a variable that can act as a classification basis for considering mid-tropospheric flow and the vertical orientation of fronts or pressure systems. The primary geopotential height trend observable from the SOM is in relation to the geopotential gradient. For high pressure dominated nodes, when the distance between isohypses (lines of constant geopotential height) is large, this is associated with less dynamical motions and stagnation conditions, and consequently, higher $PM_{2.5}$ anomaly. When the LADCO region falls under a tighter geopotential gradient the result is a lower $PM_{2.5}$ anomaly as seen in nodes (1,1) and (1,2). This tighter geopotential height gradient indicates stronger advection within the region that clears out pollutants. However, this trend does not seem entirely robust for lower pressure dominated nodes as the LADCO region appears to be within a loose geopotential gradient in nodes (0,3) and (0,4) whilst the $PM_{2.5}$ anomaly is near zero and even slightly negative. However, it has yet to be proven if this is an artifact of our averaging methodology, where these nodes may contain samples where both positive and negative extremes lead to a near zero average and should be examined in future work. In any case, considering the position of the jet streak, and the established westerly flow in nodes (2,3) and (2,4), our gradient trend appears to inverse for low pressure dominated nodes, with tighter gradients leading to higher $PM_{2.5}$ anomaly. Although this trend may prove to not be linear in nature, the following mental model is provided for air quality forecasters.



Figures 18a-b Mental model to aid in operational forecasting of air quality given the 850hPa geopotential height field.

3.3 LADCO SOM VERTICAL PROFILE ANALYSIS

Following the procedure described in section 2.5, vertical profiles for each node were generated. **Figure 19** visualizes the results from this procedure.

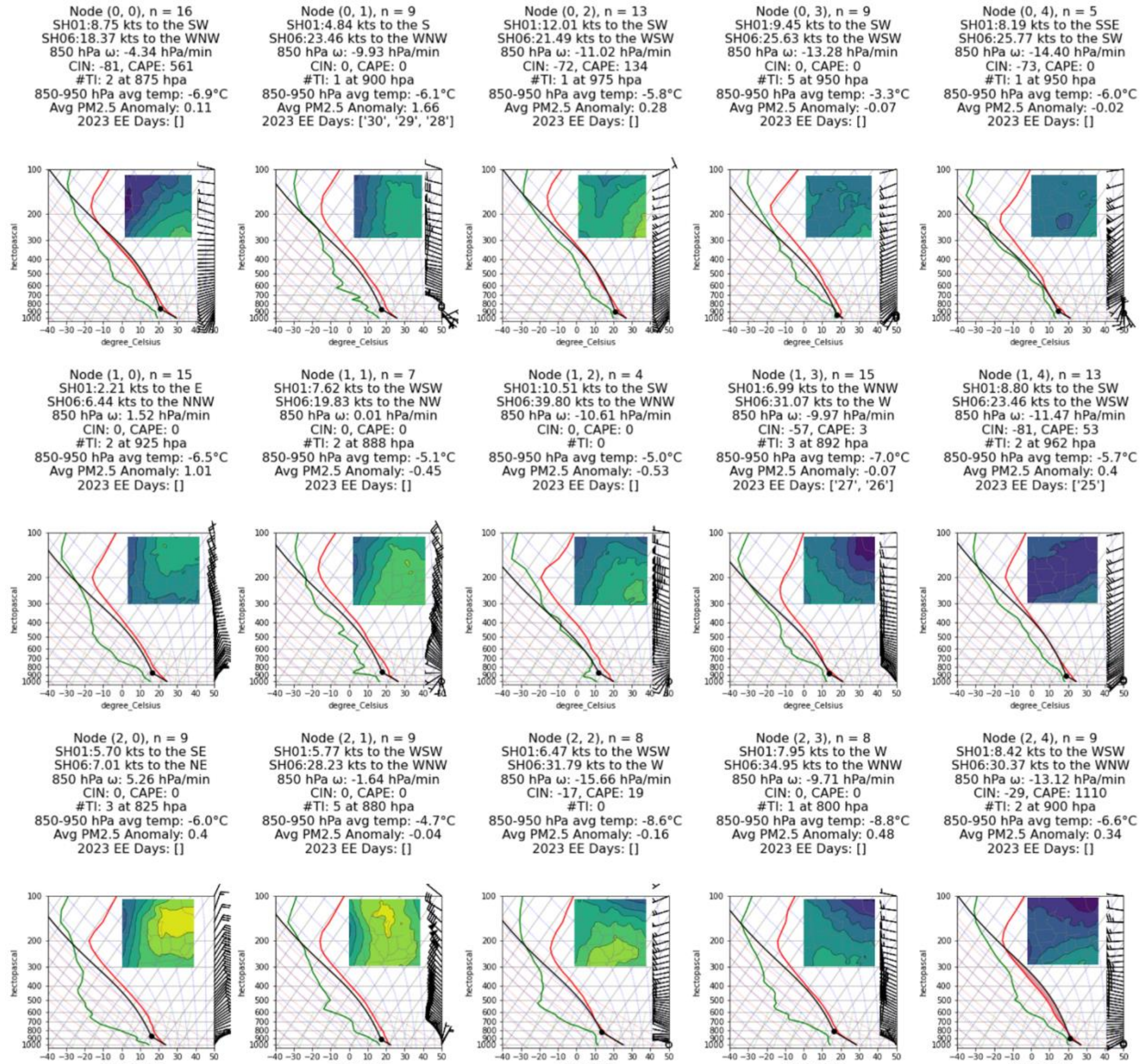


Figure 19 Averaged vertical profiles for each node within LADCO SOM with MSLP pattern in upper right corner.

