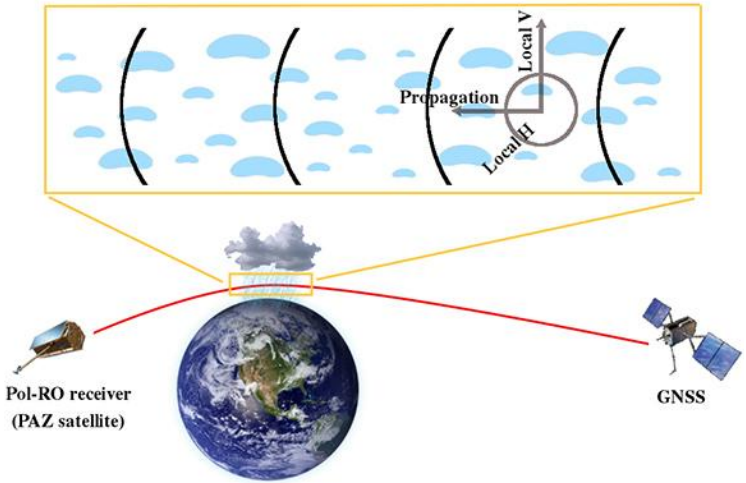
Cluster Analysis of PRO and Liquid & Ice Water Paths from GPM Microwave Data



Jonas Katona

NASA Jet Propulsion Laboratory, CA, USA
Applied Mathematics Program, Yale University, CT, USA

Joint work with Manuel de la Torre (JPL) and Terence Kubar (JPL/UCLA)

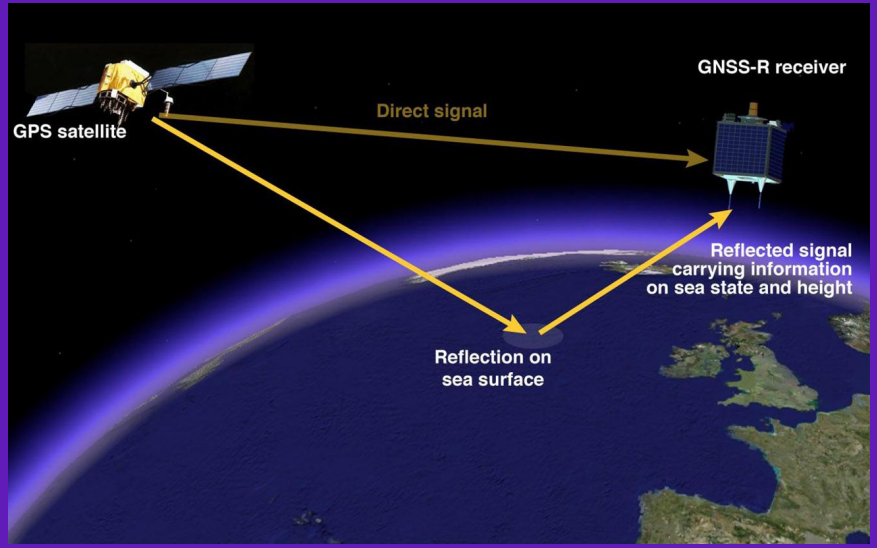
Outline

1. **Introduction to GNSS and PRO and motivation**
2. Data and methods
3. Results and clustering analyses
4. Summary

GNSS

GNSS (Global Navigation Satellite System) refers to the collection of Earth-orbiting satellites that periodically send radio signals indicating their positions.

Satellites occult from a low Earth-orbiting satellite with a GNSS receiver. The GNSS radio signal received has been refracted and bent by the atmosphere.



occult (v.): to cut off from view by interposing something.

GNSS

The bending angle is caused by the atmospheric refractivity gradient in the region through which the signal traveled

- The degree of bending can be calculated using geometry and the change in frequency of the signal between when it is emitted and received (Doppler shift)

For a usual atmospheric air composition with approximately 78 percent nitrogen and 21 percent oxygen containing water, refractivity N is equal to

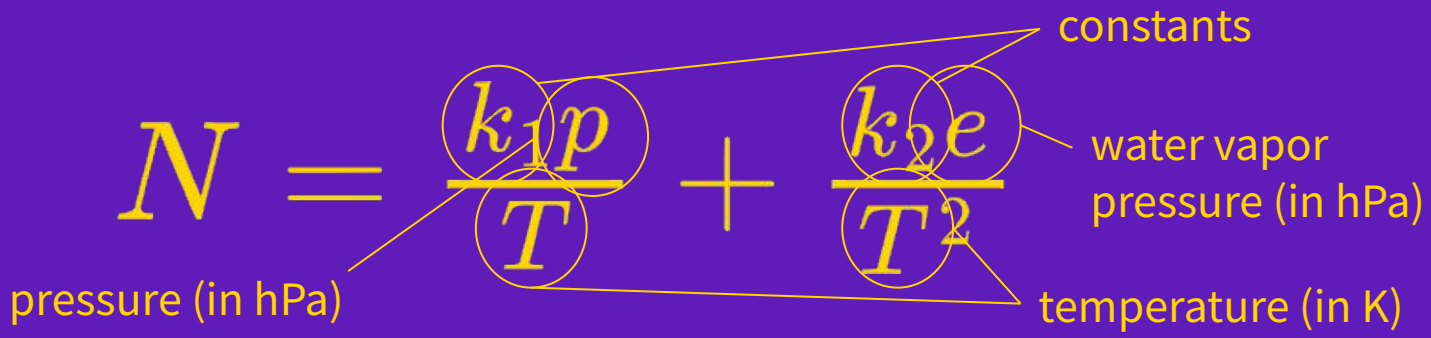
$$N = \frac{k_1 p}{T} + \frac{k_2 e}{T^2}$$

GNSS

The bending angle is caused by the atmospheric refractivity gradient in the region through which the signal traveled

- The degree of bending can be calculated using geometry and the change in frequency of the signal between when it is emitted and received (Doppler shift)

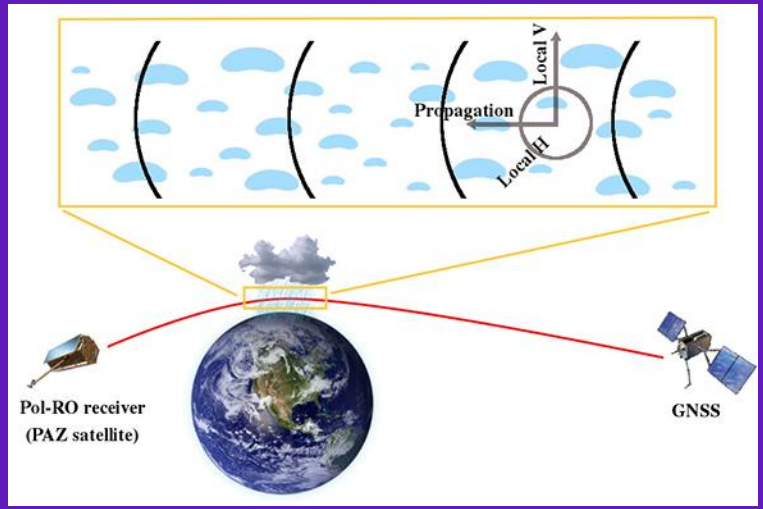
For a usual atmospheric air composition with approximately 78 percent nitrogen and 21 percent oxygen containing water, refractivity N is equal to



PRO (Polarimetric radio occultations)

Atmospheric anisotropies like precipitating droplets and ice crystals induce a phase difference between the horizontal (H) and vertical (V) components of a circularly polarized radio signal.

