



Cluster analysis of vertical polarimetric radio occultation profiles and corresponding liquid and ice water paths from Global Precipitation Measurement (GPM) microwave data

Jonas E. Katona^{1,2}, Manuel de la Torre Juárez², Terence L. Kubar^{2,3}, F. Joseph Turk², Kuo-Nung Wang², and Ramon Padullés⁴

¹Applied Mathematics Program, Yale University, New Haven, CT, United States of America

²NASA Jet Propulsion Laboratory, California Institute of Technology, La Cañada Flintridge, CA, United States of America

³Joint Institute for Regional Earth Systems Science and Engineering, University of California at Los Angeles, Los Angeles, CA, United States of America

⁴Institut de Ciències de l'Espai, Consejo Superior de Investigaciones Científicas, Barcelona, Spain

Correspondence: Jonas E. Katona (jonas.katona@yale.edu) and Manuel de la Torre Juárez (mtj@jpl.nasa.gov)

Received: 29 April 2024 – Discussion started: 12 August 2024

Revised: 23 November 2024 – Accepted: 27 December 2024 – Published: 26 February 2025

Abstract. Polarimetric radio occultations (PROs) of the Global Navigation Satellite System are able to characterize precipitation structure and intensity. Prior studies have shown the relationship between precipitation and water vapor pressure columns, known as the “precipitation pickup.” Less is known about the relationship between the vertical distributions of temperature and moisture globally within precipitating scenes as measured from space. This work uses cluster analysis of PRO to explore how the vertical distributions of temperature and moisture – combined into PRO refractivity – relate to vertical distributions of precipitation and moisture variables. We evaluate the ability of *k*-means clustering to find relationships among PRO polarimetric phase difference, refractivity, liquid water path (LWP), ice water path (IWP), and water vapor pressure using over 2 years of data matched between the Global Precipitation Measurement (GPM) mission and the radio occultations (ROs) and heavy precipitation (HP) demonstration mission on board the Spanish Paz spacecraft (ROHP-PAZ). A polytropic potential refractivity model for polytropic atmospheres is introduced to ascertain how different vertical thermodynamic profiles that can occur during different precipitation scenarios are related to changes in the polytropic index and thereby vertical heat transfer rates. The cluster analyses suggest a relationship between the amplitude and shape of deviations from the potential refractivity model and water vapor pressure. These analyses also confirm

a positive correlation between vertical shapes of polarimetric phase difference and both LWP and IWP. For certain values, the coefficients of the polytropic potential refractivity model flag physical vs. nonphysical retrievals and indicate when a profile has little to no moisture. The study reveals a similar relationship between the clustering for these coefficients and different water vapor pressure profiles.

1 Introduction

General circulation models need to represent the spatiotemporal structure of precipitation for accurate predictions of climate variability and deep convective structures. Models and observations show a relationship in the probability densities between the precipitation and column water vapor relationship known as the precipitation pickup. Emmenegger et al. (2022) conclude that the majority of the models' convection onset statistics display some degree of temperature dependence in the column water vapor value of the pickup and collapse approximately to a common critical column relative humidity value across saturation mixing ratio bins. However, prior results suggest that the onset of convective instability has a complex dependence on temperature. The vertical structure of temperature and moisture, as well as the entrainment of free-tropospheric air, affects the buoyancy of a rising

convective plume, yielding an onset moisture–temperature dependence slightly different than that of bulk saturation. This work explores how the vertical structure of temperature and moisture, combined into refractivity measured by radio occultation (RO), relates to distributions of precipitation and moisture variables.