Figure 19 presents several notable trends, the first of which can be seen by looking at the moisture profiles across all SOM nodes. Although there are exceptions, generally mid-level moisture around the 500hPa level varies greatly, and across the SOM, entire moisture profiles get more saturated starting from the lowest left node (2,0) (dryest) to the uppermost right node (0,4) (most saturated). Notable exceptions to this rule are nodes (1,1) and (1,2) which appear to be surrounded by vertical profiles that are drier. However, these two nodes (that both have very negative $PM_{2.5}$ anomaly) uniquely seem to have a low-level dry layer near the 900hPa level. **Figure 20** displays a zoomed in version of **Figure 19** with these two nodes highlighted. Although nodes (1,1) and (1,2) do not contain an above average number of identified temperature inversions, these dry conditions at the surface could indicate stronger mixing at the surface which acts to disperse pollutants upward, or the drier conditions at the surface could work to reduce the rate chemical reactions that lead to the formation of secondary $PM_{2.5}$ production such as interaction with sulfate and nitrate aerosols (which may be particularly applicable given these sounding originate over Chicago, IL). However, more work is necessary to see if this is indeed the case. Another similarity these nodes seem to share is comparatively less steep 850-950hPa environmental lapse rates, a trend shared by nodes (0,3) and (2,1) which are both accompanied by slightly negative $PM_{2.5}$ anomaly.

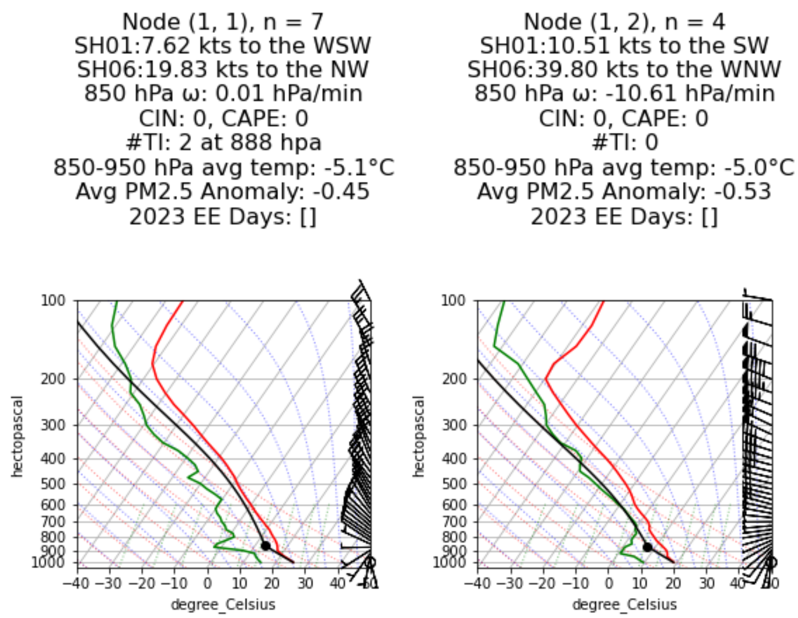


Figure 20 Averaged vertical profiles for nodes (1,1) and (1,2) with low level dry layer near 900hPa.

Another trend that is apparent from **Figure 19** corresponds to the direction and strength of winds at and near the surface level. Nodes (2,0), (1,0), and (0,1) all have slow surface winds that blow eastward, and all have elevated $PM_{2.5}$ anomaly, with nodes (1,4) and (2,4) also having significantly positive $PM_{2.5}$ anomaly however calm winds at the surface as opposed to eastward.

Also shared among high pressure dominated nodes that boast significantly high $PM_{2.5}$ anomaly is low 0-1km shear. Nodes (2,0), (1,0), and (0,1) again all seem to have this in common. With all three nodes having a high pressure dominated MSLP condition, low 0-1km shear points plainly to a correlation of $PM_{2.5}$ impacts with stagnation at the surface. Nodes (2,0) and (1,0) share the additional similarity of positive 850hPa ω values. An indicator for downward vertical motion in the atmosphere.

Figure 21 presents a zoomed in view with a grouping of the high $PM_{2.5}$ anomaly nodes mentioned above for easier reference.

