

DECODING MIDWEST JUNE PM_{2.5} EVENTS: A SELF-ORGANIZING MAP APPROACH TO METEOROLOGICAL ANALYSIS

Victor Geiser

LADCO Meteorology Intern Summer 2024

July 24th, 2024



PROJECT BACKGROUND

- Periodically, fire smoke is transported into the Midwest and results in unhealthy concentrations of fine particulate matter (PM_{2.5})
- The US Midwest's central placement within the North American continent makes it a common place for a diverse range of meteorological and chemical processes
- The identification of common meteorological setups associated with PM_{2.5} concentrations is applicable to both the fields of air quality and meteorology

PROJECT GOALS

- Find common weather patterns that occur alongside high PM_{2.5} days in the month of June for the LADCO region and analyze their potential for long range pollution transport
- Interpret the results of our SOM (Self Organizing Map) and apply this understanding to an exceptional PM_{2.5} event that occurred on June 25-30th 2023
- Compare the synoptic weather conditions in the Midwest during air pollution episodes with and without the influence of wildfire smoke



RESEARCH QUESTION

How can Self Organizing Maps (SOMs) be used to identify meso-scale meteorological conditions associated with high PM_{2.5} and fire smoke impacted conditions in the LADCO region?

JUNE 25-30th CANADIAN WILDFIRE EXCEPTIONAL EVENT

Source:

LADCO Exceptional Event TSD

- AirNowTech
- NOAA's Hazard Mapping System (HMS)
- NOAA/NASA GOES-16

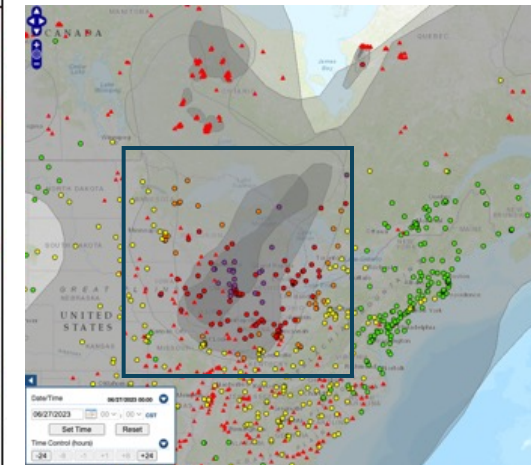
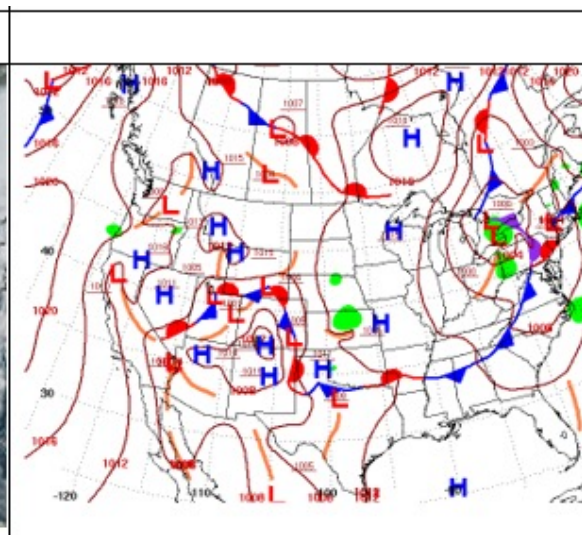
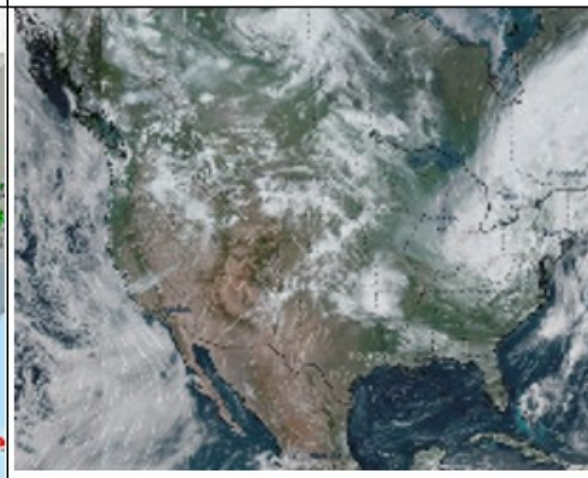
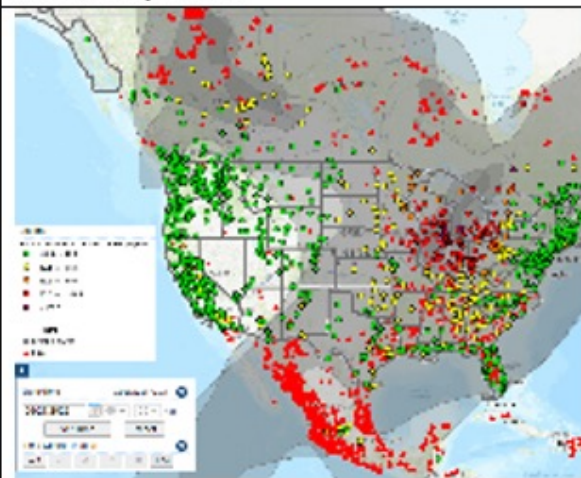


Chicago Skyline June 27 Jamie Keleter Davis/Bloomberg via Getty Images

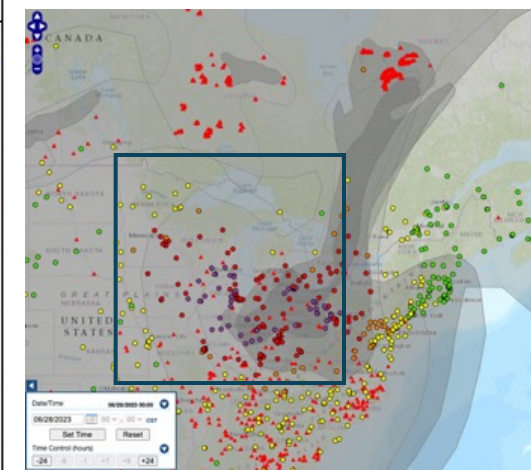
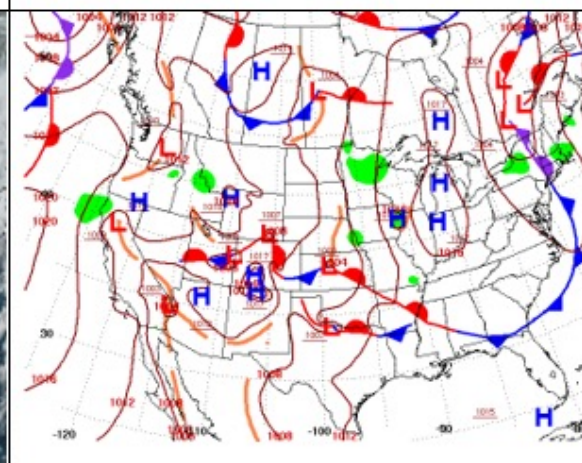
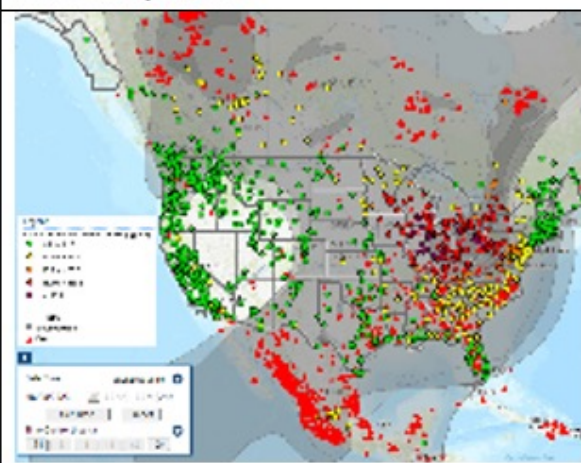


2023 JUNE 25-30 EXCEPTIONAL PM EVENT

June 27, 2023



June 28, 2023

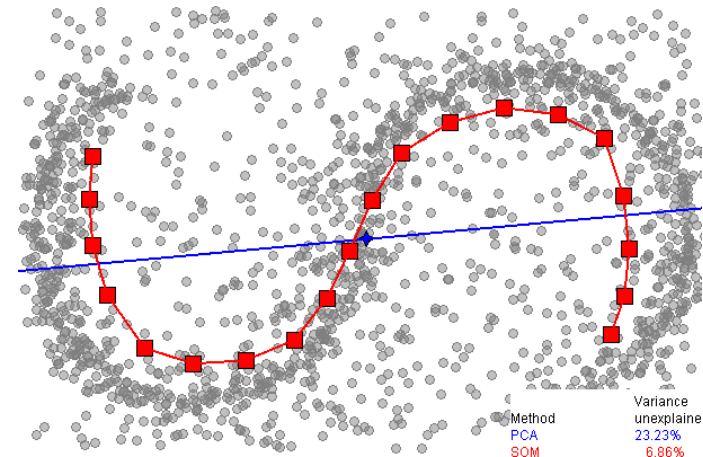
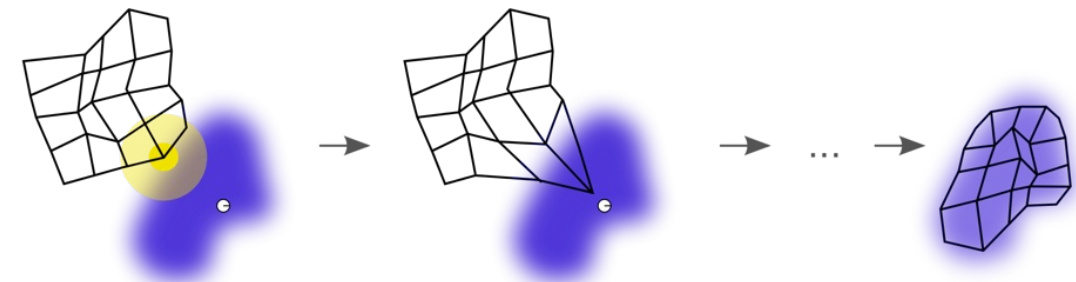




SELF ORGANIZING MAPS (SOMs) ALGORITHM

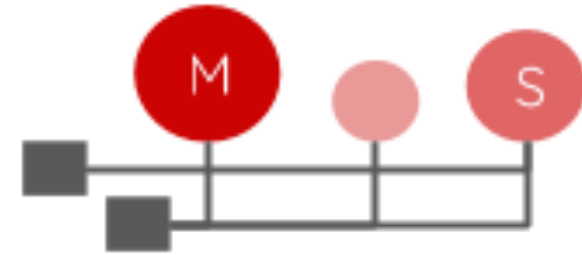
SELF ORGANIZING MAPS ALGORITHM

- **Self Organizing Maps (SOMs)** were originally proposed in 1982 by **Teuvo Kohonen**.
 - Artificial neural network.
 - Finds lower dimensional relationships in high dimensionality data.
 - Attempts to preserve the original structure (topology) of the data.
 - Doesn't make any underlying assumptions about the data (such as PCA that assumes a linear relationship).



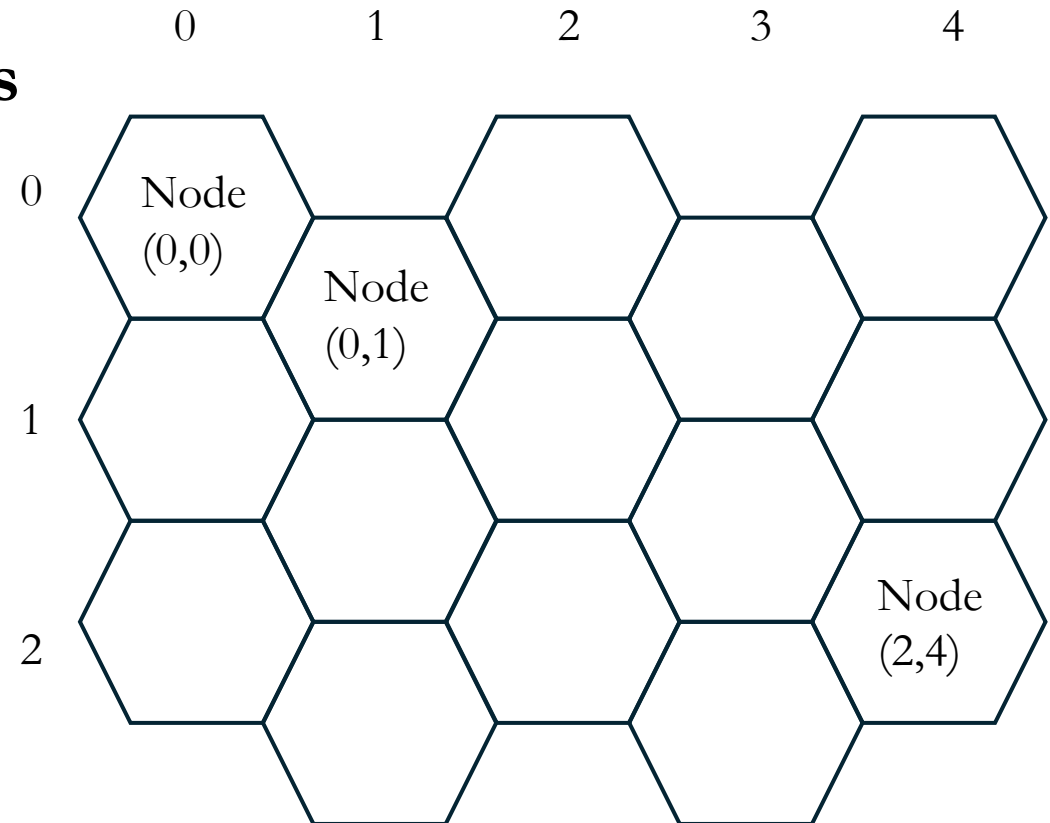
SELF ORGANIZING MAPS IMPLEMENTATION

- “MiniSOM”
- Written in python by “JustGlowing” + contributors
- Accessible on Github
 - <https://github.com/JustGlowing/minisom>
- Minimalistic implementation of Self Organizing Maps that relies only on the Numpy library
- Cited more than 300 times



WHAT KIND OF SOM ARE WE USING?