$\Delta\phi$ (the polarimetric phase difference) is the difference between H and V and can measure the amount of precipitation or ice.



Motivation of this study

Main goal: Better understand the relationships between $\Delta\phi$ and thermodynamic variables of interest and clarify how these can be used to inform predictions better.

- In particular, how can $\Delta\phi$ be related with the presence of ice or precipitation?
- Do different types of clouds help relate $\Delta\phi$ with other precipitation- or moisture-related variables such as liquid water path (LWP), ice water path (IWP), and water vapor pressure?

We explore the ability of cluster analysis to help achieve these goals.

Outline

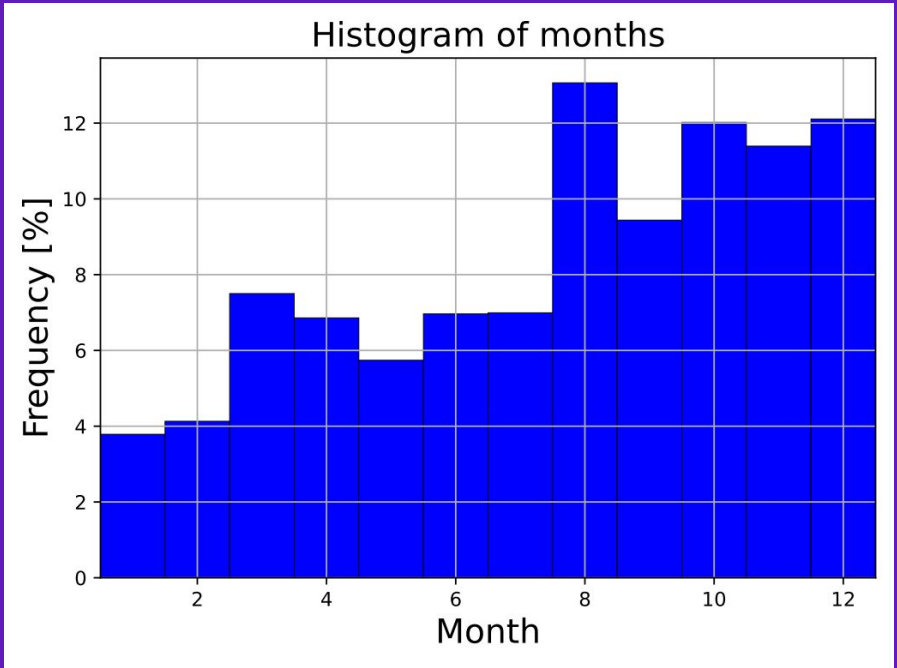
1. Introduction to GNSS and PRO and motivation
2. **Data and methods**
3. Results and clustering analyses
4. Summary

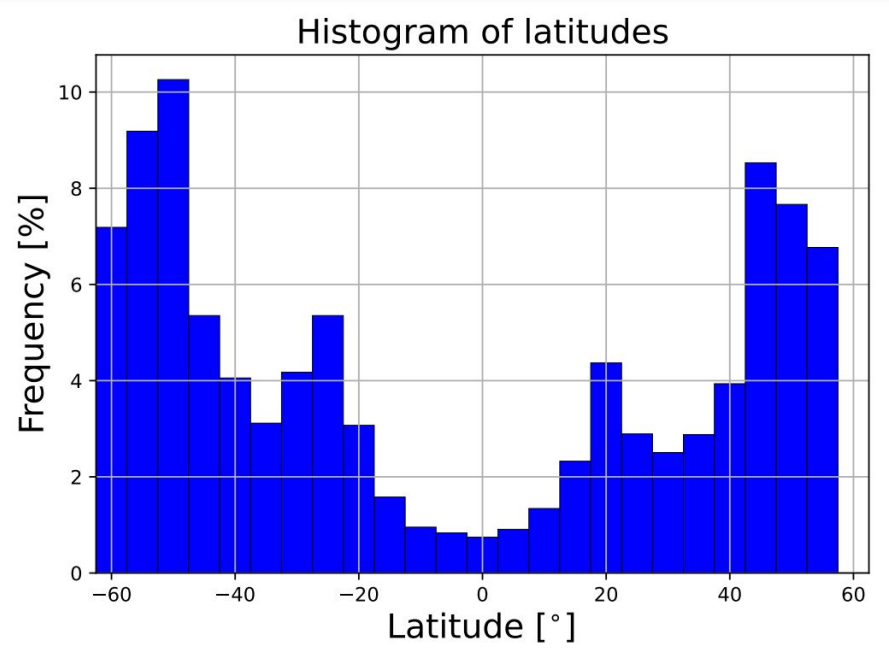
Datasets: sampling and coverage

1. Level 2 Global Precipitation Measurement (GPM) data (collected and prepared by Joe Turk)
    - From the NASA Goddard Space Flight Center
    - LWP and IWP
2. Level 2 Radio Occultation and Heavy Precipitation data from the PAZ satellite (prepared by Kuo-Nung Wang and Ramon Padullés)
    - From the PAZ satellite (ROHP-PAZ)
    - Refractivity and $\Delta\phi$

All variables are given as functions of height at different latitudes, longitudes, and times. These were matched across the two datasets from July 26th, 2018 to August 22nd, 2020 yielding 6706 total coincidences.