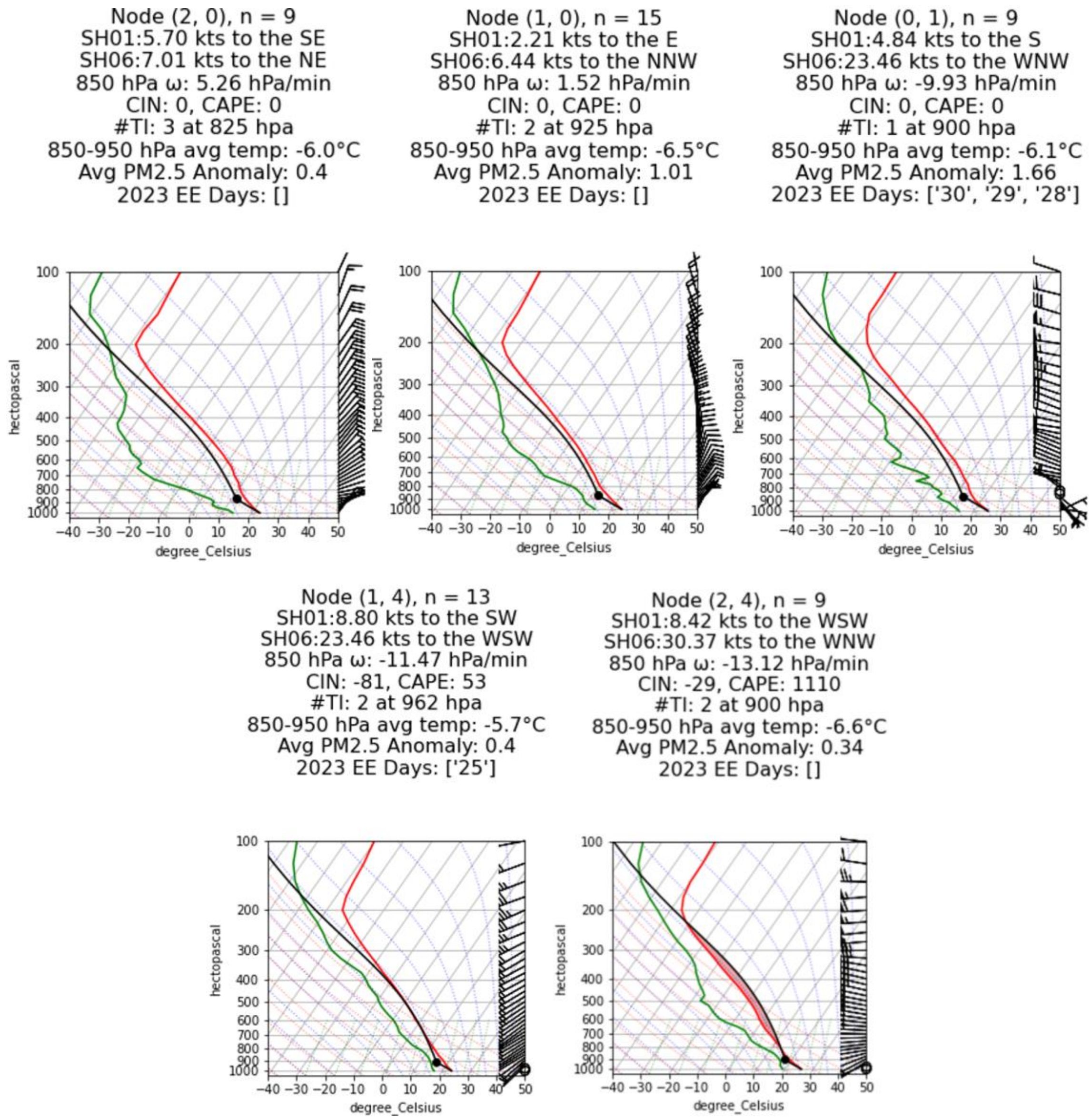


Figure 21 Averaged vertical profiles for nodes (2,0), (1,0), (0,1), (1,4) and (2,4). The top row is characterized by slow and eastward surface winds, bottom row with calm conditions at the surface.

4. APPLICATIONS

Although pipelines and code for operational use have yet to be implemented, LADCO SOM (or an improved version of the SOM in the future) has potential for operational use in the field of air quality forecasting. By inputting the conditions of, for example, a forecast hour +48 HRRR run initialized at 0z, a classification of the modeled atmospheric conditions can be outputted by LADCO SOM. Doing so would allow decision makers an initial forecast of the expected air quality given the atmospheric conditions modeled for the future. Currently LADCO SOM only has knowledge of June $PM_{2.5}$ events across multiple years, an updated operational model would most likely require data spanning multiple months and years, unless it was determined that a more specialized SOM for a specific time period performed better. Perhaps two versions of a SOM are given different data, one corresponding to cold season months and one corresponding to warm season months. Each SOM could be asked to classify the conditions for each day, and the two SOMs running in parallel could cover any days that might be during transitions between seasons, or any anomalously warm or cool days. Moreover, separate SOMs given data corresponding to ENSO patterns (El Niño & La Niña) may also provide additional insight into broader climatological trends specific to the LADCO region.

Air quality forecasters may also find use in inputting the current atmospheric conditions, or the atmospheric conditions from a past event to give an indication for whether or not that particular meteorological setup is synonymous with a certain type of $PM_{2.5}$ anomaly. In this way it is possible to examine the relative anomaly (anomaly detection) of certain events given knowledge of data within the same seasonal period.