- **Abbreviated SOM Hyperparameters:**
 - **SOM size of 3 rows and 5 columns**
 - **Gaussian neighborhood function**
 - **Hexagonal topology**
 - **Euclidean distance function**
 - **200 training iterations**
 - **Randomly initialized**

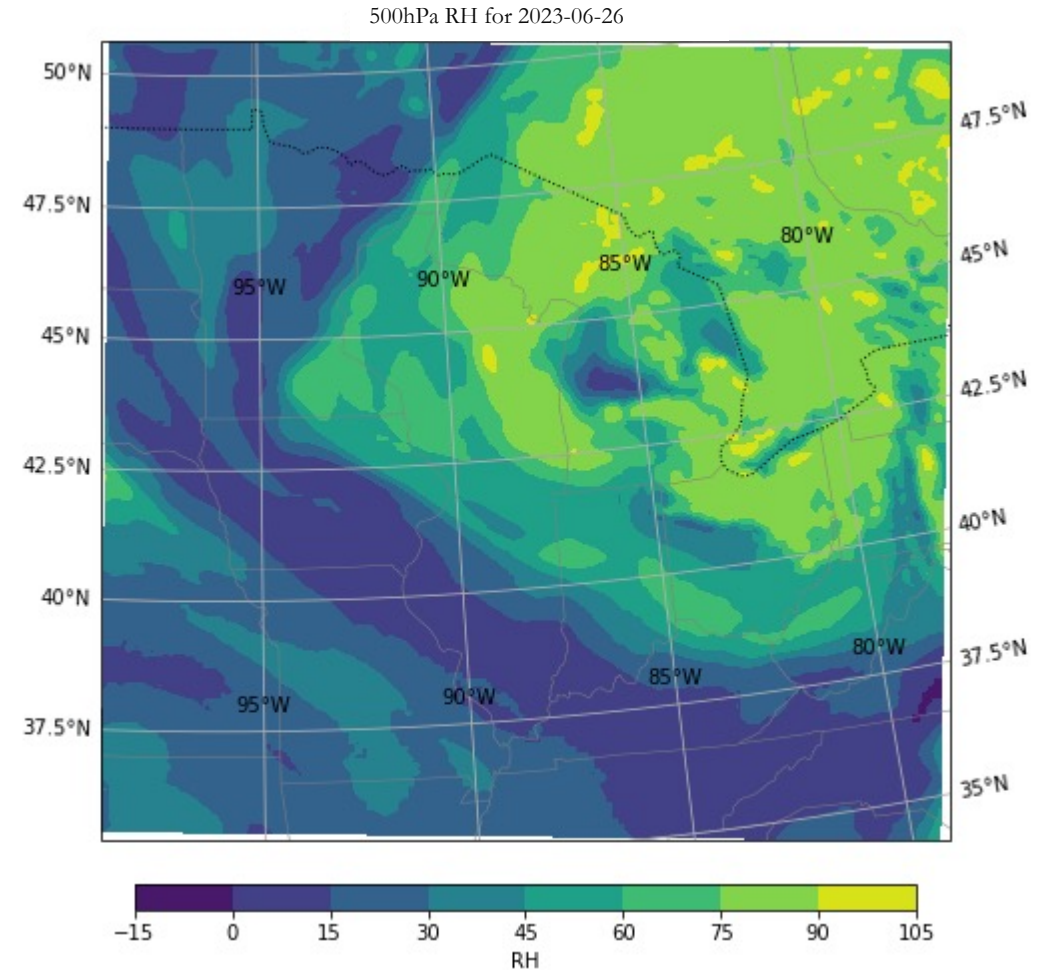




SOM INPUT DATA

LADCO DATA BACKGROUND

- Daily 4km HRRR-NAM reanalysis data
- Midwestern extent
- 149 samples comprising of all June days between 2019 and 2023*



* = June 1st, 2023, is missing due to an incomplete HRRR run



MULTIVARIATE SPATIAL MINISOM

- We want to include all of our variables. However, we can't just input this data into MiniSOM as is. {ValueError}
- The tabular oriented data structure of MiniSOM is not conducive to our inherently spatial data.
- What is the best way to make our data look “tabular”?

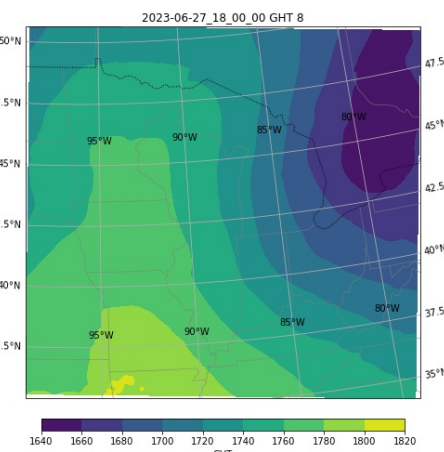
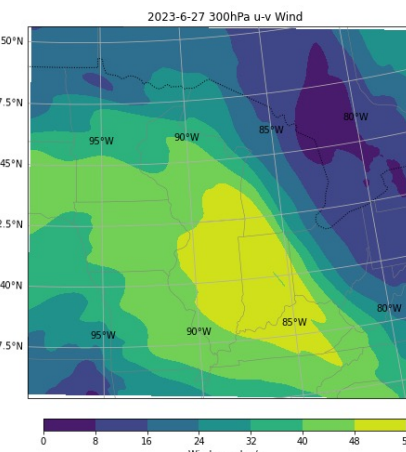
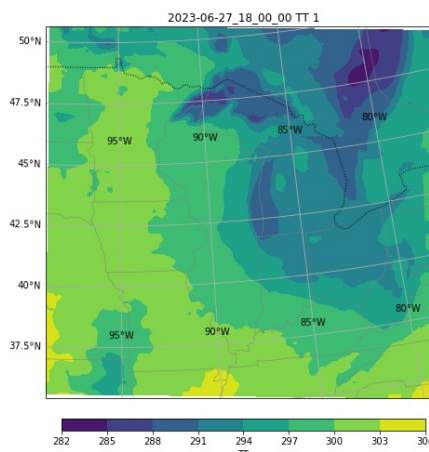
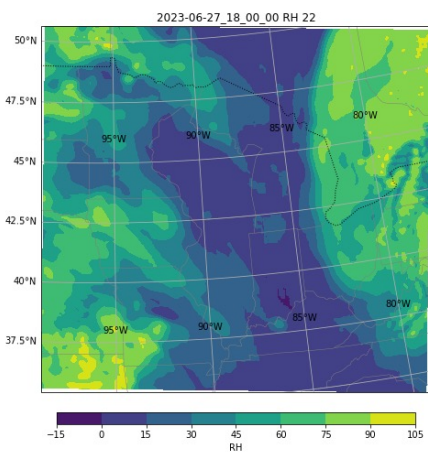
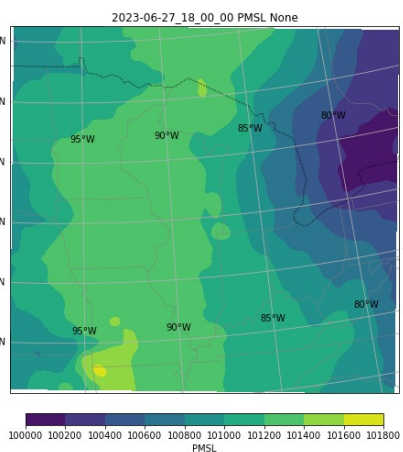
A possible solution? -> Vectorization

Before	→	After
[[a, b], [c, d]]		[a, b, c, d]



METEOROLOGICAL VARIABLES USED

- Mean sea level pressure
- 500hPa Relative Humidity
- Surface Temperature
- 300hPa U-wind and 300hPa V-wind
- 850hPa Geopotential Heights



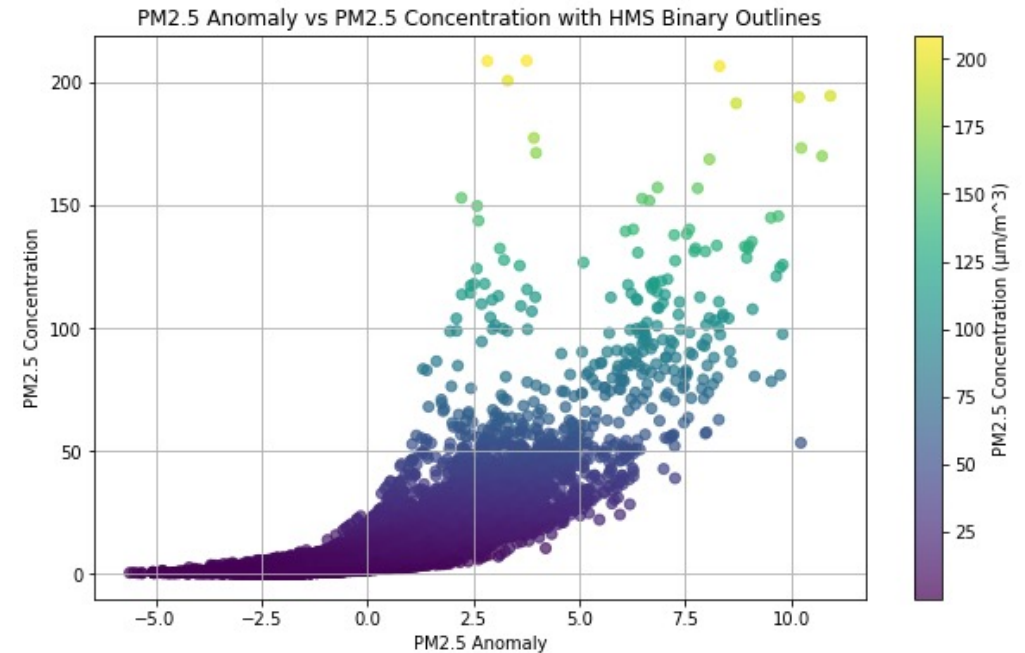


NODE PM_{2.5} METRICS PART 1 OF 4

- “n”
 - Number of samples classified into that node
- “Smoke Days”
 - Number of samples within a particular node with both identified smoke aloft and significantly elevated PM_{2.5} at the surface ($\sigma > 1$)

NODE PM_{2.5} METRICS PART 2 OF 4

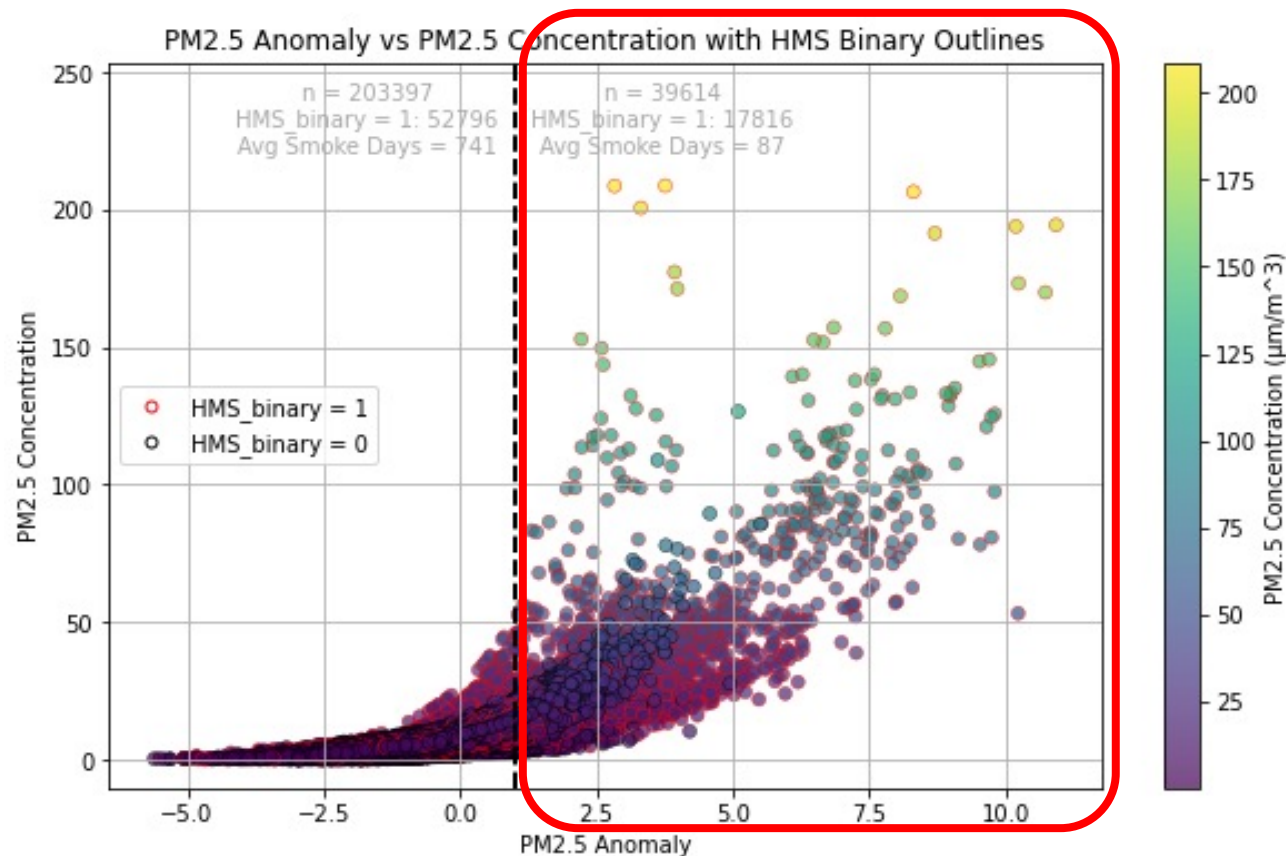
- “Avg PM”
 - Node averaged PM_{2.5} concentration for all ground monitors across all days classified into that node
- “Avg PM anom”
 - A standardized value of log transformed PM_{2.5} concentration.
 - Measures the relative PM_{2.5} anomaly of a PM_{2.5} measurement, and when averaged and applied to the SOM, a particular node.





NODE PM_{2.5} METRICS PART 3 OF 4

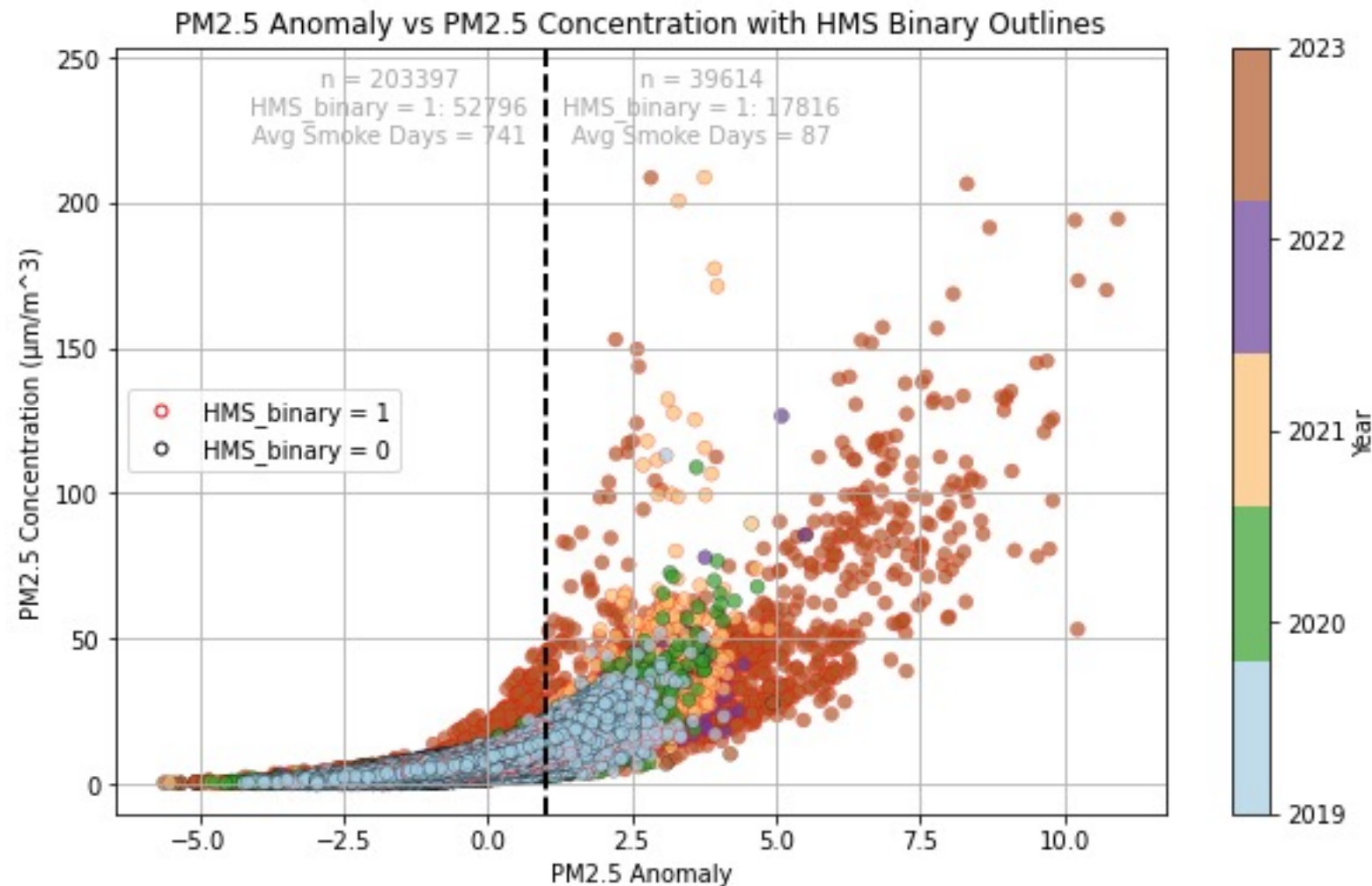
- “Avg Res1PM”
 - An average residual value of PM_{2.5} that is applied on only the samples that have a Standardized Value > 1 AND where HMS Binary = 1
 - It is taken considering an average of whole node (all days and samples)
- Tells us how much PM_{2.5} concentration increased beyond the high end of what is normal