Datasets: sampling and coverage





Refractivity model

To identify thermodynamic changes from refractivity profiles, we look for deviations from an ideal refractivity profile, $\widehat{N}$, derived from the following assumptions:

1. $N = \frac{k_1 p}{T} + \frac{k_2 e}{T^2}$
2. Linear temperature profile with height (does not require adiabaticity)
3. Ideal gas law
4. Constant specific humidity (standard assumption for an undersaturated profile)
5. Hydrostatic equilibrium

$$\widehat{N}(z) = \frac{N(z_0)}{\left(1 - c_1 (z - z_0)\right)^2} \times \left\{ (1 - c_2)\left[1 - c_1 (z - z_0)\right]^{c_0} + c_2 \right\}$$

fit coefficients: $c_0 = \frac{g}{R\widehat{\Gamma}} + 1, \quad c_1 = \frac{\widehat{\Gamma}}{\widehat{T}_0}, \quad c_2 = \frac{k_2 \hat{e}_0}{N(z_0)\widehat{T}_0^2}$

Refractivity model

To identify thermodynamic changes from refractivity profiles, we look for deviations from an ideal refractivity profile, $\widehat{N}$, derived from the following assumptions:

1. $N = \frac{k_1 p}{T} + \frac{k_2 e}{T^2}$
2. Linear temperature profile with height (does not require adiabaticity)
3. Ideal gas law
4. Constant specific humidity (standard assumption for an undersaturated profile)
5. Hydrostatic equilibrium

$$\widehat{N}(z) = \frac{N(z_0)}{\left(1 - c_1 (z - z_0)\right)^2} \times \left\{ (1 - c_2)\left[1 - c_1 (z - z_0)\right]^{c_0} + c_2 \right\}$$

fit coefficients: $c_0 = \frac{g}{R\widehat{\Gamma}} + 1, \quad c_1 = \frac{\widehat{\Gamma}}{\widehat{T}_0}, \quad c_2 = \frac{k_2 \hat{e}_0}{N(z_0)\widehat{T}_0^2}$

Refractivity model

We want to study deviations in measured refractivity from $\widehat{N}$ because these could indicate changes in mixing ratio, precipitation, or non-equilibrium physics.

Hence, $\widehat{N}$ is only trained where the assumptions would hold: from z_0=2.5 km to 200 m below the estimated tropopause.

- Tropopause estimated by finding where $\left|\frac{\partial T}{\partial z}\right|$ is minimized for all heights above 5 km and where the temperature is within 10 K of the cold-point tropopause

$$\widehat{N}(z) = \frac{N(z_0)}{(1-c_1(z-z_0))^2} \times \{(1-c_2)[1-c_1(z-z_0)]^{c_0} + c_2\}$$

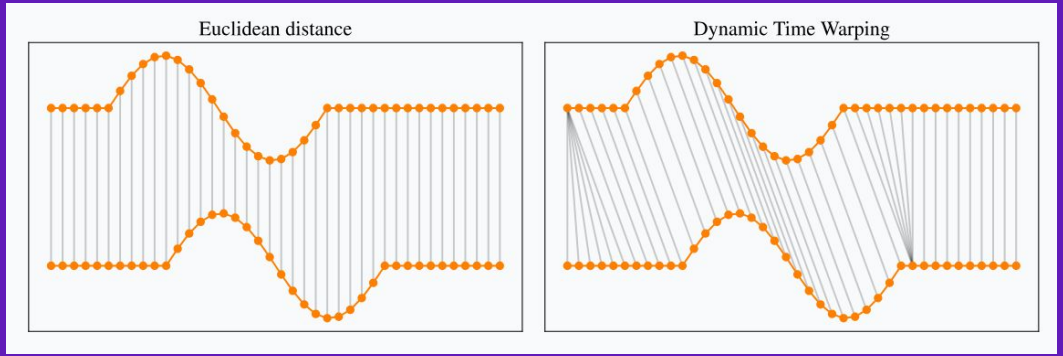
Given all the assumptions, the fit coefficients need not be physical.

k-means clustering

We run k-means clustering with k=8 clusters on the following variables:

- RO measured variables: $\Delta\phi$ (2.5 to 10 km), $N - \widehat{N}$ (2.5 to 8 km), the three fit coefficients (c_0, c_1, and c_2)
- GPM variables and ancillary data: RO+model-derived water vapor pressure (2.5 to 10 km) and GPM+RO ray path computed LWP (1 to 10 km), IWP (1 to 10 km), and total (liquid+ice) water path (1 to 10 km)

We use time series k-means clustering, which is the same as standard k-means except the Euclidean distance is replaced with the dynamic time warping metric.

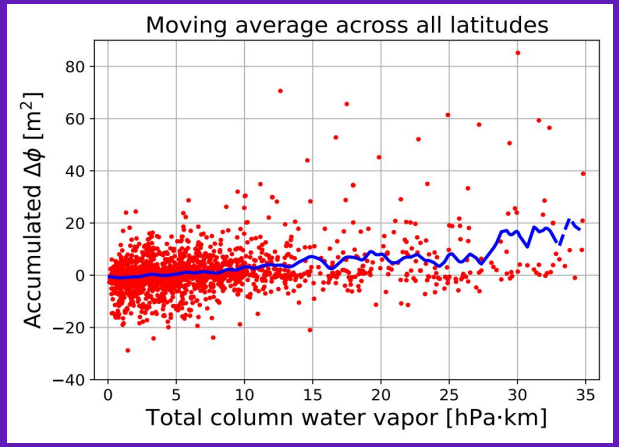


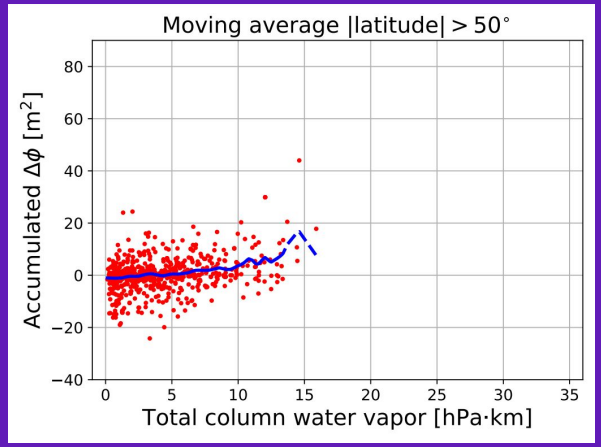
(RO altitude plays the role of time thus enabling vertical shifts in patterns)