Much like the insights gained from this study, given a larger and temporally comprehensive set of data, an updated version of LADCO SOM may be able to discover harder to detect meteorological relationships and classify these similarities into representative analogs, that may have transferable relationships between pressure dominance regardless of season, or seasonal relationships regardless of pressure dominance.

There also exists room for policy evaluation within LADCO SOM (or a future version with an expanded dataset), although perhaps it is not the most direct way of doing so. One way to accomplish a policy evaluation using a SOM would be to provide a SOM with historical meteorological data (and optionally air quality data) before policy implementation to get a baseline understanding of typical patterns and relationships identified by the SOM initially. Then after a policy has been implemented, either look for shifts in patterns that indicate changes in quality under similar meteorological conditions, or (if given air quality data) look to see if similar clusters exist whose primary distinction is on the basis of $PM_{2.5}$. If possible, then examine further similar nodes to analyze what time period samples commonly mapped to each node are from. If two nearby nodes have a similar meteorological setup, but one node has a lower average $PM_{2.5}$ concentration, and the samples within that node come from a time after the policy was implemented, this could be a sign that a certain policy was effective. Although perhaps a more actionable response might be to then perform a more rigorous comparative analysis of each node, now informed by the SOM of the data's meteorological similarity.

5. FUTURE IMPROVEMENTS

This section will outline aspects of LADCO SOM that should be targeted for improvement in the future. Many are simple enhancements, among them different (or just more) input data distributions as mentioned in **section 4**. However, of the improvements mentioned below, several would significantly alter (and improve) the current functionality of LADCO SOM. While worthwhile, the time required to perform such modifications is left to future work. This section will be broken down into subsections relating to each potential improvement.

5.1 LADCO SOM + KRIGGED $PM_{2.5}$

In the later stages of the project a question occurred that was: would the results of the SOM clustering significantly improve given knowledge of $PM_{2.5}$ concentrations on the ground? To answer this question, we explored a few different options, although just imputing an average PM2.5 variable tacked on as a column on the end of the dataset was not successful, owed to the dataset's high dimensionality. For this a PM2.5 variable with the same dimensions as a single meteorological variable needed to be considered. This was accomplished via the creation of a "krigged" (spatially interpolated) $PM_{2.5}$ dataset. A visualization of what this dataset looks like when plotted over the study region is displayed in **Figure 5**.

The krigged $PM_{2.5}$ dataset was able to be successfully integrated into LADCO SOM's code, however the results were not necessarily informative to the research question. **Figure 22** presents the results of LADCO SOM being run with all 6 meteorological variables + the krigged PM2.5 dataset as input variables.

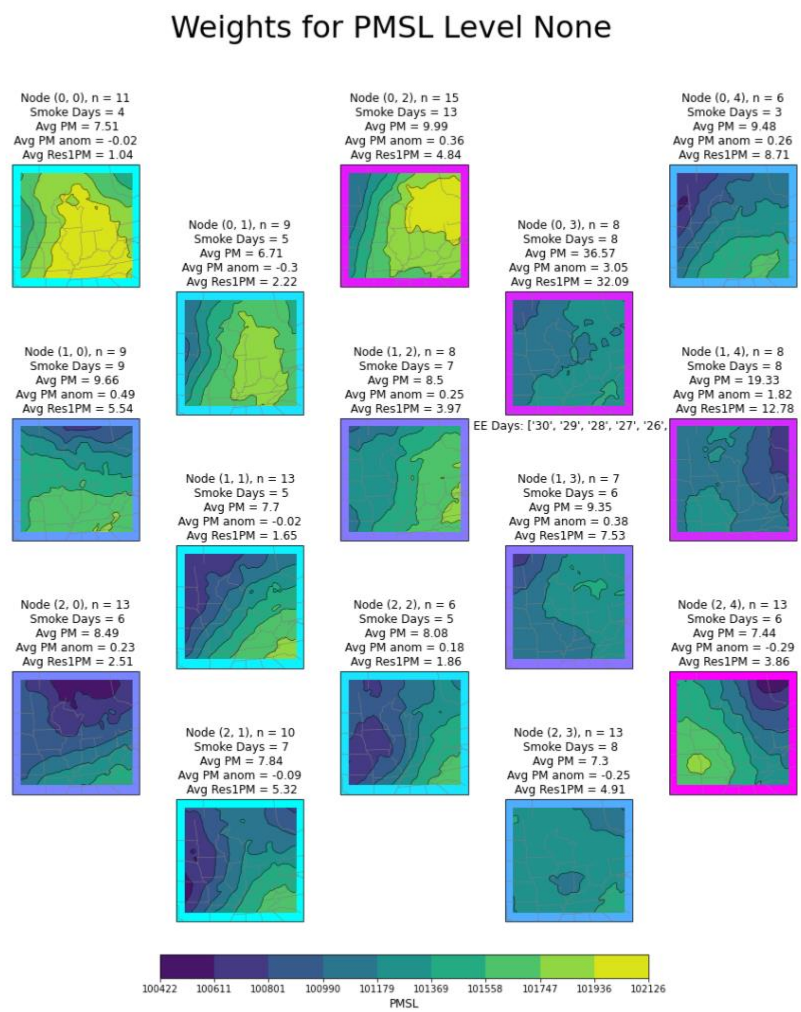


Figure 22 The weights for MSLP within LADCO SOM with the krigged PM2.5 variable.