NODE PM_{2.5} METRICS PART 4 OF 4

- All PM Metrics by year:



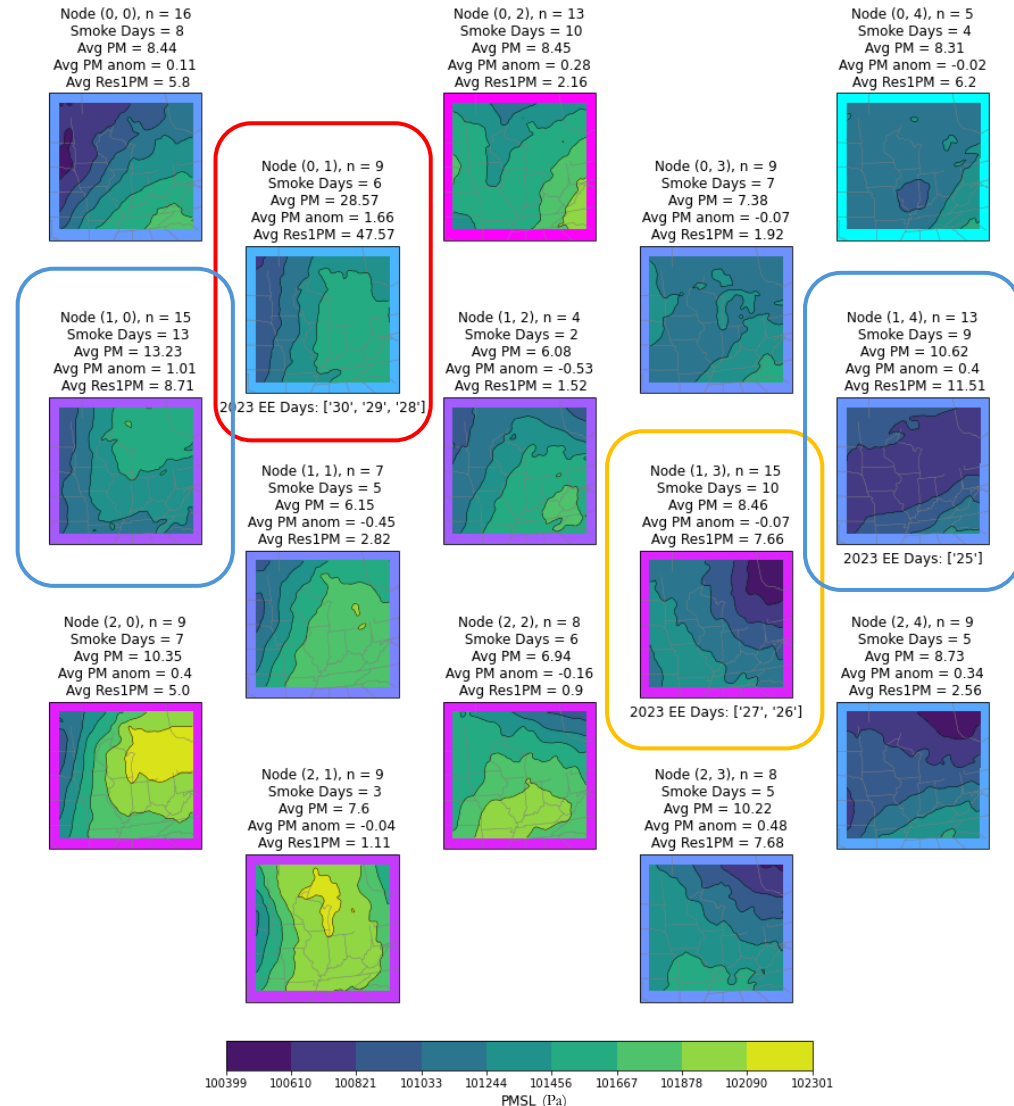


SOM: METEOROLOGICAL ANALYSIS AND CONCLUSIONS

MEAN SEA LEVEL PRESSURE

Notable features:

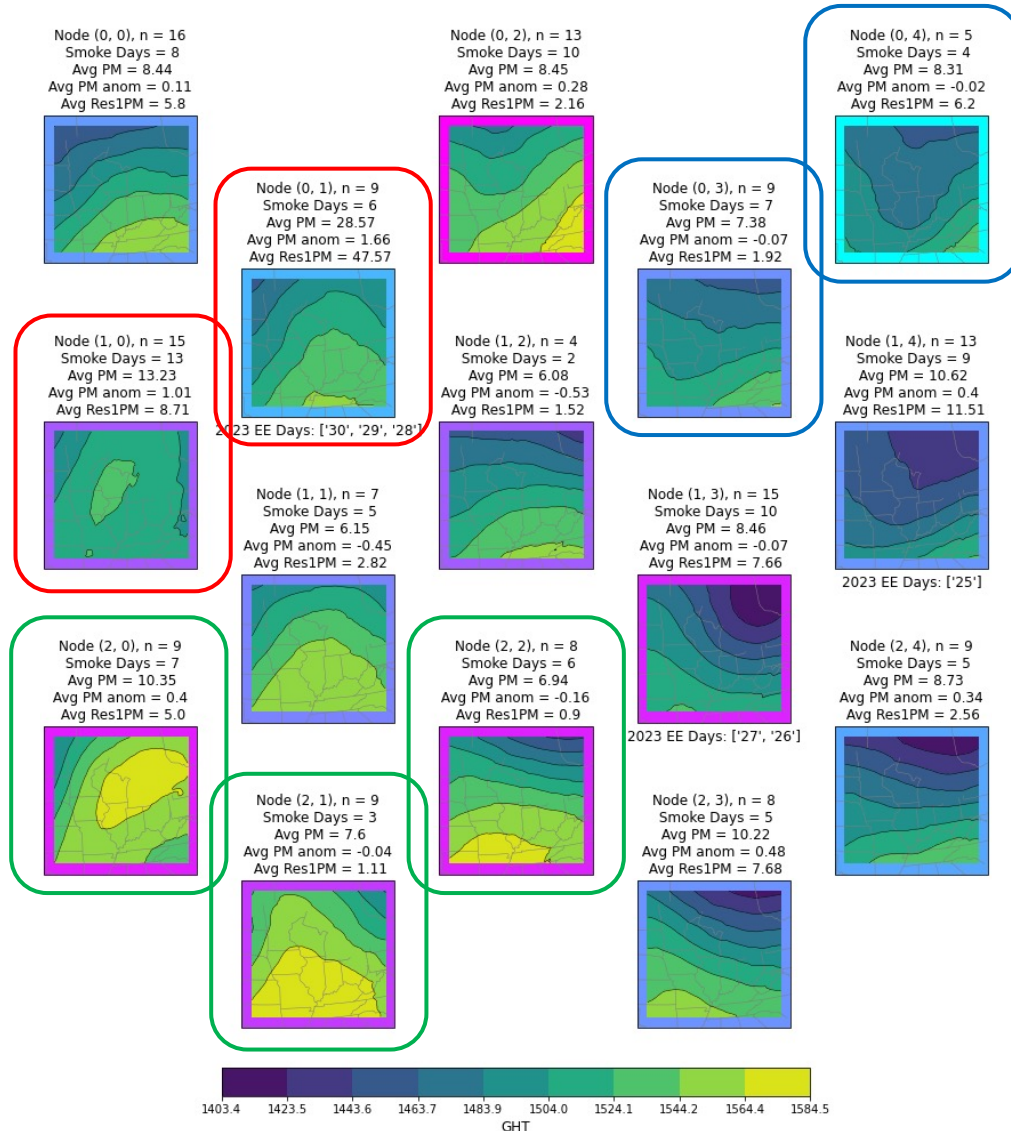
1. **Node (0,1) Stagnation conditions with a blocking high pressure system over SE of LADCO region.**
 - a. Associated with the highest PM2.5 Anomaly (sigma~ 1.7) of any node
 - b. PM2.5 increased by 47 $\mu\text{g}/\text{m}^3$ on average from the mean+1sigma
2. **Node (1,3) Smoke is transported into the LADCO region via a low-pressure system and cyclonic flow.**
 - a. Near-zero PM anomaly and lower residual value indicate that the fire smoke impact was not over the domain extent, but localized. It is an averaging artifact for a narrow pollutant transport path.
3. **Node (1,0) and Node (1,4) Higher PM residuals are present when there is no pressure dominance or within a transitory state between high and low pressure**



850hPa GEOPOTENTIAL HEIGHT

Notable features:

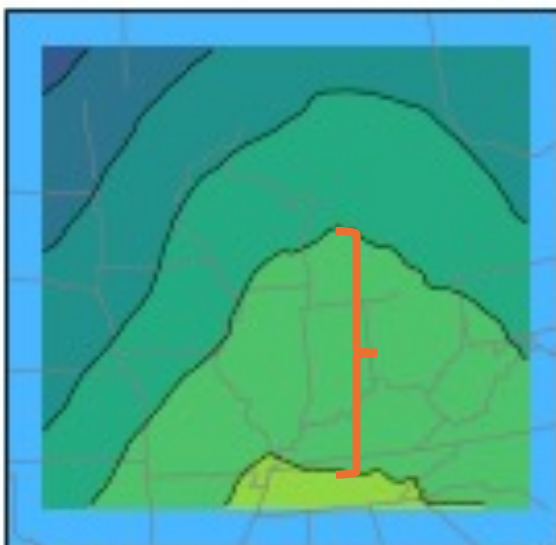
1. Positive PM_{2.5} anomalies are observed to increase when LADCO region is not within a geopotential height gradient leading to (2) and (3) and (4)
2. Greater distance between isohypses (lines of constant geopotential height) appears to correlate with positive PM_{2.5} anomalies in high-pressure dominated nodes (1,0) and (0,1)
3. Vertically stacked high pressure of node (2,0) compared to node (2,1) and node (2,2)
4. For low pressure dominated nodes this looks almost inverse in which nodes (0,3) and (0,4) both have slightly negative PM_{2.5} anomaly.





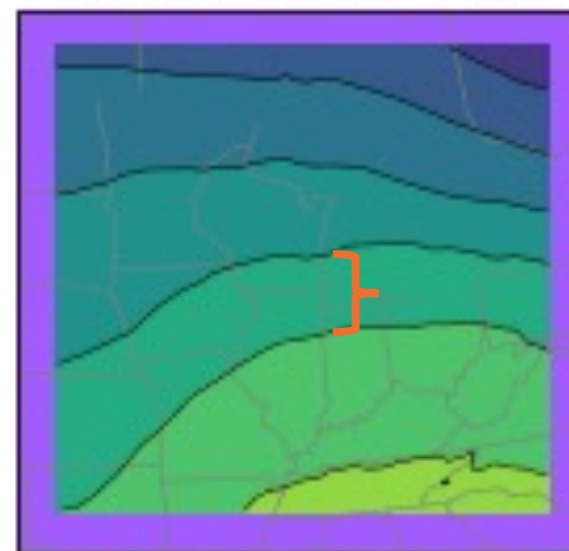
GEOPOTENTIAL GRADIENT EXAMPLE

Node (0, 1), n = 9
Smoke Days = 6
Avg PM = 28.57
Avg PM anom = 1.66
Avg Res1PM = 47.57



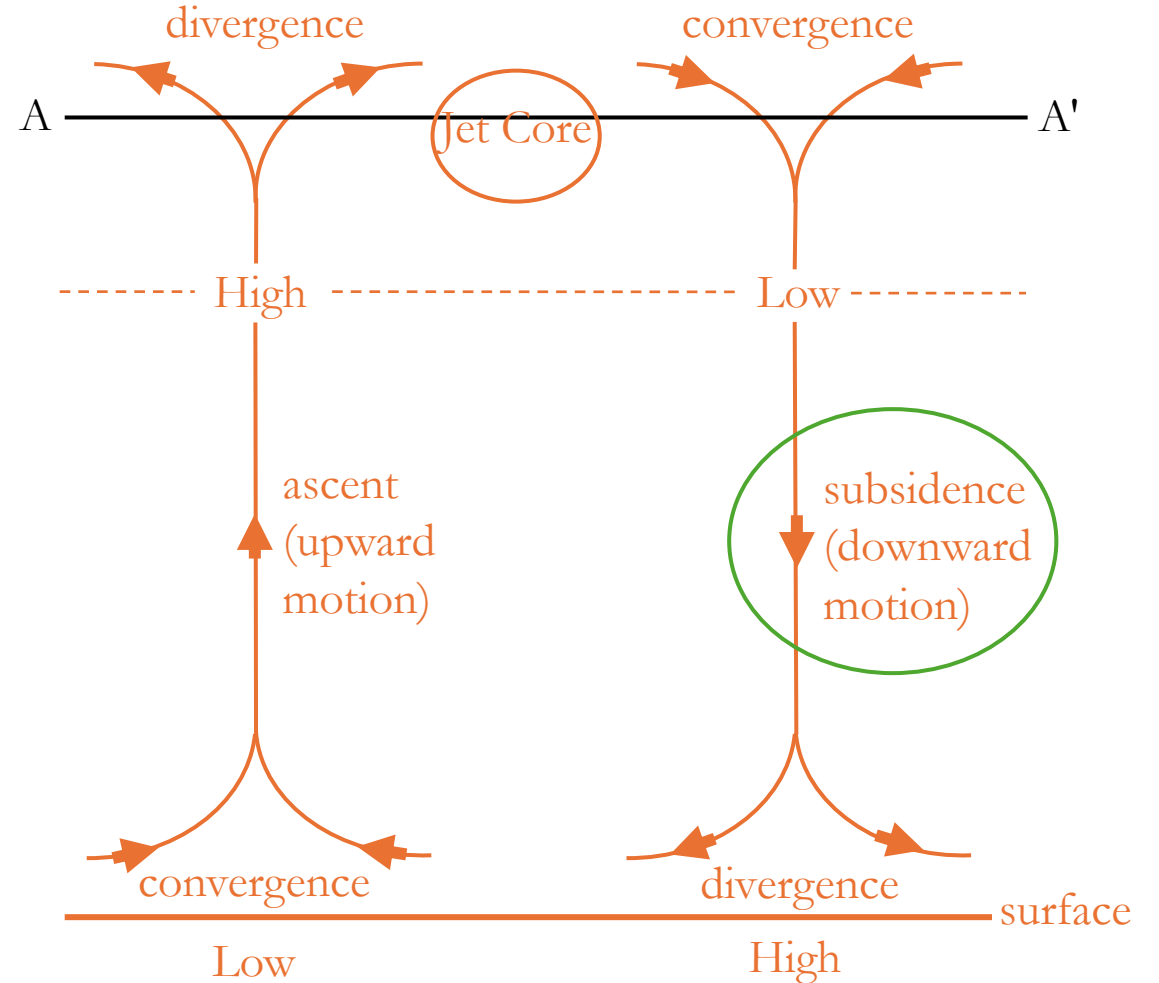
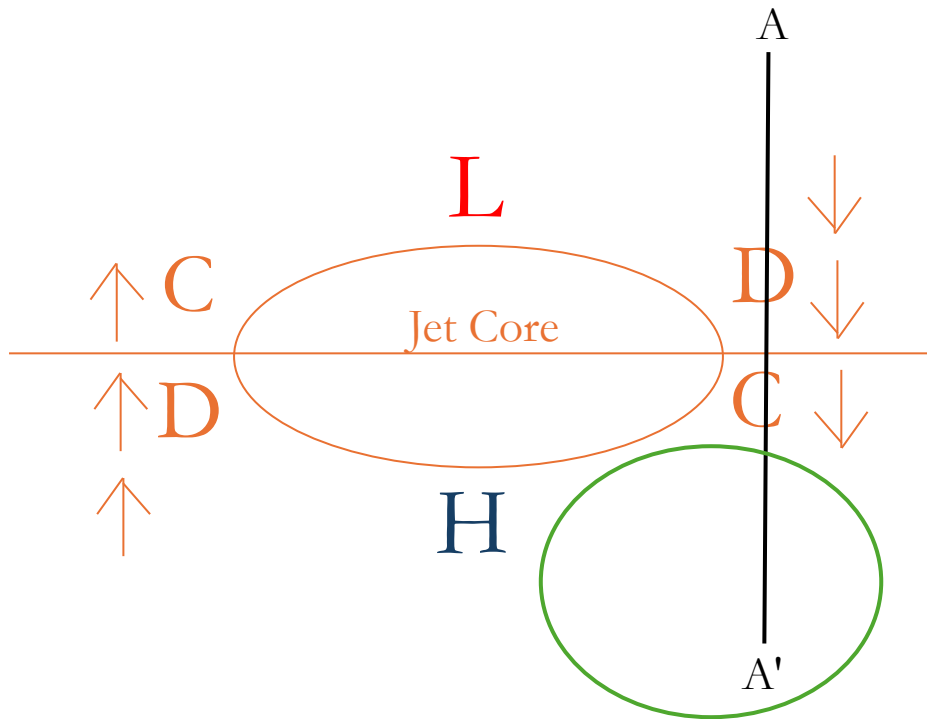
2023 EE Days: ['30', '29', '28']