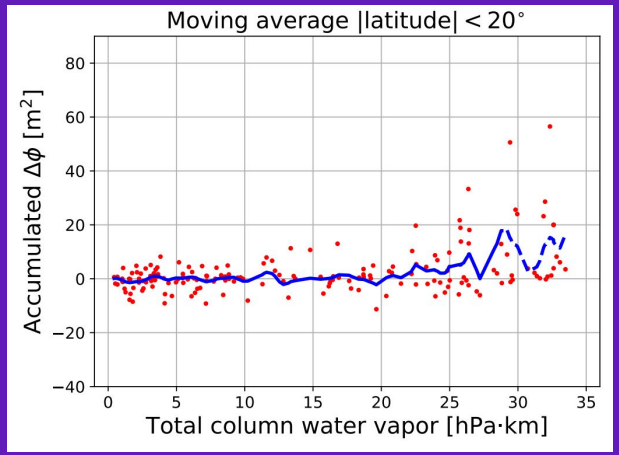
Outline

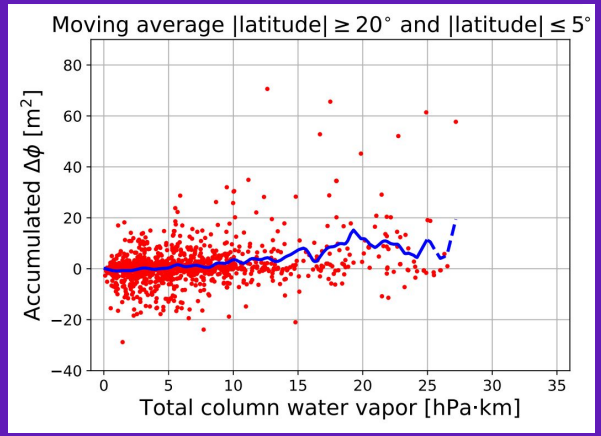
1. Introduction to GNSS and PRO and motivation
2. Data and methods
3. **Results and clustering analyses**
4. Summary

$\Delta\Phi$ vs. total column water vapor









Across all latitudes, mid-latitudes (≥20° and ≤50°), and high latitudes (>50°), the Spearman's correlation coefficients ϱ are low ($0.2<\varrho<0.3$) for the raw data but high ($\varrho>0.9$) for the moving averages.

For low latitudes (<20°), $\varrho=0.287$ for the raw data but $\varrho=0.683$ for the moving averages, largely due to a lack of statistics.

$\Delta\Phi$ vs. total column water vapor

(a) Correlation tests on the raw dataset

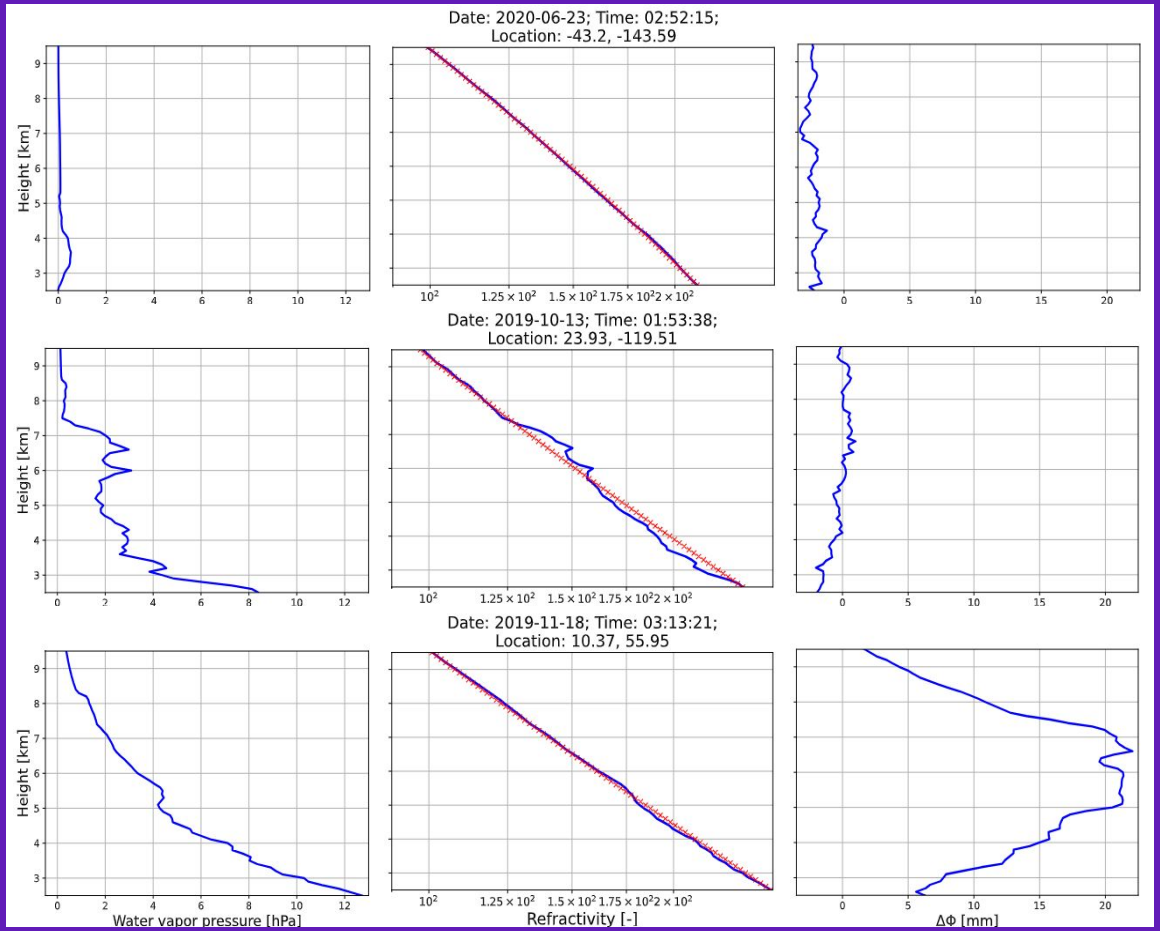
Latitudinal range: $\rightarrow$ Correlation coefficient: $\downarrow$	All	$> 50°$	$\geq 20°$ and $\leq 50°$	$< 20°$
Pearson's r	0.332	0.315	0.349	0.375
Spearman's ρ	0.216	0.223	0.206	0.287
Kendall's τ	0.147	0.151	0.139	0.194

(b) Correlation tests on the moving averages

Latitudinal range: $\rightarrow$ Correlation coefficient: $\downarrow$	All	$> 50°$	$\geq 20°$ and $\leq 50°$	$< 20°$
Pearson's r	0.940	0.901	0.921	0.708
Spearman's ρ	0.971	0.964	0.947	0.683
Kendall's τ	0.864	0.847	0.803	0.508

All p-values below 1e-9!