The resulting SOM appears to classify every day from the 2023 EE into one node. Except, this is questionable because we are certain that the meteorological conditions changed significantly throughout the course of the event period. This version of LADCO SOM model built with krigged $PM_{2.5}$ indicates that i) The 2023 EE days are being classified together into a node could be due to the extremely anomalous $PM_{2.5}$ impacts, and ii) it calls for examination of surface input variables when using the krigged $PM_{2.5}$ field. Instead of gleaning insight into the meteorological conditions associated with these impacts, we get yet another indicator that is event is extreme, which is already known. **Figure 23a** displays this undesirable behavior, pay particular attention to the scale in (µg/m^3), **Figure 23b** is generated with a high value mask to see $PM_{2.5}$ values within the first interval of **Figure 23a**.

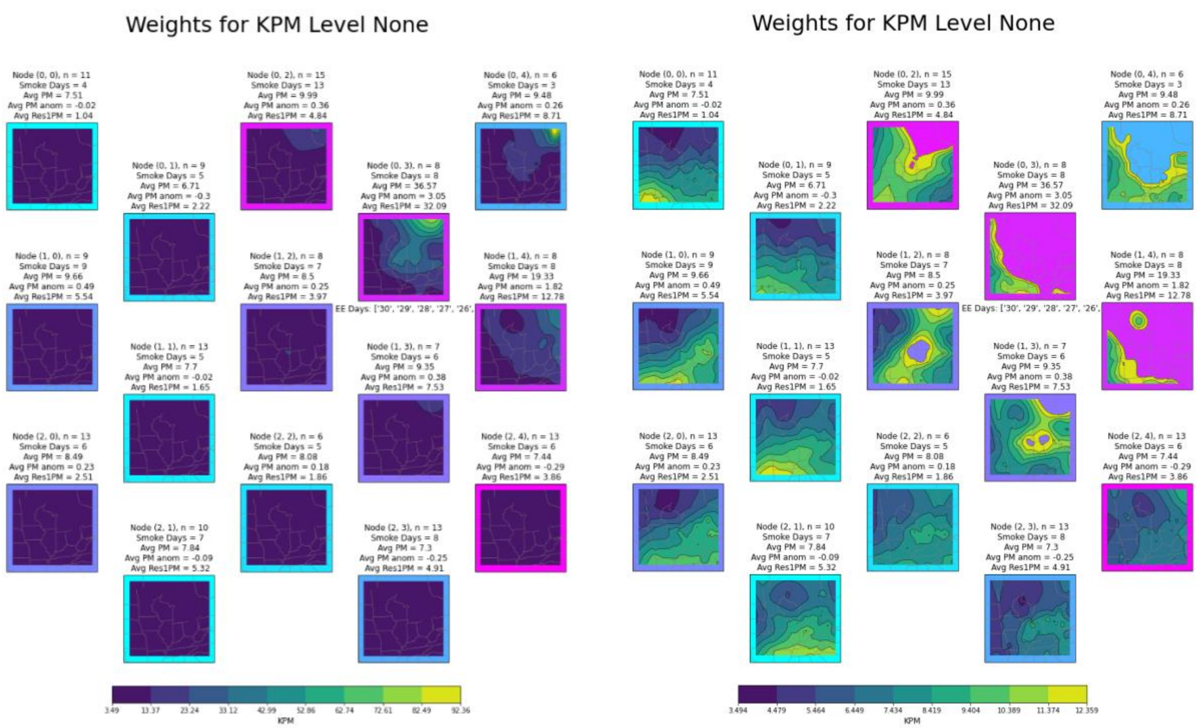


Figure 23a-b The weights for the krigged $PM_{2.5}$ variable within LADCO SOM + krigged $PM_{2.5}$.

We can see clearly that there certainly exist interesting $PM_{2.5}$ relationships with meteorology data, although we are also certain that this data is being significantly impacted by the 2023 EE. Hence, improvements to the meteorological variables + krigged $PM_{2.5}$ SOM is left for future work, a possible direction may be as simple as excluding the 2023 EE.

5.2 OZONE SOM ANALYSIS

Another SOM application that initially started off as a part of the project, but then fell off due to time constraints is an analysis of ground level ozone. Given ozone's more predictable relationship with meteorological conditions, a version of LADCO SOM that does a thorough analysis of ozone would be extremely beneficial. In its current state LADCO SOM is capable of calculating statistics of SOM-grouped ozone data, however interpretation of these results remains a challenge. **Figure 24** displays LADCO SOM with ozone statistics visualized.