Node (1, 2), n = 4
Smoke Days = 2
Avg PM = 6.08
Avg PM anom = -0.53
Avg Res1PM = 1.52



JET CORE DYNAMICS

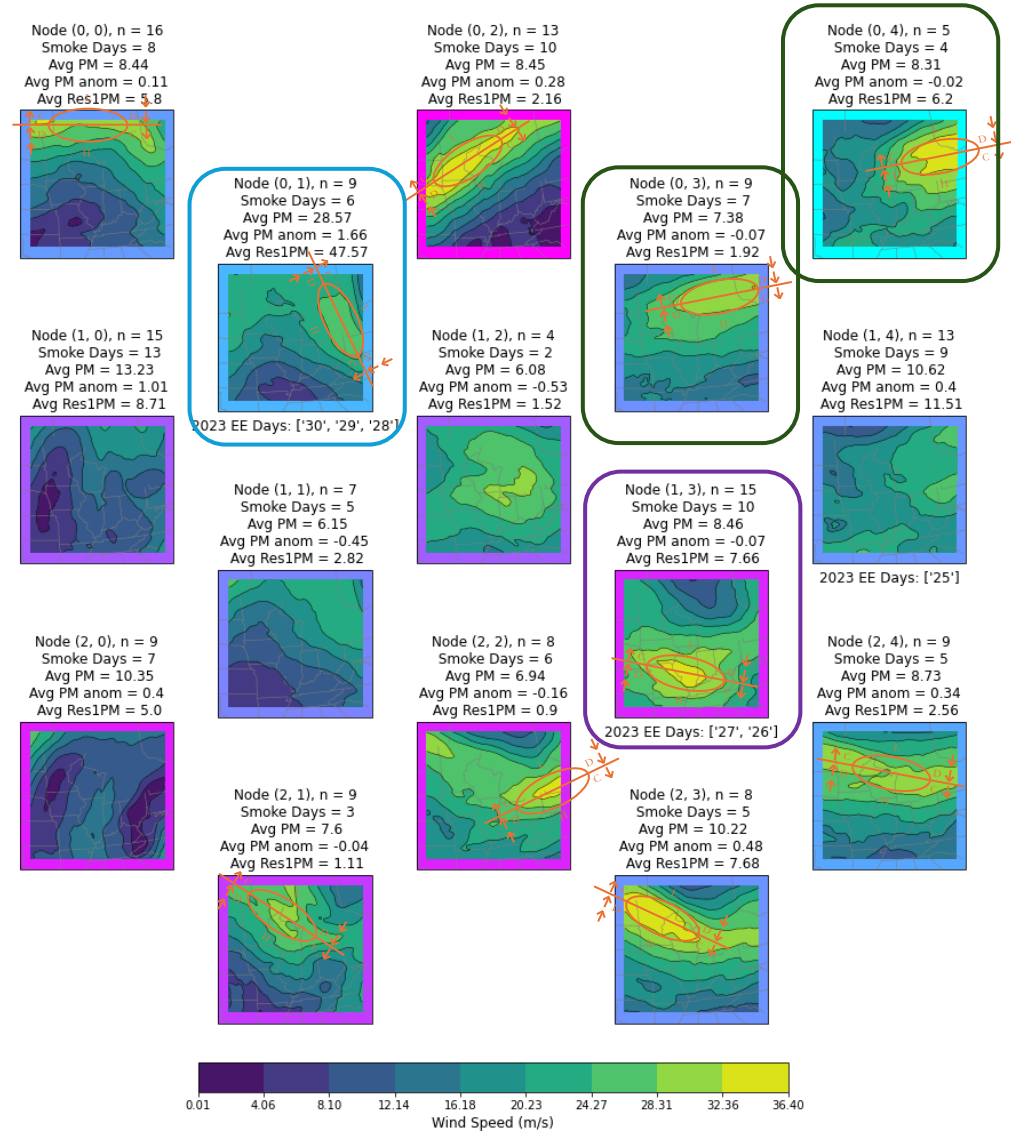
An abbreviated overview:



300hPa WIND SPEED

Notable features:

1. Node (0,1) Upper-level convergence over LADCO region associated with synoptic scale sinking air motion (Air transport to the near surface level)
2. Node (1,3) Upper-level divergence associated with synoptic scale lifting air motion. (Low deepening)
3. Node (1,3) Stronger jet streaks support stronger advective motion aloft
4. Nodes (0,3) and (0,4) have negative PM_{2.5} anomaly and are associated with upper-level divergence.





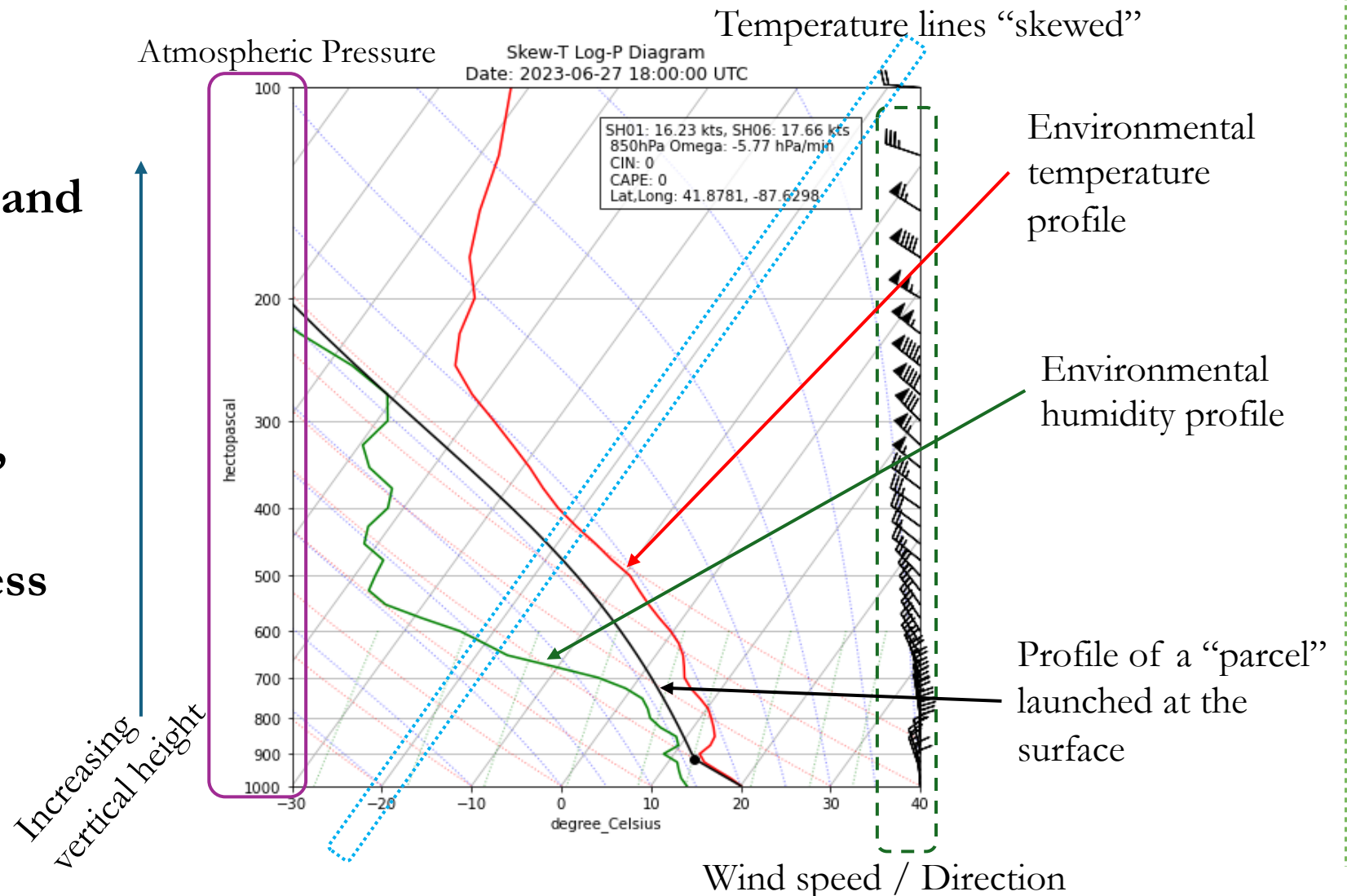
VERTICAL PROFILE ANALYSIS

MODEL DERIVED SOUNDINGS

- We would like to determine if overhead smoke is being transported down to the surface within a specific node
- To do this the generation of a vertical profile would be informative
- However, it is not explicitly an output of the SOM. Although a SOM has been applied to atmospheric soundings in the past: (Nixon et al. 2023)
- We do know, however, which samples are mapped to each node
- Thus, is it possible to create an averaged vertical profile for a particular node based off the samples classified into that node
- The results for LADCO SOM are the following...
 - All soundings are taken from Chicago, IL

QUICK RECAP: SKEW-T BASICS

- Visualization tool for assessing the vertical change in temperature and humidity with height
- Useful for assessing atmospheric stability, temperature inversions, and much more.
- Averaged profiles are less specific but still informative


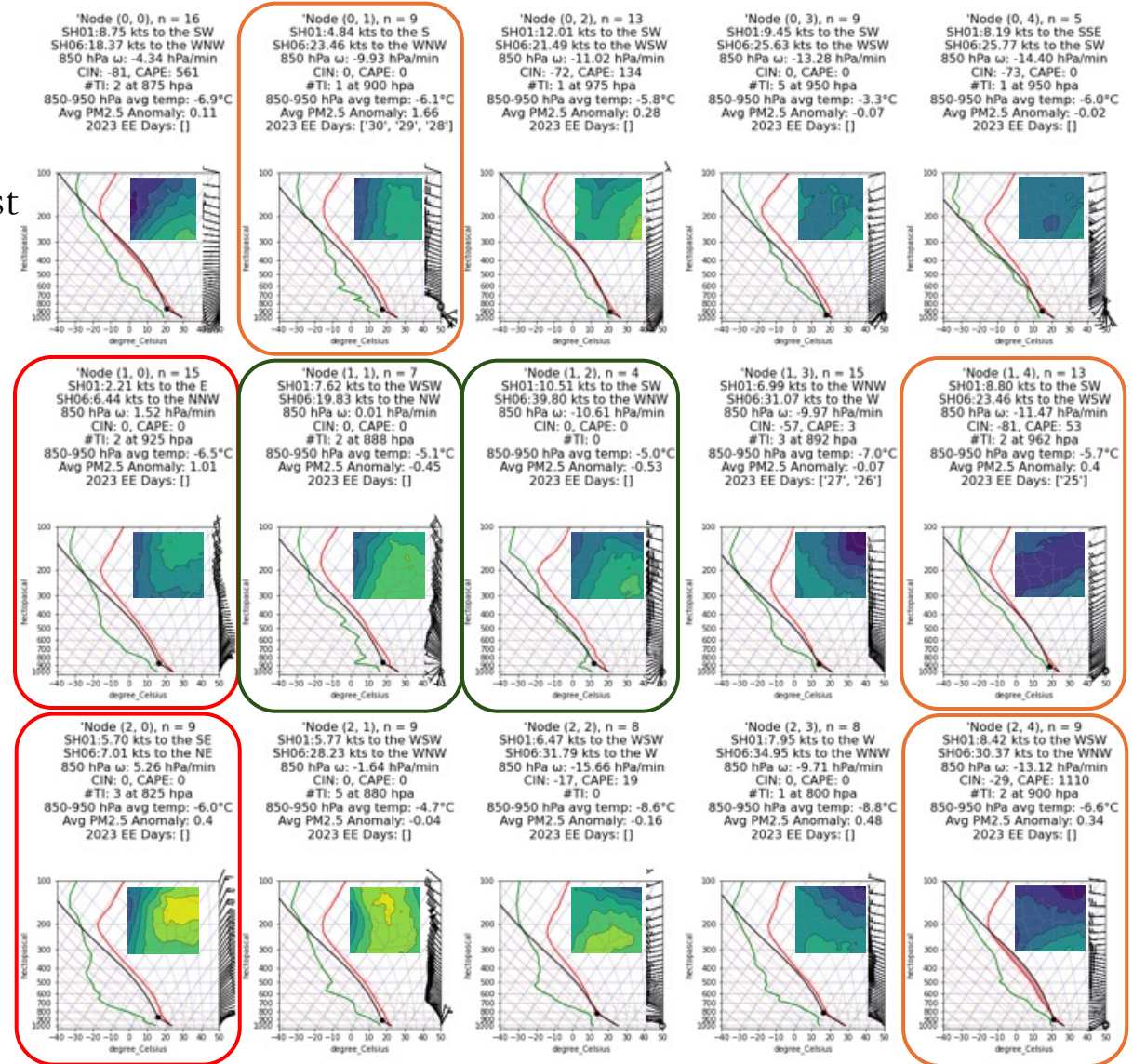


NODE VERTICAL PROFILES

Notable features:

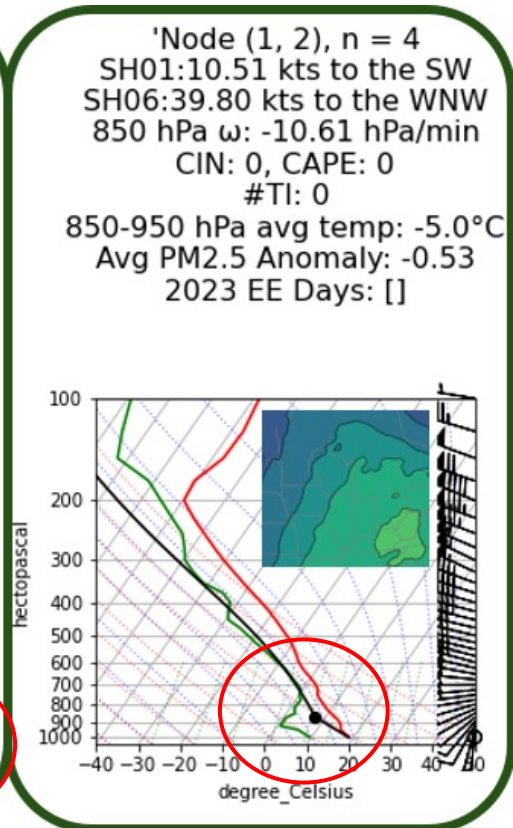
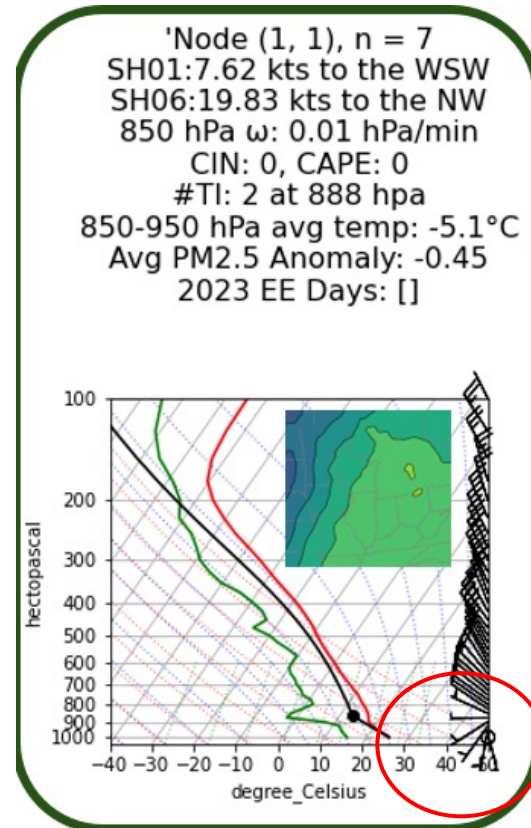
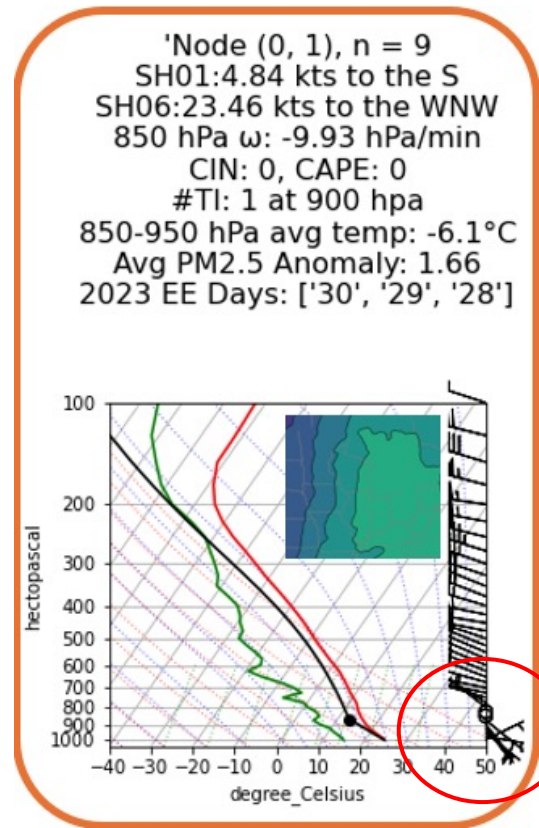
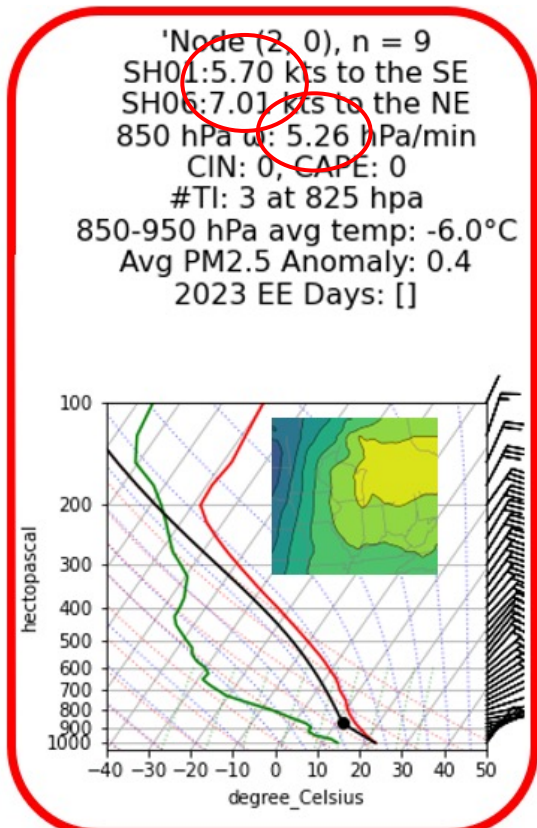
1. Mid level moisture profiles vary greatly
2. Nodes (1,0) and (2,0) both have very positive 850hPa ω
 - Low SH01 and SH06
 - High PM2.5 anomaly
3. High PM2.5 anomaly profiles commonly have calm or eastward winds at the surface
4. Strong negative PM2.5 anomalies have less steep environmental lapse rates and a dry lower levels with moisture aloft
5. There is no definitive indicator
 - However, SOM can make minute distinctions

More moist
More dry



NODE VERTICAL PROFILES CONTINUED





FURTHER EXPLORATION

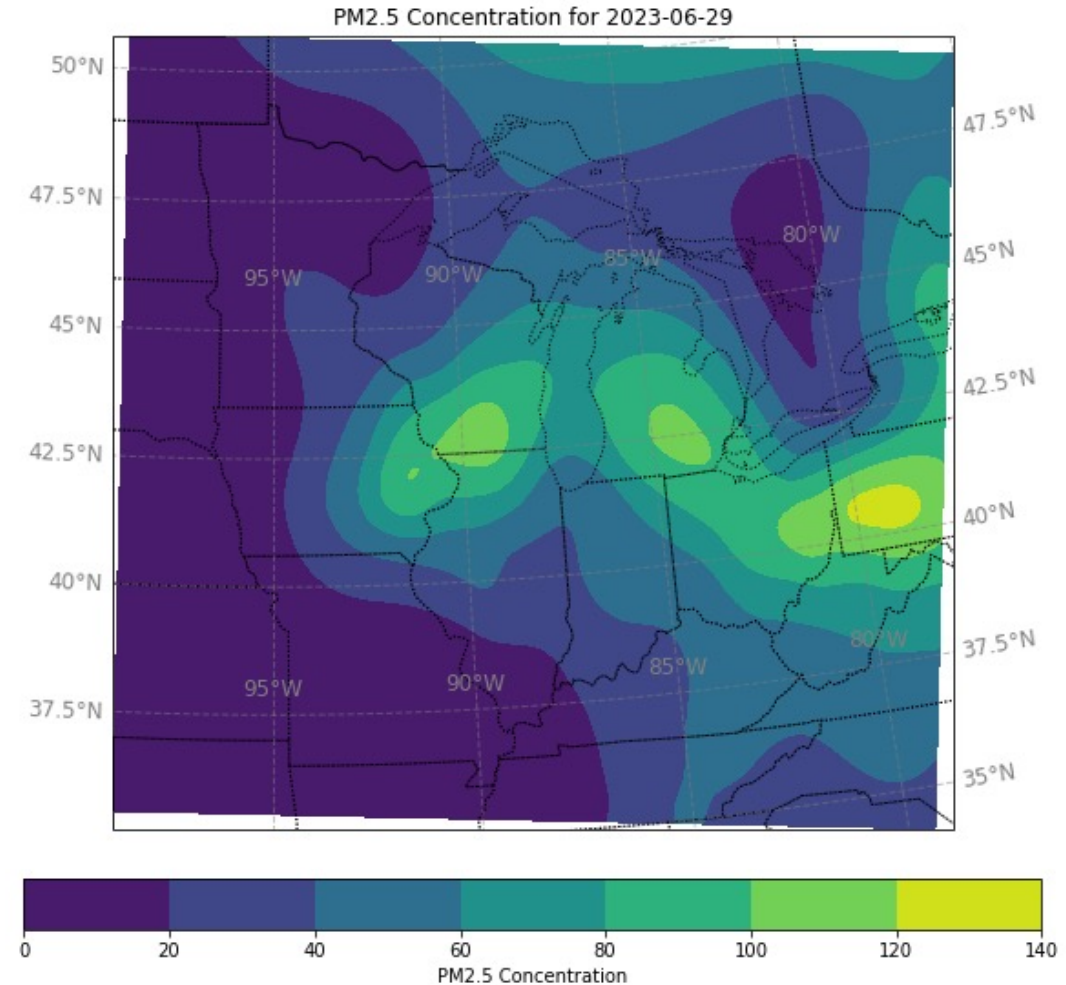
The “Extra Credit”

MOTIVATIONS FOR ADDING PM_{2.5} DATA

- Informed by the SOM, high PM_{2.5} conditions occur in a variety of different meteorological setups
- These setups have common conditions, however...
- We wanted to test if we could characterize specific properties of high PM_{2.5} episodes we would like to now give LADCO SOM node knowledge of PM concentrations at the surface
- We ran a test SOM adding a PM_{2.5} variable into the SOM. But due to our high dimensionality it needed to be a similar shape of our meteorological data

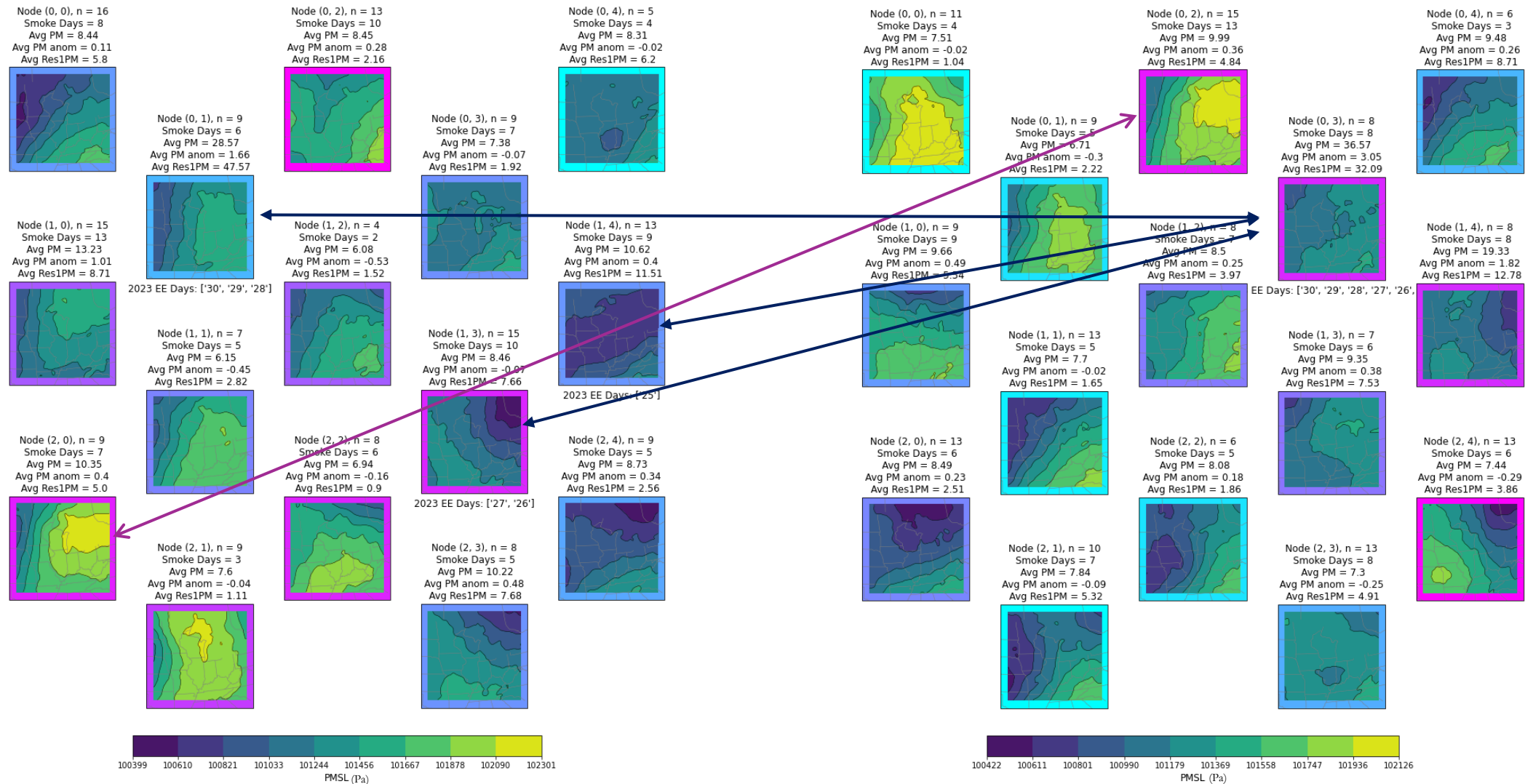
ADDING PM_{2.5} DATA INTO SOM INPUT

- Meteorology data plus “Krigged” (spatially interpolated) PM_{2.5} data
- Krigged PM_{2.5} field from PM_{2.5} concentrations measured at AQS monitors
- All June days from 2019-2023 except for 2023-06-01 due to an incomplete HRRR run



SOM WITH MET ONLY

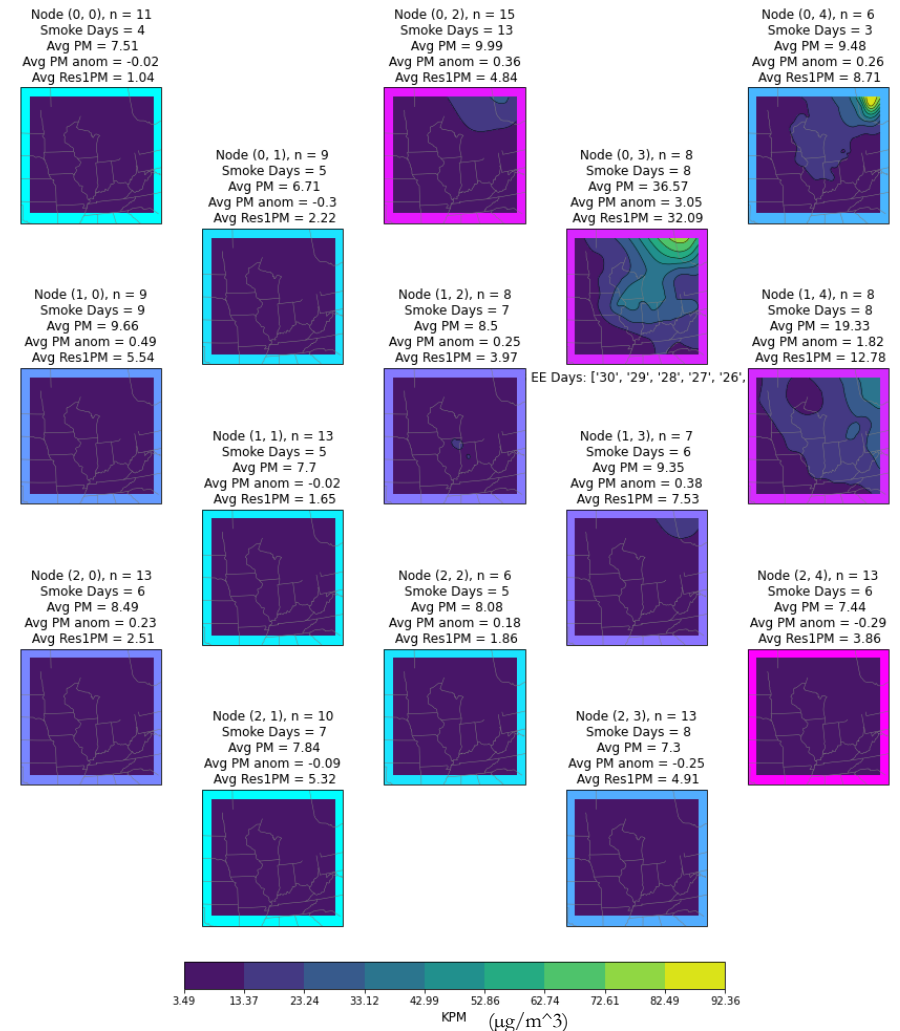
SOM WITH MET+PM2.5



KRIGGED PM_{2.5} SOM WEIGHTS

Notable features:

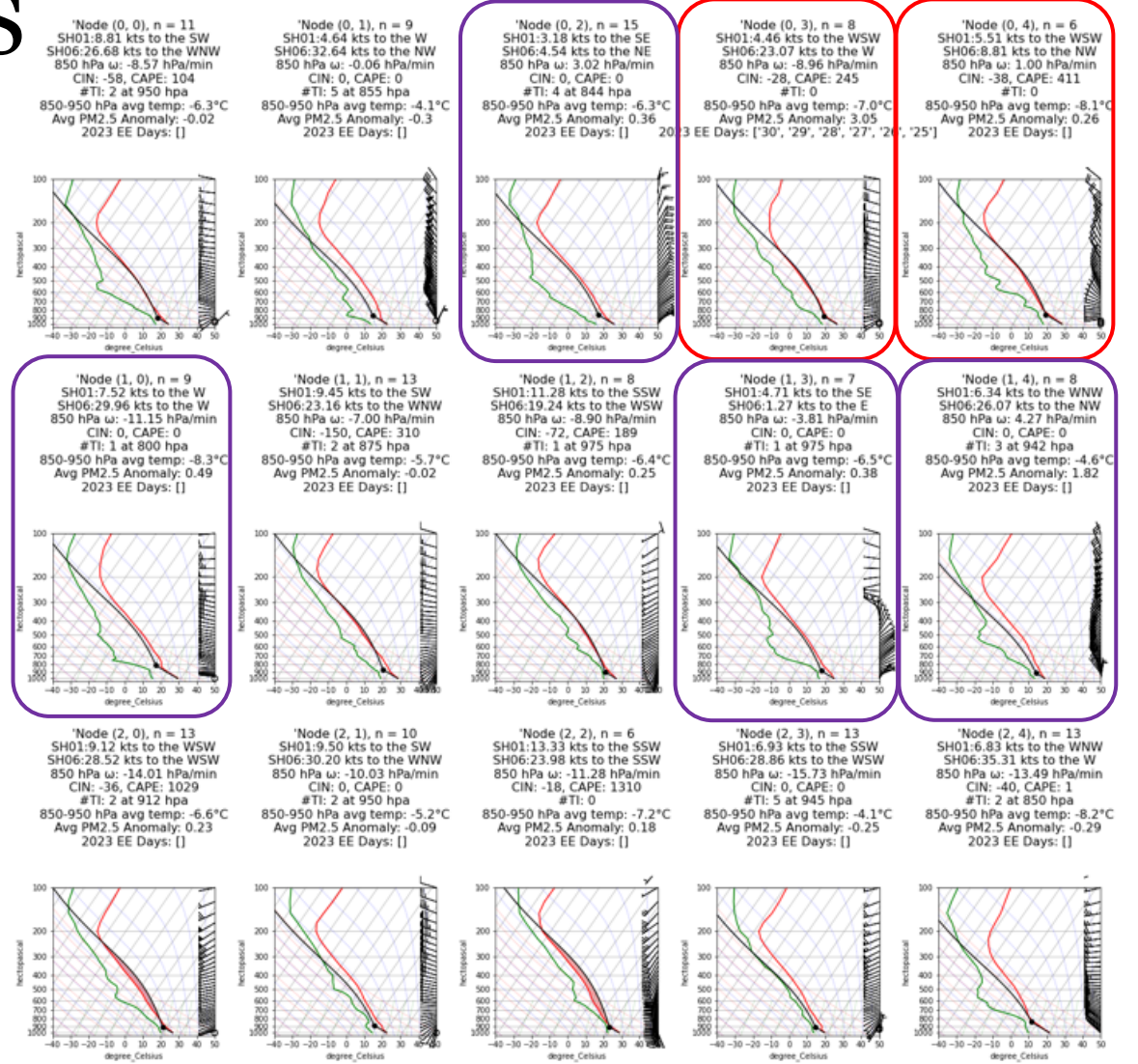
1. “Equal interval” symbology is dominated by the 2023 event
2. MN appears to experience less PM_{2.5} impacts in the month of June compared to the rest of the LADCO states
3. Notice the scale!



KPM VERTICAL PROFILES

Notable features:

1. High PM2.5 events characterized by low level shear environments
2. Additionally, high PM2.5 events seem less dependent on temperature inversions and more correlated with stagnant conditions
 - Although this could also be due to the location of the point sounding
3. High PM2.5 conditions also appear now to be clearly associated with dryer conditions in the mid levels



LADCO SOM CONCLUSIONS

- In general, weak high pressure systems are associated with stagnation conditions and higher PM_{2.5} concentrations
- Strong low pressure systems are capable of maintaining long range pollutant transport pathways in the upper atmosphere
- High PM_{2.5} nodes are strongly correlated with upper-level convergence which can lead to higher PM_{2.5} impact at the surface and slower jet streak motion which can diagnose less dynamical systems
- Considering solely meteorological variables may not present the full picture and significantly elevated PM_{2.5} impacts have occurred within a variety of meteorological modes
 - Improved PM_{2.5} integration
 - Atmospheric chemistry influences

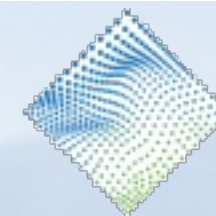
SOM APPLICATIONS TO FORECASTING

- A potential use of the LADCO SOM is in air quality forecasting
- Air quality forecasters can leverage the results from LADCO SOM to examine if PM_{2.5} impacts are typical or atypical for a June day in the LADCO region given the region's current meteorological conditions
- Further refinement to LADCO SOM may enable undiscovered insights into harder to detect meteorological relationships with trends in air quality



NEXT STEPS FOR SOM ANALYSIS

- **Additional feature analysis and examination may inform an optimal blend of input variables for different SOM configurations**
- **Considering alternative climatological periods or varying sets of training data could glean more specific (or broader if desired) trends given the right input data**
- **The meteorology + krigged PM_{2.5} SOM could be improved further. With improved knowledge of PM_{2.5} conditions, more actionable results for determining the relative anomaly of current (or a specific event's) meteorological conditions can be achieved**



LADCO | LAKE MICHIGAN
AIR DIRECTORS CONSORTIUM

Email: victor.geiser@colostate.edu

Questions?



Backup Slides!

SOM ALGORITHM

• Three main concepts:

1. **Best Matching Unit / Activation Distance**
2. **Neighborhood Function / Decay**
3. **Learning Rate / Decay**

Input vector	Initial 3x2 SOM	
$[1, 2, 3]$	$[5, -3, 7]$	$[8, 8, 8]$
	$[8, -4, 2]$	$[13, 4, 9]$
	$[0, 3, 5]$	$[5, 6, -5]$

Input vector	Initial 3x2 SOM	
$[1, 2, 3]$	$[5, -3, 7]$	$[8, 8, 8]$
	$[8, -4, 2]$	$[13, 4, 9]$
	$[0, 3, 5]$	$[5, 6, -5]$

Initial 3x2 SO	SOM weights after 1 iteration
$[5, -3, 7]$ $[8, 8, 8]$	$[4.5, -2.5, 6.5]$ $[7.5, 7.5, 7.5]$
$[8, -4, 2]$ $[13, 4, 9]$ ->	$[7, -3, 1]$ $[12, 3, 8]$
$[0, 3, 5]$ $[5, 6, -5]$	$[1, 2, 4]$ $[4, 5, -4]$

LADCO SOM DISTANCE MATRIX

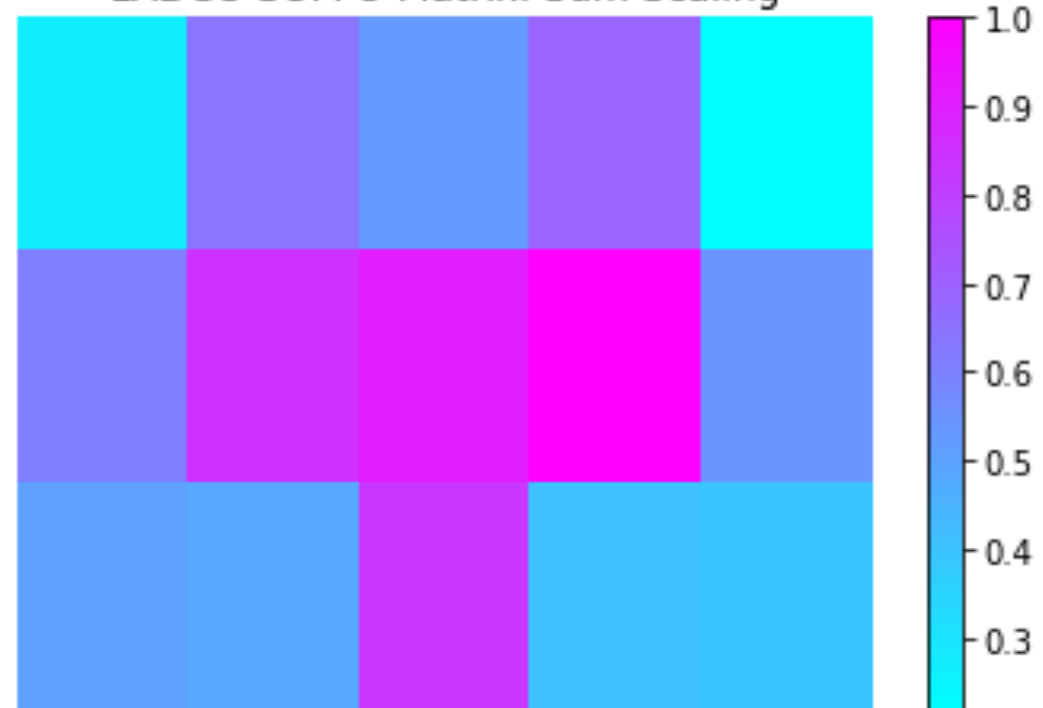
Examines relative node heterogeneity
(used in LADCO SOM)

LADCO SOM U-Matrix: Mean Scaling



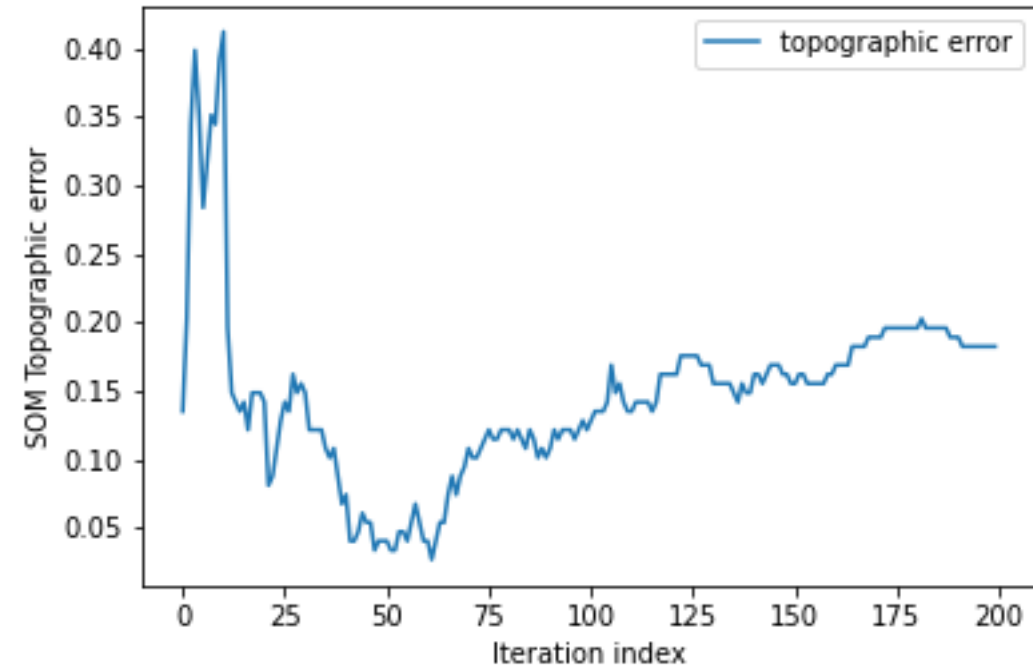
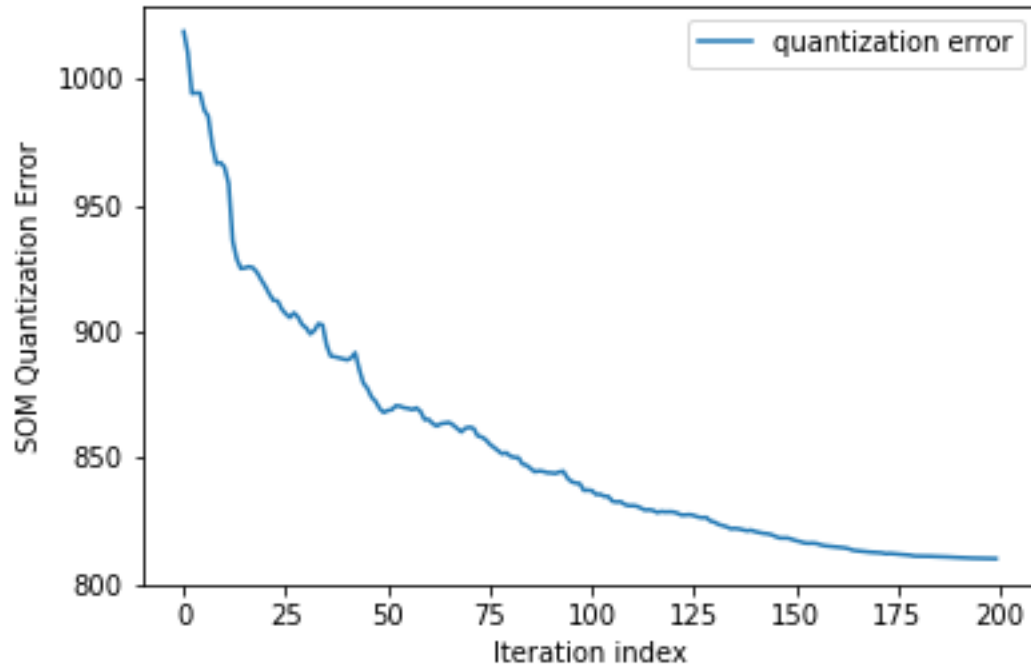
Informs boundaries between
potential node clusters

LADCO SOM U-Matrix: Sum Scaling



SOM LEARNING CURVES

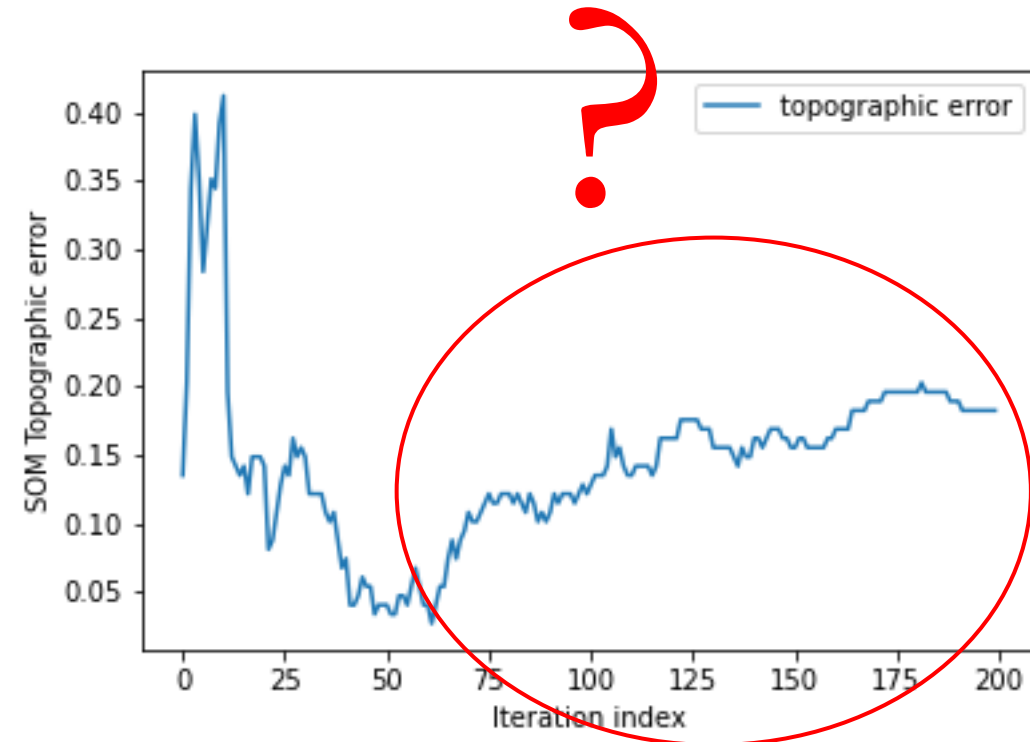
- This is how we can tell our SOM is “learning”!