$N - \hat{N}$ vs. water vapor pressure

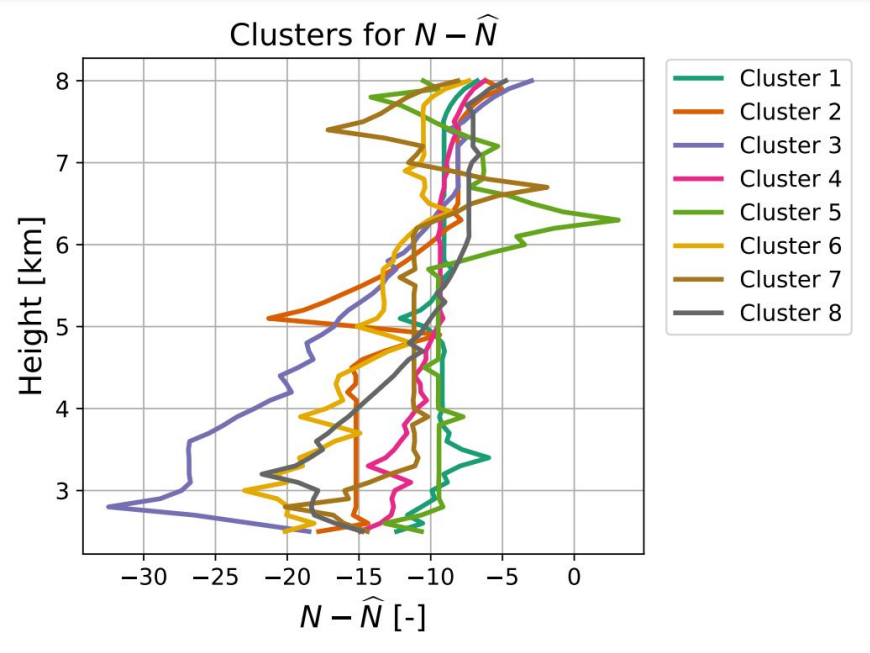


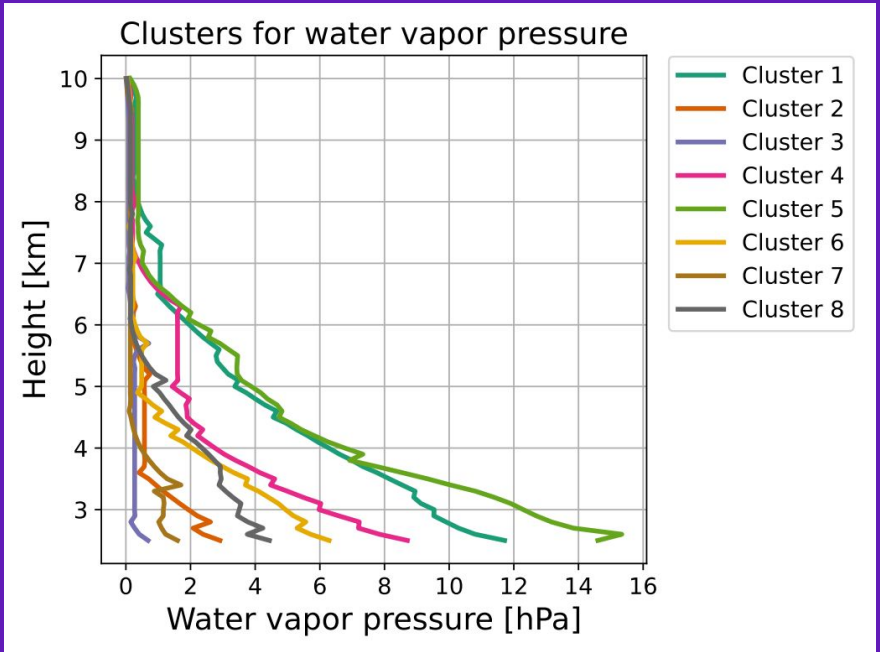
There is a correlation between anomalies in $N - \hat{N}$ and anomalies in e.

From top to bottom:

1. Low moisture, no apparent precipitation
2. Some moisture, no apparent precipitation
3. High moisture, high precipitation

$N - \hat{N}$ vs. water vapor pressure





$N - \hat{N}$ clusters distinguish between different moist thermodynamics (e) but it is complex to describe the precise relationship.

$N - \hat{N}$ clusters do not relate directly to clusters in e.

$N - \widehat{N}$ vs. water vapor pressure

$N - \widehat{N}$: most negative … most positive

closest to flat

e: driest … wettest

$N - \widehat{N}$: →		6.66%	5.34%	7.06%	16.97%	3.75%	38.69%	18.12%	3.41%
e: ↓		3	2	6	8	7	4	1	5
21.35%	3	0.00%	0.29%	0.00%	5.09%	0.00%	30.86%	41.19%	0.90%
20.53%	7	**0.46%**	**1.73%**	**0.66%**	14.45%	2.88%	27.63%	31.66%	23.98%
19.63%	2	8.56%	10.12%	**0.66%**	**35.45%**	8.23%	23.05%	14.81%	23.53%
12.87%	8	14.35%	**24.57%**	**2.18%**	21.91%	12.76%	11.12%	7.15%	**28.51%**
10.89%	6	**30.09%**	**32.66%**	7.42%	15.91%	**29.22%**	4.90%	3.91%	14.93%
7.27%	4	**30.56%**	20.52%	19.43%	6.09%	**24.28%**	1.59%	0.60%	7.69%
6.16%	1	14.35%	8.96%	**55.46%**	0.82%	18.52%	0.36%	0.09%	0.00%
1.29%	5	1.16%	0.87%	**14.19%**	0.09%	3.70%	0.00%	0.00%	0.45%

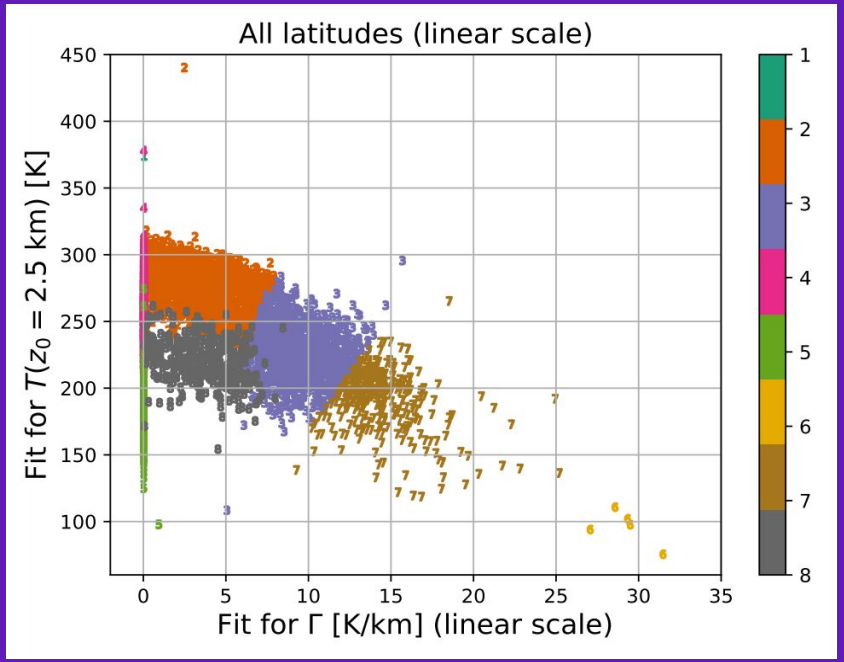
Clusters are ordered from smallest to largest in magnitude for each row and column.