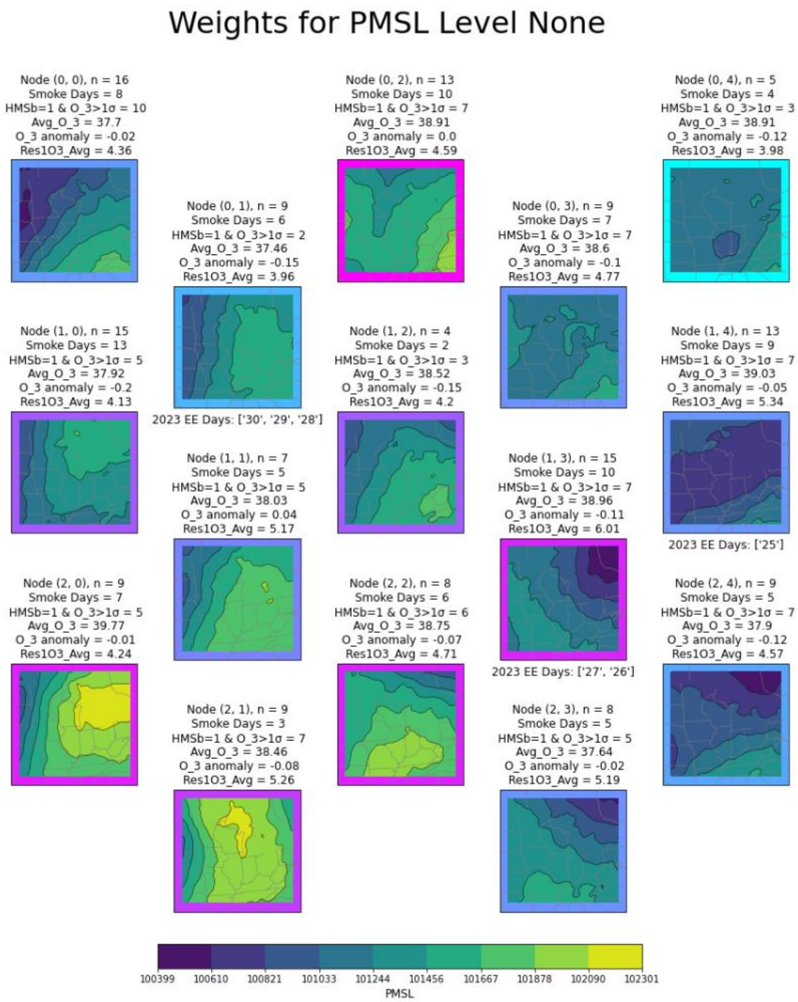


Figure 24 The weights for MSLP within LADCO SOM with calculated ozone statistics.

As the data currently stands the highest ozone anomaly is 0.04 associated with node (1,1) which isn't particularly anomalous, although interestingly is associated with a node whose vertical profile has a dry layer near the surface. The shortfalls of LADCO SOM in ozone classification perhaps lay within low ozone variability in the month of June for the LADCO region? Therefore, an increasingly seasonal dataset may prove useful for further ozone analysis. Particularly for the ozone case, incorporation of a krigged ozone dataset based off of ground monitors may provide the SOM with extra knowledge of ozone concentrations to cluster nodes off of.

5.3 LADCO SOM WITH ADDITIONAL METEOROLOGICAL VARIABLES

A potential way to address the current issues with the SOM presented in 5.1, is to add more meteorological variables to offset the proportion of the input space the krigged $PM_{2.5}$ takes up.

While this significantly adds to the dimensionality of the SOM (and substantially to the overall runtime), considering that dimensionality already presents an issue within the study, a test was ran considering an expanded array of meteorological variables. While the same six meteorological variables were used within as introduced in **section 2.2** this expanded version of the SOM, LADCO SOM was given the data for these variables at the "critical levels" within the atmosphere. The full array of considered variables is as follows in the code:

```
variables = [
    ('PMSL', None),
    ('RH',1), ('RH',4), ('RH',8), ('RH',14), ('RH',22), ('RH',32),
    ('TT',1), ('TT',4), ('TT',8), ('TT',14), ('TT',22), ('TT', 32),
    ('UU',1), ('UU',4), ('UU',8), ('UU',14), ('UU', 22), ('UU',32),
    ('VV',1), ('VV',4), ('VV',8), ('VV',14), ('VV',22), ('VV', 32),
    ('GHT',1),('GHT',4),('GHT',8),('GHT',14),('GHT',22),('GHT',32),
    ('KPM', None)
]
```

Where variables only available at the surface have a level "None" and the associated pressure levels with model levels are as such:

1 = Surface level (near 1000hPa)

4 = 950hPa

8 = 850hPa

14 = 700hPa

22 = 500hPa

32 = 250hPa

Figure 25 demonstrates how the 2023 EE days are no longer classified into a single node, and meteorological variables are once again primarily used for distinctions between nodes. The most apparent problem the expanded SOM presents is exponentially higher dimensionality where each input vector has of length ~6Million "columns". For this reason, it is hard to gauge whether classifications made the expanded LADCO SOM are accurate, visualizations of other variables within the expanded LADCO SOM present somewhat contradictory information to those discussed within this report, although the legitimacy and verification of these results is questionable and hence left to future work. An initial step for improving the LADCO SOM model is to conduct an exploratory analysis on a representative set of input variables guided be Principal Component Analysis or other dimensionality reduction techniques prior to building SOM models.