SOM LEARNING CURVES CONT. 1 of 2

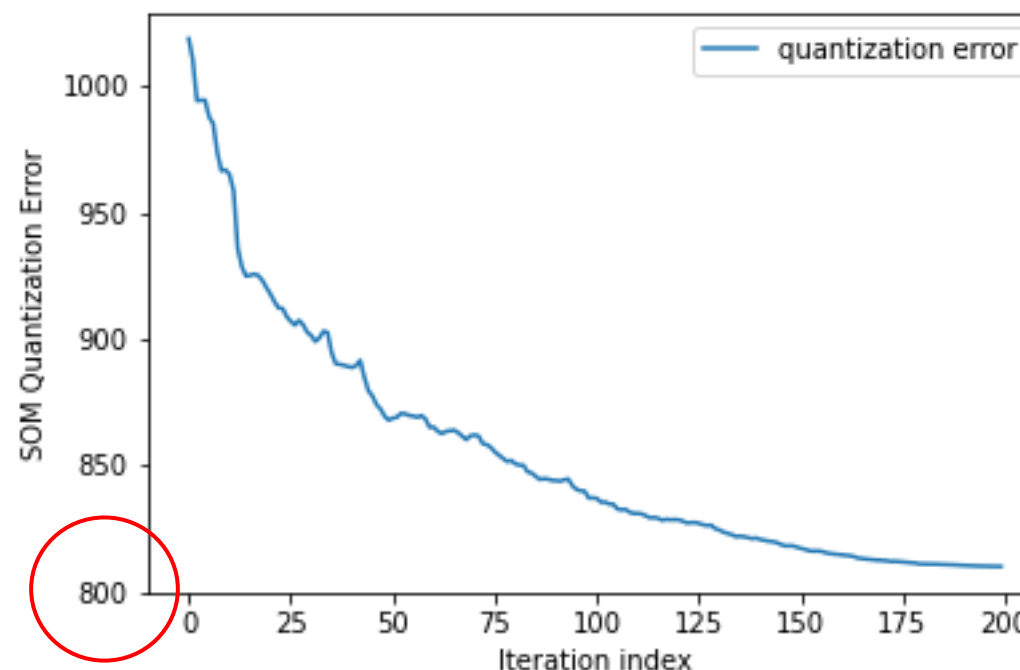
- Why does our topographic error increase in the middle of our training iterations?
 - Forest, F., et al (2021)
- “Topographic error shows the trade-off between self-organization ... and the resulting clustering quality.”
- “A practitioner could thus choose to use an early stopping strategy ... but it would harm the quality of the clustering.”



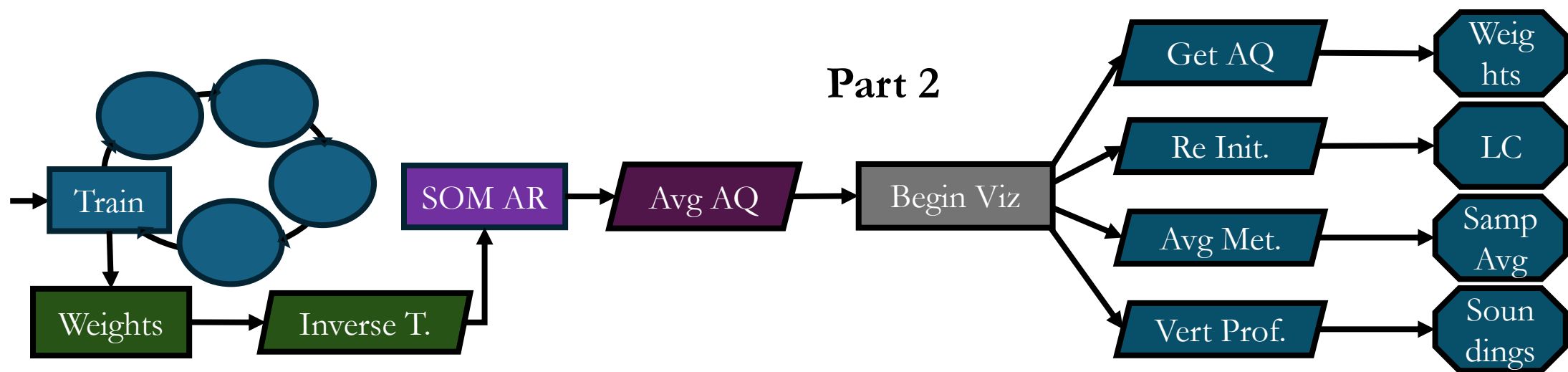
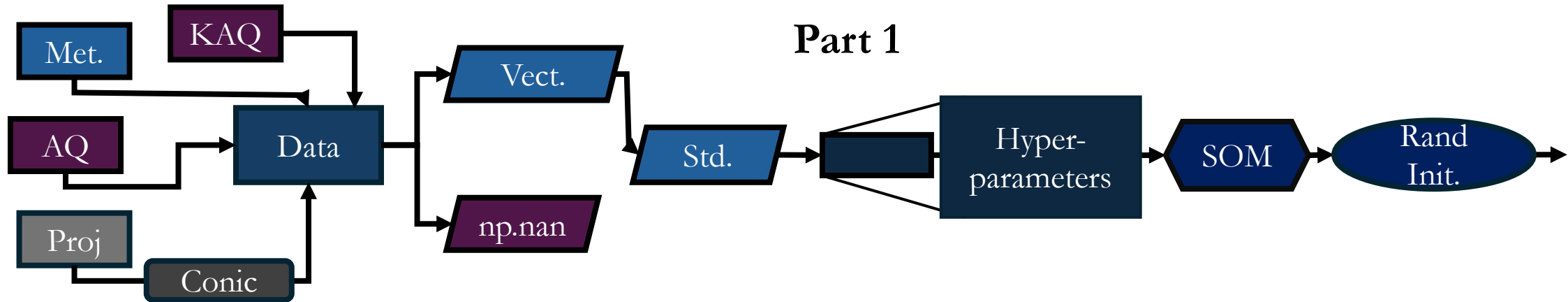


SOM LEARNING CURVES CONT. 2 of 2

- Why does quantization error stop at 800? That seems abnormally high?
- Correct!
- The answer lies in how the data is represented in the SOM
- Our data is vectorized
- 420 latitude x 444 longitude = 186480 (for one variable) ... * 6 = 1118880 “columns” for one sample



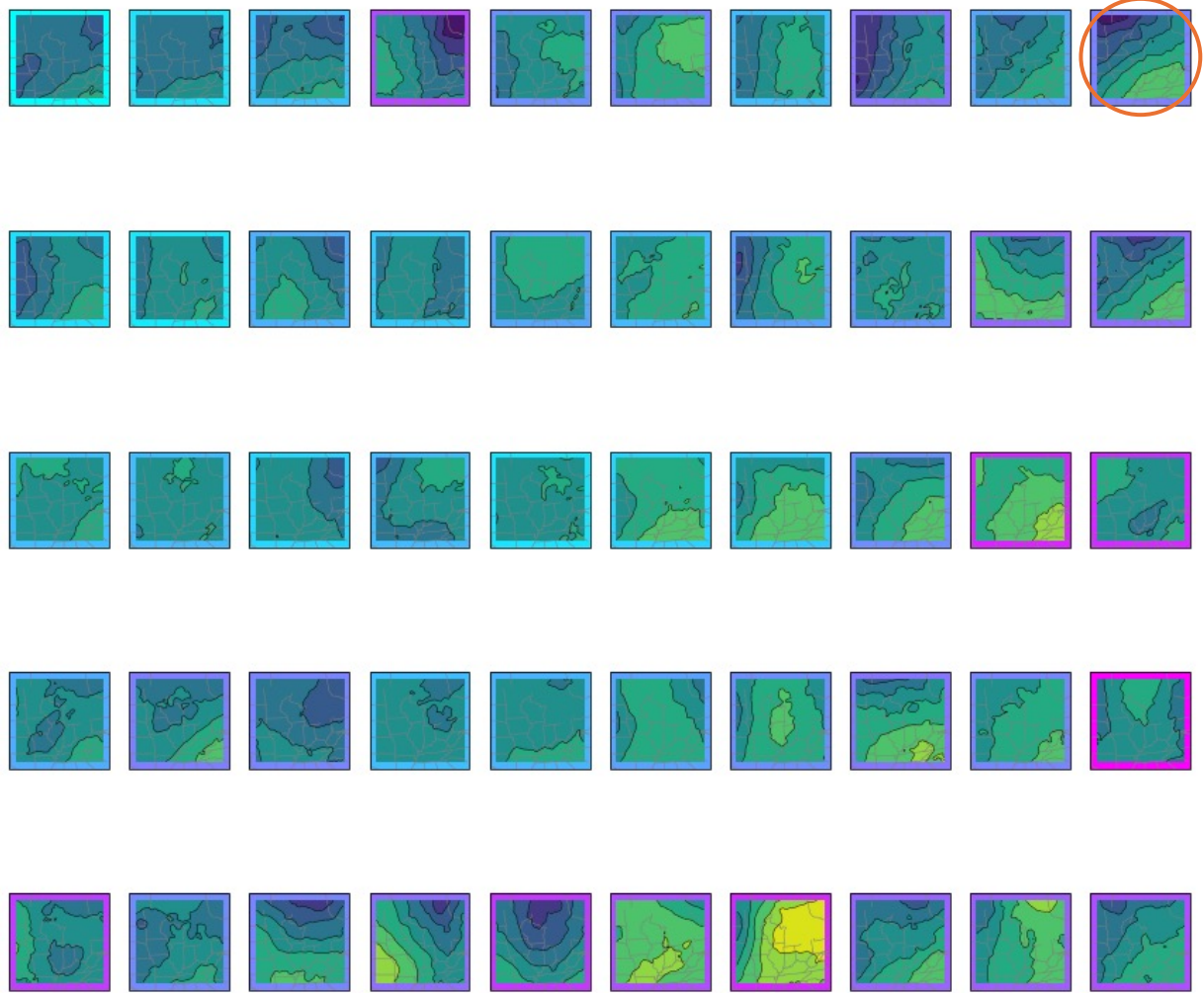
FULL PYTHON WORKFLOW



PRESERVING TOPOLOGY

- More similar “nodes” are closer to one another within the SOM

This is a single SOM “node”





EXPLANATION OF PARAMETERS

- “SH01” – 0-1km wind shear and direction
- “SH06” – 0-6km wind shear and direction
- “850hPa ω ” – Vertical velocity above the PBL*
- “CIN, CAPE” – Convective Inhibition, Convective Available Potential Energy
- “#TI” – Number of temperature inversions ($2^{\circ}\text{C} / 100\text{hPa}$)
- “850-950hPa avg temp” – Average temperature difference between the 850 hPa and 950hPa levels.
- “Avg PM2.5 Anomaly” – The same as before

500hPa RELATIVE HUMIDITY

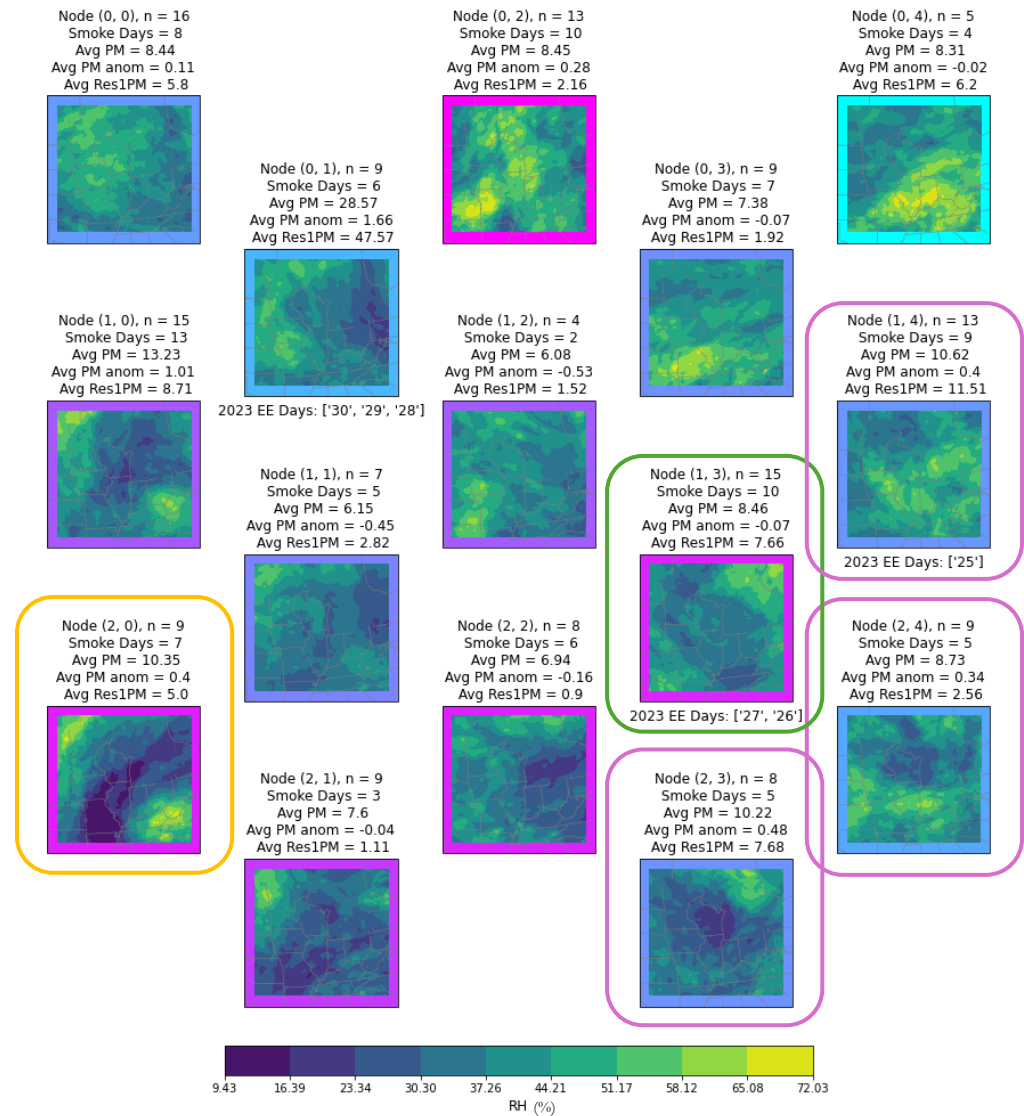
Notable features:

1. No direct and significant statistical relationship between avg 500hPa RH and PM Anomaly

- Spearman Correlation = -0.067 (slightly negative) P-value = 0.81 (>>0.05 – trend is not very significant)
 - Higher RH → Lower PM

2. High PM2.5 can occur in a variety of different mid-level moisture profiles

- Node (2,0) Displays a dry conveyor belt associated with elevated PM Anomaly (increase of 5mg/m³). Node (2,1) to a lesser extent
- Node (1,3) Low pressure dominated node with more N/S transport and lower anomaly
- Nodes (1,4) (2,4) (2,3) Low pressure dominated nodes with more W/E transport and higher anomaly



*Lower lake surface temperatures are masked out of visualization to make land surface temperatures easier to read

SURFACE TEMPERATURE

Notable features:

1. Two primary surface temperature patterns

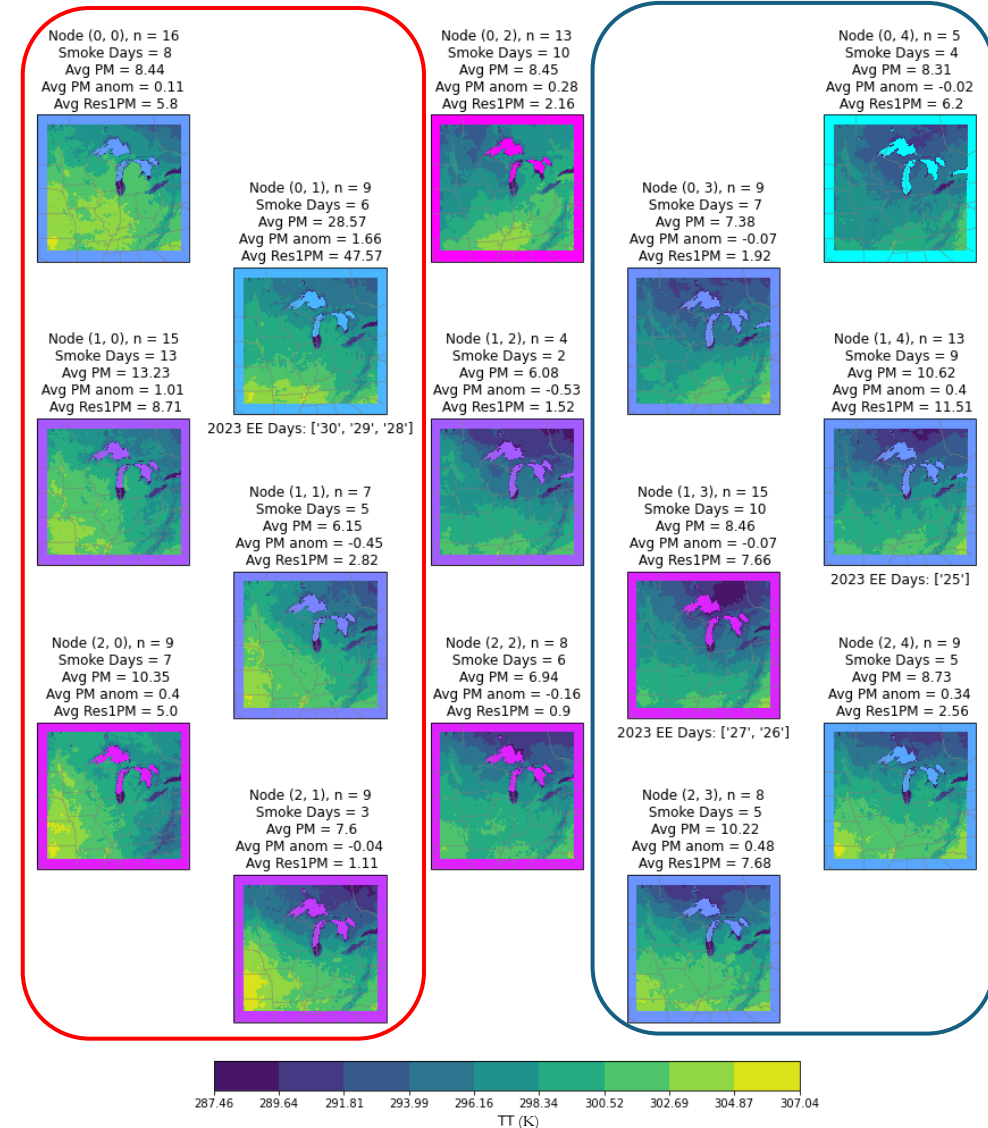
a. **Warmer southern temperatures extending northward**

b. **Cooler northern temperatures extending southward**

- Distinctions between nodes are primarily on the order of the magnitude of this extension
- Middle column is an exception!

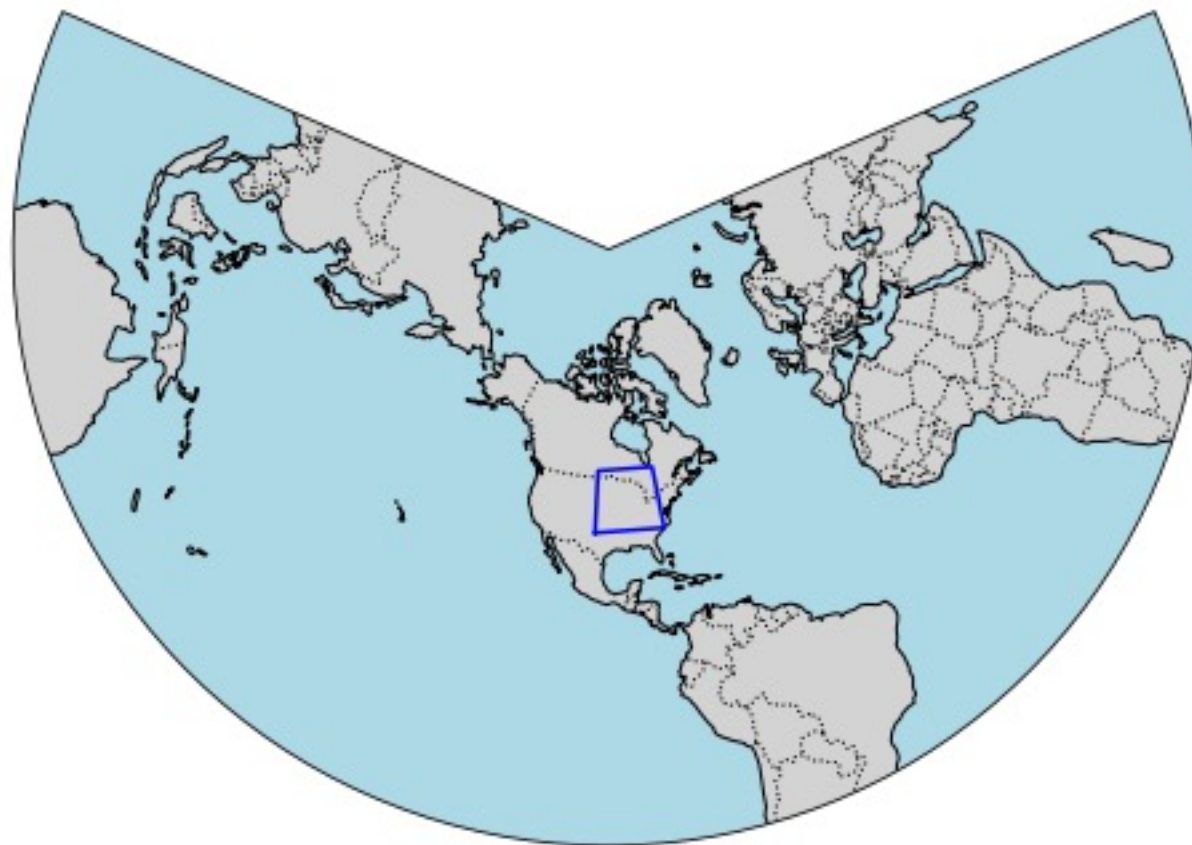
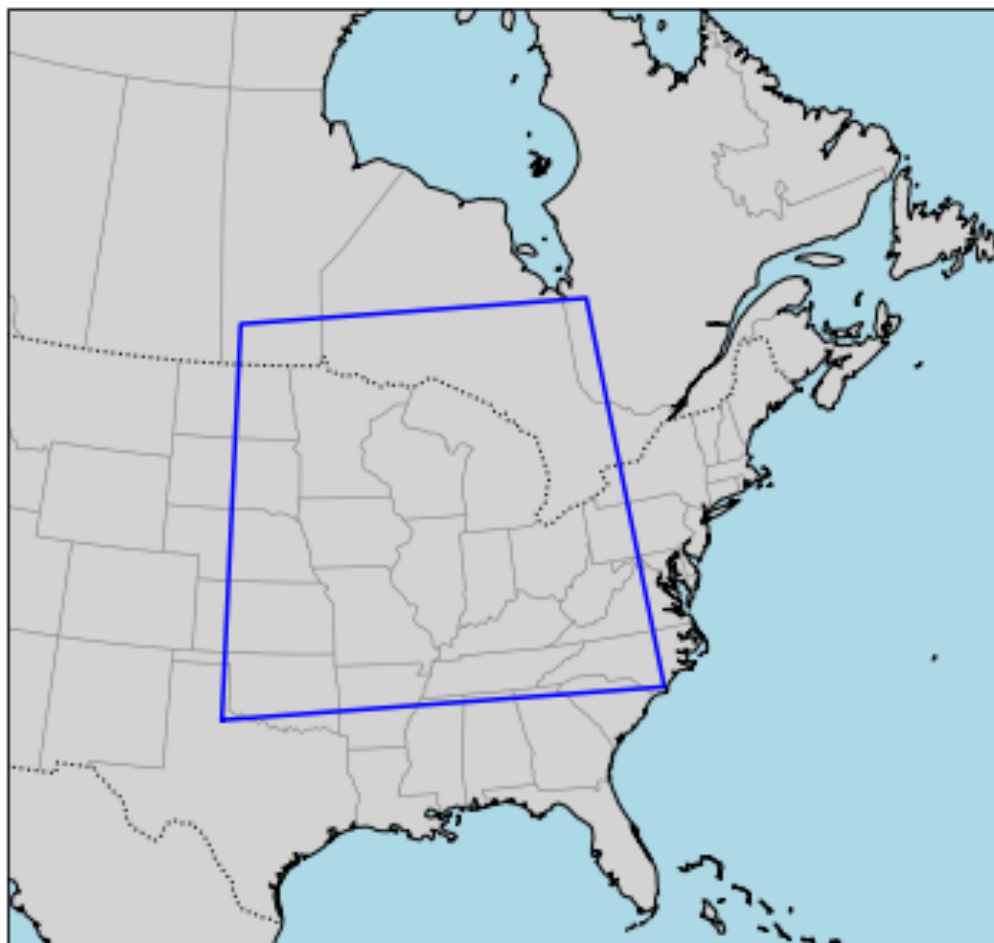
2. More significant statistical relationship between avg surface temperature and PM Anomaly

- Spearman Correlation = 0.44 (moderately positive) P-value = 0.099 (significant at the 10% range)



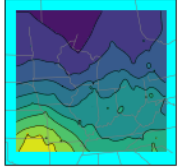


DATA EXTENT

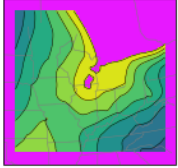




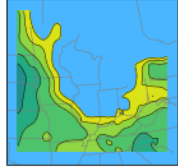
Node (0, 0), n = 11
Smoke Days = 4
Avg PM = 7.51
Avg PM anom = -0.02
Avg Res1PM = 1.04



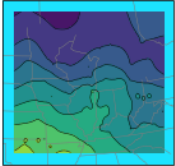
Node (0, 2), n = 15
Smoke Days = 13
Avg PM = 9.99
Avg PM anom = 0.36
Avg Res1PM = 4.84



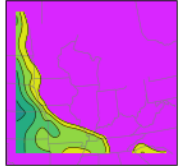
Node (0, 4), n = 6
Smoke Days = 3
Avg PM = 9.48
Avg PM anom = 0.26
Avg Res1PM = 8.71



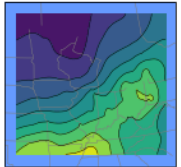
Node (0, 1), n = 9
Smoke Days = 5
Avg PM = 6.71
Avg PM anom = -0.3
Avg Res1PM = 2.22



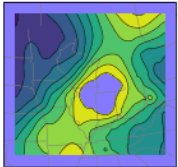
Node (0, 3), n = 8
Smoke Days = 8
Avg PM = 36.57
Avg PM anom = 3.05
Avg Res1PM = 32.09



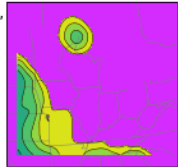
Node (1, 0), n = 9
Smoke Days = 9
Avg PM = 9.66
Avg PM anom = 0.49
Avg Res1PM = 5.54



Node (1, 2), n = 8
Smoke Days = 7
Avg PM = 8.5
Avg PM anom = 0.25
Avg Res1PM = 3.97

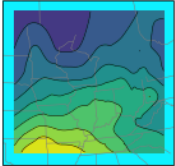


Node (1, 4), n = 8
Smoke Days = 8
Avg PM = 19.33
Avg PM anom = 1.82
Avg Res1PM = 12.78

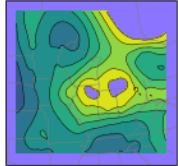


EE Days: ['30', '29', '28', '27', '26']

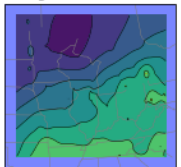
Node (1, 1), n = 13
Smoke Days = 5
Avg PM = 7.7
Avg PM anom = -0.02
Avg Res1PM = 1.65



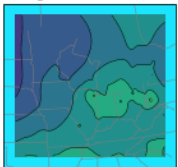
Node (1, 3), n = 7
Smoke Days = 6
Avg PM = 9.35
Avg PM anom = 0.38
Avg Res1PM = 7.53



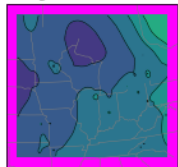
Node (2, 0), n = 13
Smoke Days = 6
Avg PM = 8.49
Avg PM anom = 0.23
Avg Res1PM = 2.51



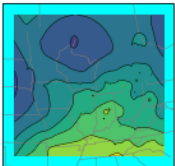
Node (2, 2), n = 6
Smoke Days = 5
Avg PM = 8.08
Avg PM anom = 0.18
Avg Res1PM = 1.86



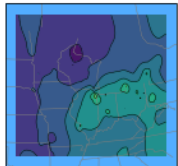
Node (2, 4), n = 13
Smoke Days = 6
Avg PM = 7.44
Avg PM anom = -0.29
Avg Res1PM = 3.86



Node (2, 1), n = 10
Smoke Days = 7
Avg PM = 7.84
Avg PM anom = -0.09
Avg Res1PM = 5.32



Node (2, 3), n = 13
Smoke Days = 8
Avg PM = 7.3
Avg PM anom = -0.25
Avg Res1PM = 4.91



Adjusted Symbology of met + krigged pm2.5 SOM for visualization purposes