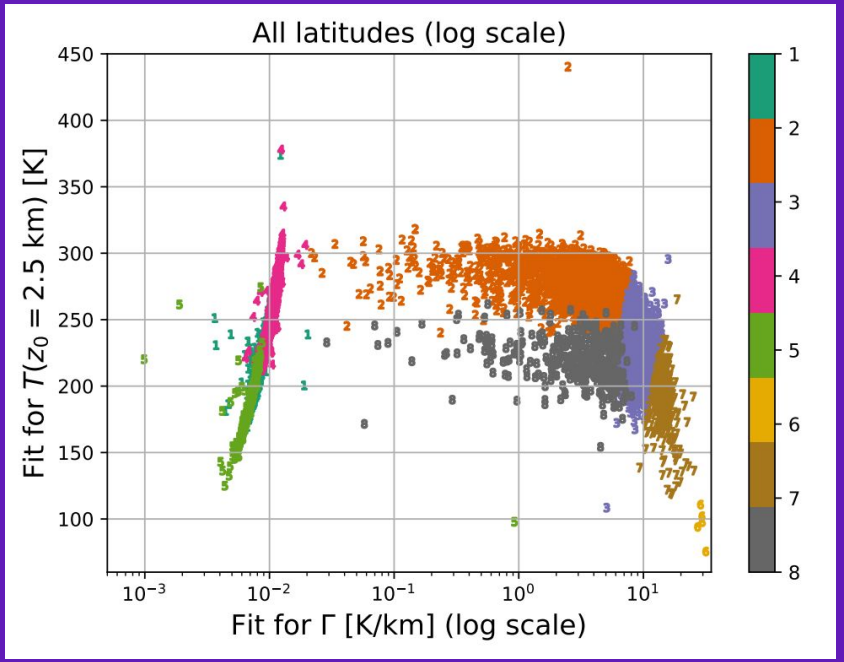
Red: Smallest percent in each row

Green: Largest percent in each row

Bold: Over 1.5 SDs away from the mean for each row

Model coefficients and cluster groups





The colors and numbers correspond to the clusters over (c_0,c_1,c_2).
For $\widehat{\Gamma} > 0.1$, we have a Pearson correlation coefficient of -0.697, a Spearman rank correlation coefficient of -0.676, and a Kendall rank correlation coefficient of -0.497.
(Remember: This is a fit of N to $\widehat{N}$. Fit coefficients need not be physical!)

Model coefficients and cluster groups

most physical

c: →	21.73%	24.13%	16.19%	19.89%	7.60%	0.07%	4.82%	5.57%
e: ↓	1	2	3	4	5	6	7	8
21.35%	3	3.37%	43.54%	34.65%	11.55%	2.75%	40.00%	34.67%	3.75%
20.53%	7	9.34%	29.13%	29.49%	18.98%	8.06%	**0.00%**	**35.91%**	8.85%
19.63%	2	19.78%	16.57%	15.21%	**27.23%**	22.59%	**0.00%**	15.79%	16.09%
12.87%	8	15.73%	6.12%	9.31%	16.05%	**22.40%**	**40.00%**	6.50%	**21.18%**
10.89%	6	17.17%	2.66%	6.54%	11.18%	**27.11%**	20.00%	1.86%	18.50%
7.27%	4	**16.21%**	1.24%	3.13%	6.30%	12.57%	**0.00%**	1.55%	11.26%
6.16%	1	15.11%	0.37%	0.46%	6.98%	3.34%	**0.00%**	0.62%	**18.23%**
1.29%	5	**3.09%**	0.19%	0.65%	1.58%	0.20%	**0.00%**	0.62%	1.88%