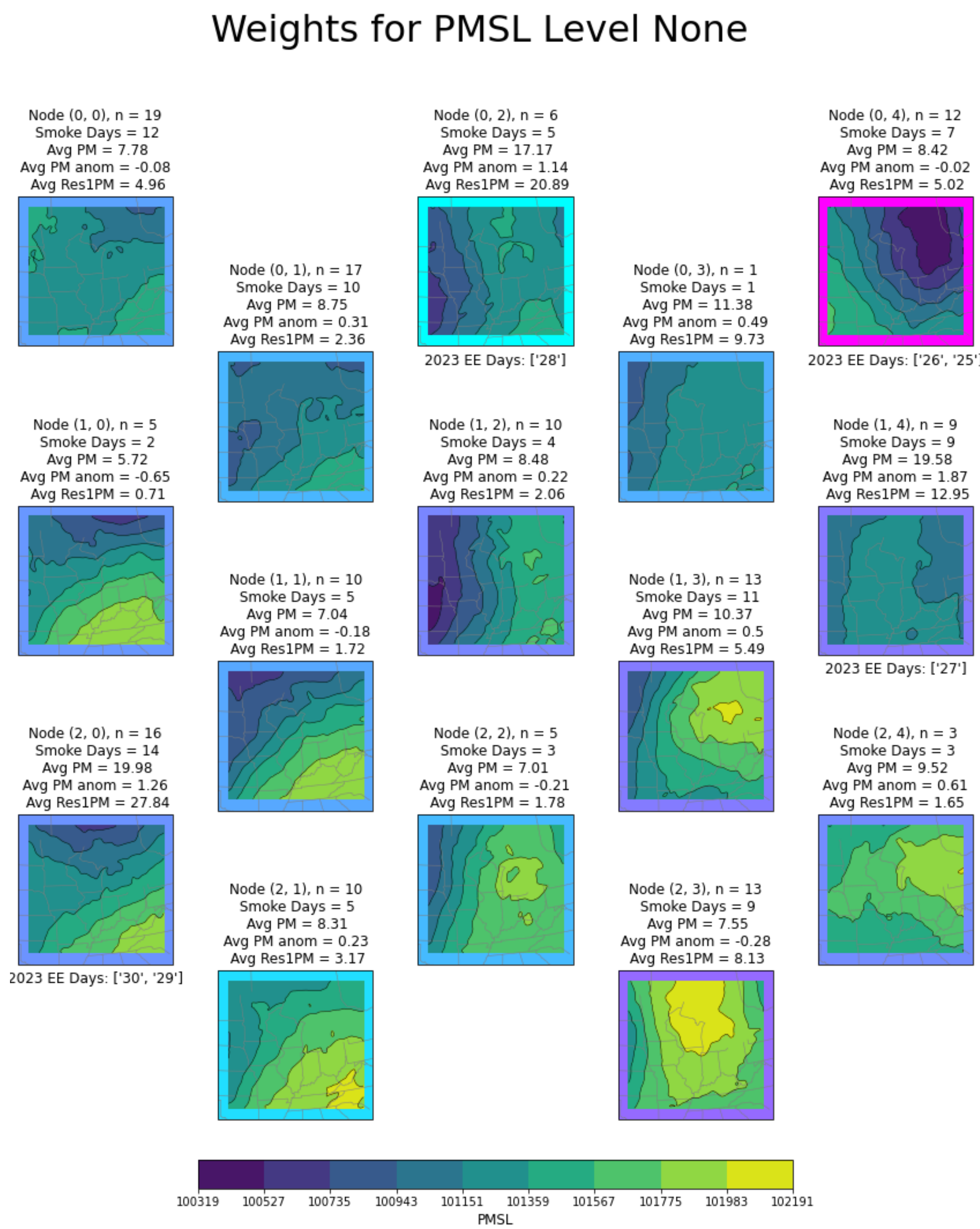


Figure 25 The weights for MSLP within the expanded LADCO SOM.

5.4 DIMENSIONALITY AND QUANTITATIVE ANLYSIS IMPROVMENTS

Owing to the extremely high dimensionality of our input data, quantitative clustering metrics have a harder time diagnosing proper hyperparameters for LADCO SOM. The SOM presented within this paper has the following clustering metrics:

Calinski-Harabasz Score: 14.7473, Silhouette Score: -0.0242, Davies-Bouldin Index: 2.4361

While these metrics are generally used to evaluate different hyperparameter configurations, it is plainly noticeable that our Silhouette score is negative (when higher values for Silhouette score are supposed to represent better clustering); and while this might be cause for concern in other cases, it was not mentioned earlier in the report as there is and explainable reason for this. **Figure 8** displayed the best QE achieved by LADCO SOM is around 824, and this is due to our highly dimensional data. Our Silhouette score is negative for a similar reason. Silhouette score measures how similar a point is to its own cluster compared to other clusters, which is this case leads to all points being roughly equidistant from each other, which is resulting in a negative Silhouette score. Similar interpretations can be reached when considering Calinski-Harabasz Score and the Davies-Bouldin Index.

Calinski-Harabasz Score: This score evaluates the ratio of the sum of between-cluster dispersion and within-cluster dispersion. High dimensionality is leading to increased within-cluster dispersion due to the curse of dimensionality, where distances between points become less meaningful as dimensionality increases.

Davis-Bouldin Index: This index evaluates the average similarity ratio of each cluster with its most similar cluster, considering cluster centroids. Considering the dimensionality of input data, our centroid might not be a good representation of the cluster, leading to higher index values.

In all three cases, the results align with challenges posed by high dimensionality. How then is LADCO SOM being evaluated? To this we point to trial and error during hyperparameter adjustment, manual inspection, and using close to default settings, with the initial requirement that every step (including the results) is understood, explainable, or interpretable. We also know from (Hewitson and Crane 2002) that the results of the SOM are less dependent on the data conforming to a specific distribution or underlying model.

Since no methods exist for perfectly reconstructing high dimensional data spaces within non-linear manifold representation learning, and that LADCO SOM relies on data of the same shape to reconstruct any useful visuals, we are currently at a stalemate with this dimensionality.

However, potential improvements on the front of LADCO SOM's dimensionality could come in the form of applying additional dimensionality reduction techniques before consideration by the SOM such as t-Distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection. Although in effect we are not trying to explicitly reduce the dimensionality of our data (as in get rid of less informative columns) and instead are trying to reduce the amount of data it takes to represent them. In this way LADCO SOM can still be thought of as on par with these dimensionality reduction techniques in regards to the presentation of analogs for meteorological conditions, that describe (in much fewer cases) a representative map for classifying the meteorological conditions of future $PM_{2.5}$ events.

6. CONCLUSIONS

The report has demonstrated how Self-organizing maps (SOMs) can provide additional insight into classifying meteorological conditions and their associated $PM_{2.5}$ impacts and provided justification for modes of vertical transport capable of carrying fire smoke to the surface. Moreover, based on a known extreme event such as the late June 2023 event we confirm that the SOM's behavior is both predictable and explainable. Finally, we present applications for this research in the field of air quality forecasting and analysis and demonstrate the need for future research on the topic.

ACKNOWLEDGEMENTS

The author acknowledges the work of Tsengel Nergui for her preparation of all input data for LADCO SOM including the krigged $PM_{2.5}$ data, as well as her support throughout the duration of the project and beyond. The author also acknowledges the support of all LADCO staff for their positive encouragement and productive criticism and conversations.

REFRENCES

Carrasco Kind, Matias, and Robert J. Brunner. 2014. "SOMz: Photometric Redshift PDFs with Self-Organizing Maps and Random Atlas." *Monthly Notices of the Royal Astronomical Society* 438 (4): 3409–21. https://doi.org/10.1093/mnras/stt2456.

Deboeck, Guido, and Teuvo Kohonen. 2013. *Visual Explorations in Finance: With Self-Organizing Maps*. Springer Science & Buisness Media.

Forest, Florent, Mustapha Lebbah, Hanene Azzag, and Jérôme Lacaille. 2021. "Deep Embedded Self-Organizing Maps for Joint Representation Learning and Topology-Preserving Clustering." *Neural Computing and Applications* 33 (24): 17439–69. https://doi.org/10.1007/s00521-021-06331-w.

Hewitson, B. C., and R. G. Crane. 2002. "Self-Organizing Maps: Applications to Synoptic Climatology." *Climate Research* 22 (1): 13–26.

Hrust, Lovro, Zvjezdana Bencetić Klaić, Josip Križan, Oleg Antonić, and Predrag Hercog. 2009. "Neural Network Forecasting of Air Pollutants Hourly Concentrations Using Optimised Temporal Averages of Meteorological Variables and Pollutant Concentrations." *Atmospheric Environment* 43 (35): 5588–96. https://doi.org/10.1016/j.atmosenv.2009.07.048.

Hulle, Van, and M Marc. 2012. "Self-Organizing Maps." *Handbook of Natural Computing* 1:585–622.

Keyser, Daniel, and M. A. Shapiro. 1986. "A Review of the Structure and Dynamics of Upper-Level Frontal Zones." *Monthly Weather Review* 114 (2): 452–99. https://doi.org/10.1175/1520-0493(1986)114<0452:AROTSA>2.0.CO;2.

Kohonen, Teuvo. 1982. "Self-Organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics* 43 (1): 59–69. https://doi.org/10.1007/BF00337288.

Törönen, Petri, Mikko Kolehmainen, Garry Wong, and Eero Castrén. 1999. "Analysis of Gene Expression Data Using Self-organizing Maps." *FEBS Letters* 451 (2): 142–46. https://doi.org/10.1016/S0014-5793(99)00524-4.

Vesanto, J., and E. Alhoniemi. 2000. "Clustering of the Self-Organizing Map." *IEEE Transactions on Neural Networks* 11 (3): 586–600. https://doi.org/10.1109/72.846731.