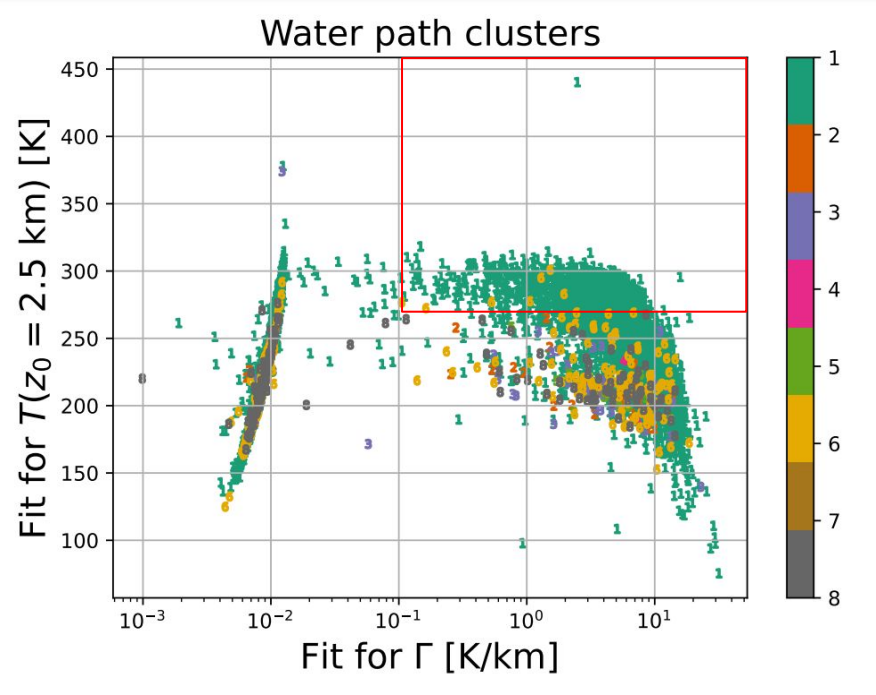
e: driest

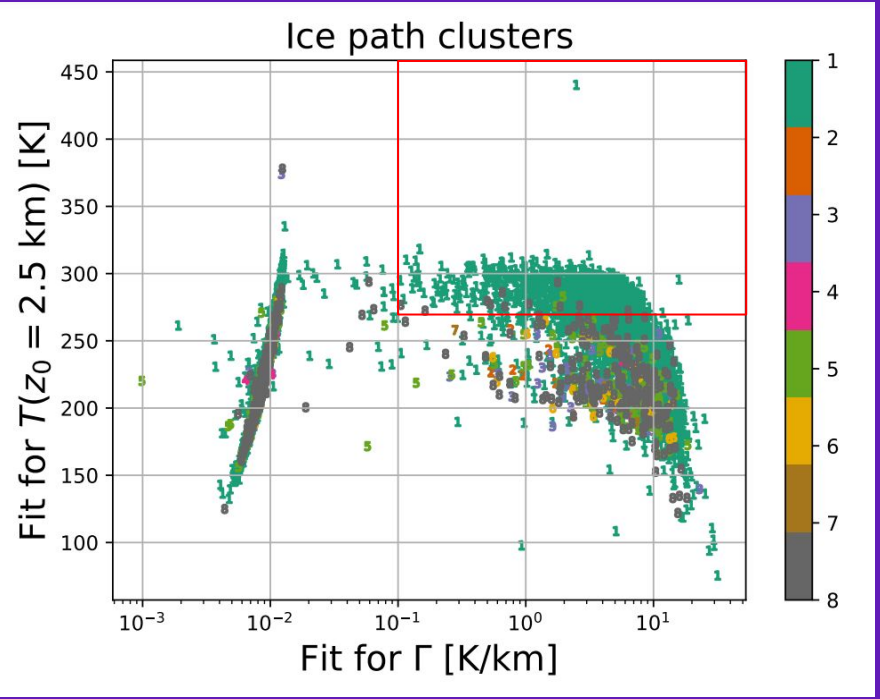
.
.
.

wettest

Red: Smallest percent in each row
Green: Largest percent in each row
Bold: Over 1.5 SDs away from the mean for each row

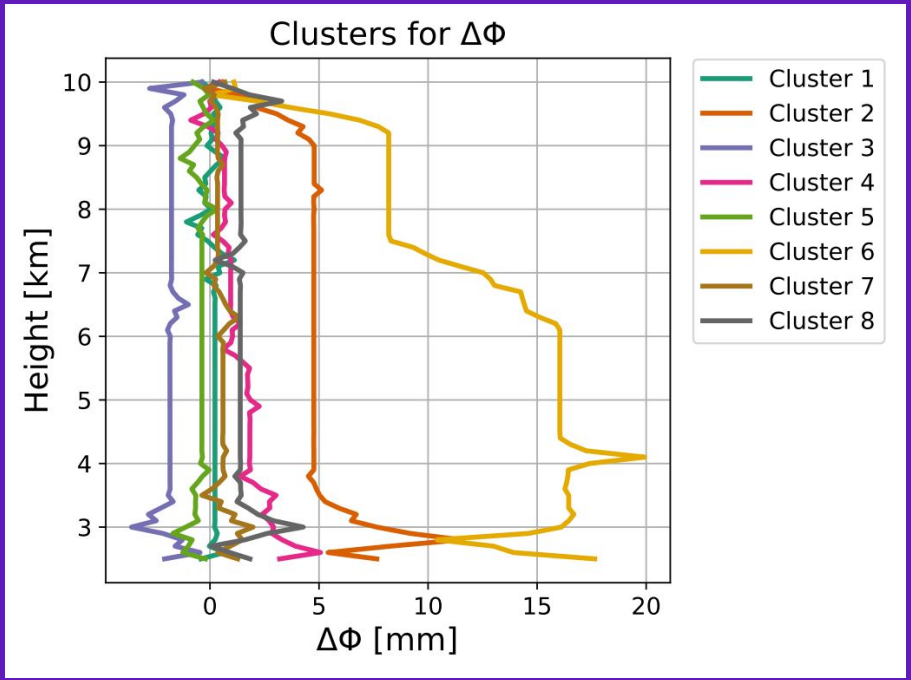
Model coefficients and cluster groups

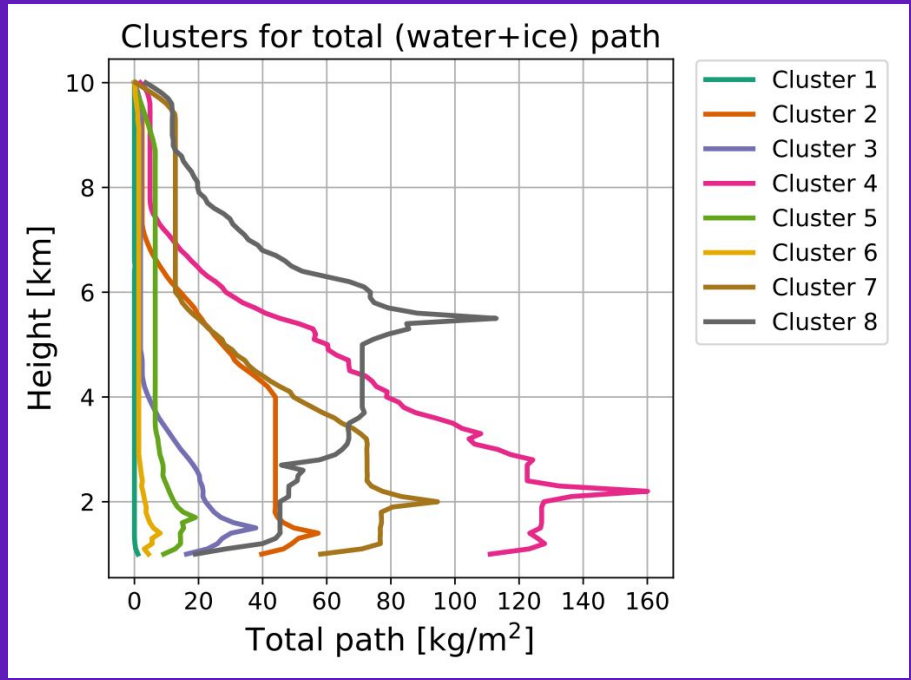




The fit values for T_0 and Γ can often be used to rule out water precipitation and ice.
(Remember: This is a fit of N to $\widehat{N}$. Fit coefficients need not be physical!!)

$\Delta\Phi$ vs. total (liquid & ice) water path





These are the centroids for the $\Delta\phi$ and total (liquid+ice water) path clusters.
They are mostly distinguished based on their overall magnitude.

$\Delta\Phi$ vs. total (liquid & ice) water path

$\Delta\phi$: most negative/zero … most positive

LWP+IWP: driest

.
.
.

wettest

$\Delta\Phi$: $\rightarrow$ / Total path: $\downarrow$		5.20% 3	25.32% 5	34.62% 1	23.45% 7	3.87% 8	5.74% 4	1.56% 2	0.24% 6
84.55%	1	93.02%	92.96%	93.26%	84.14%	60.71%	**28.92%**	**3.33%**	**0.00%**
8.46%	6	4.98%	5.81%	5.09%	12.09%	15.63%	**24.40%**	3.33%	0.00%
3.99%	5	1.99%	1.09%	1.15%	2.88%	17.41%	**29.82%**	12.22%	14.29%
1.79%	3	0.00%	0.07%	0.45%	0.66%	5.36%	**14.46%**	**25.56%**	7.14%
0.82%	2	0.00%	0.00%	0.05%	0.15%	0.45%	2.41%	**35.56%**	**42.86%**
0.22%	7	0.00%	0.00%	0.00%	0.00%	0.45%	0.00%	**12.22%**	**14.29%**
0.09%	8	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	**4.44%**	**7.14%**
0.07%	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**3.33%**	**14.29%**

(similar for LWP and IWP separately)

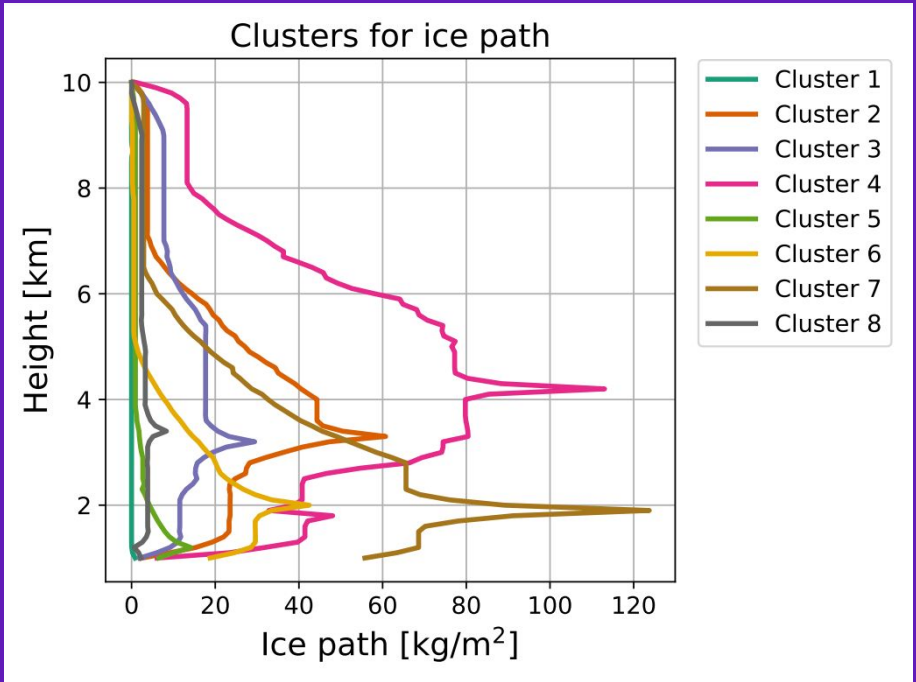
Clusters are ordered from smallest to largest in magnitude for each row and column.

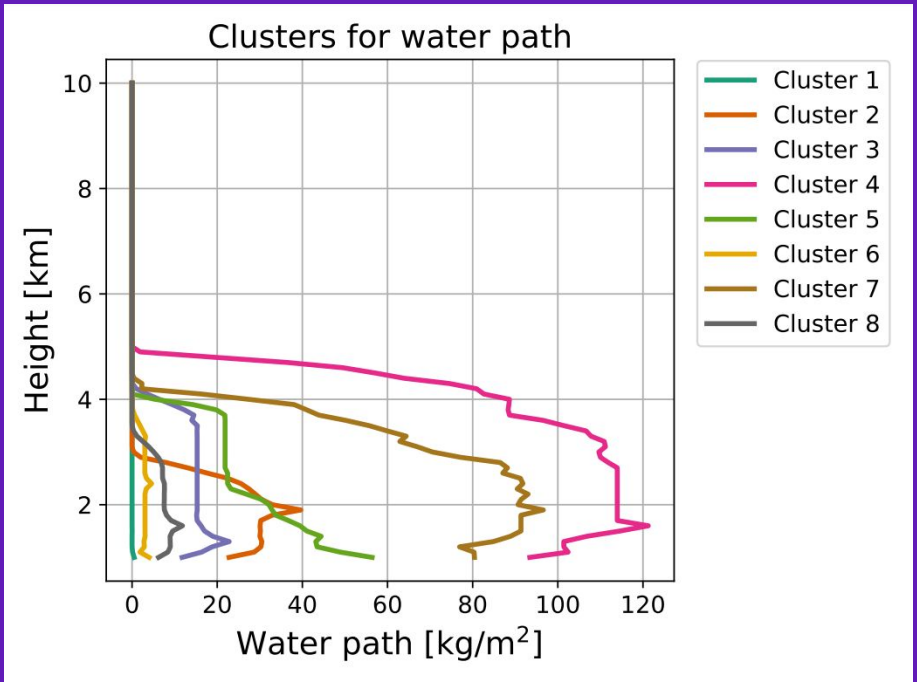
Red: Smallest percent in each row

Green: Largest percent in each row

Bold: Over 1.5 SDs away from the mean for each row

Other clustering analyses





Clustering identifies different vertical distributions, e.g., for LWP and IWP.
For the same vertical distribution, clustering can also separate small and large paths.

Outline

1. Introduction to GNSS and PRO and motivation
2. Data and methods
3. Clustering analyses
4. **Summary**

Summary

What does cluster analysis allow us to do?

- Automate the classification and analysis of physical phenomena found across multiple profiles
- Condense, store, and assimilate thermodynamic information across many profiles
- Utilize PRO data to quickly rule out or confirm physical phenomena without requiring more detailed (and generally more expensive) data retrieval
- Readily identify faulty retrievals and suspicious data without needing to manually check profiles

We still need more sophisticated statistical techniques to *quantify* the uncertainty in using clustering for certain variables to predict other variables, and to *reduce* the uncertainty, we can only go so far with RO-derived variables alone.

Summary

Food for thought:

- How can RO measured variables distinguish between liquid and ice water?
- How can RO measured variables distinguish between different types of clouds?
- How does the precipitation threshold with total column water vapor vary across different latitudes, longitudes, and times of the year?

Summary

Food for thought:

- How can RO measured variables distinguish between liquid and ice water?
- How can RO measured variables distinguish between different types of clouds?
- How does the precipitation threshold with total column water vapor vary across different latitudes, longitudes, and times of the year?

Thank you all for listening! Any questions or comments?