Global Navigational Satellite System (GNSS) satellites orbiting Earth periodically send circularly polarized radio signals indicating their positions globally. As these satellites occult from a low-Earth-orbiting satellite with a GNSS receiver, the radio signal they receive has been refracted and bent by the atmosphere. The bending angle is caused by the atmospheric refractivity gradient in the region where the signal traveled. The degree of bending can be calculated using the geometry between the emitting satellite and a receiver, as well as the shift in signal phase between when the signal is emitted and received. GNSS ROs provide refractivity, N , which is related to pressure (p , in hPa), temperature (T , in K), and water vapor pressure (e , in hPa) as follows (e.g., Smith and Weintraub, 1953; Kliore et al., 1974) for an atmospheric air composition with approximately 78 % nitrogen and 21 % oxygen containing water:

$$N = \frac{k_1 p}{T} + \frac{k_2 e}{T^2}, \quad (1)$$

where $k_1 = 77.6$ and $k_2 = 3.73 \times 10^5$ are typically given without dimensions. However, N is expressed in refractivity units, N units; hence, k_1 would be understood in N units \cdot K hPa $^{-1}$, and k_2 in N units \cdot K 2 hPa $^{-1}$.

Quantities derived from RO have demonstrated high accuracy and resolution in space (e.g., Kursinski et al., 1997; Huang et al., 2010; Son et al., 2017). RO temperatures derived from refractivity have been shown to be of similar quantitative accuracy to temperatures directly measured by radiosondes, which are mostly limited to land (e.g., Nishida et al., 2000; Randel et al., 2003; Schmidt et al., 2004; Kim and Son, 2012).

One of the most powerful applications of RO has been in understanding climatic trends – including intraseasonal-to-interannual atmospheric modes of variability such as the quasi-biennial oscillation (QBO), Madden–Julian oscillation (MJO), and El Niño–Southern Oscillation (ENSO) – as they relate to atmospheric structure over the tropics (Scherllin-Pirscher et al., 2021), especially in the upper-troposphere–lower-stratosphere (UTLS) region (Schmidt et al., 2004; Lackner et al., 2011; Johnston et al., 2018, 2022). RO observations have also been used to uncover and measure the upper-level thermal structures of deep convection in tropical storms both alongside and without precipitation radar data (Biondi et al., 2012; Xian and Fu, 2015; Scherllin-Pirscher et al., 2021). In the context of precipitation events, Johnston et al. (2018, 2022) studied the impacts of deep convection and precipitation on the thermodynamic structure of the UTLS region by collocating RO temperature profiles with data from the Global Precipitation Measurement (GPM) mis-

sion and Tropical Rainfall Measuring Mission (TRMM) in both the tropics and midlatitudes.

Equation (1) shows that using RO refractivity data to retrieve thermodynamic variables such as temperature, pressure, and water vapor remains underconstrained. Water vapor information is extracted from refractivity by assuming that the temperature profiles from a given weather analysis – typically either the European Centre for Medium-Range Weather Forecasts (ECMWF) or the National Centers for Environmental Prediction (NCEP) – are accurate at the location of each RO profile, even in cases where the RO and model refractivity may differ (e.g., Kursinski et al., 1997; Kuo et al., 2001). Two common methods for extracting water vapor information from RO refractivity are the 1D-Var method (Wee et al., 2022), which iteratively refines retrievals by combining RO data with background atmospheric model information through a variational data assimilation process, and the direct method (Hajj et al., 2002), which derives retrievals based on hydrostatic equilibrium and an assumed model or background temperature profile. To avoid relying on model water vapor pressure as an assumed background a priori, this study uses the direct retrieval method using temperature profiles provided by NCEP.

An inaccurate refractivity profile from the analysis will lead to erroneous water vapor retrievals. Because RO makes a more valuable contribution to model improvement precisely in the profiles where the weather analysis and RO differ, the relationship between water vapor and refractivity has a higher error bar, particularly in the most useful profiles. Moreover, GNSS RO measurements are sensitive to variations in temperature and water vapor within clouds (Kuo et al., 2001; Huang et al., 2010) but require other observables to confirm the presence of clouds and understand their structure.

Polarimetric radio occultation (PRO) provides a way to expand the applications of standard RO. PRO measures the response of circularly polarized GNSS radio signals to atmospheric anisotropies like precipitating droplets and ice crystals, as these induce a phase difference between the horizontal (H) and vertical (V) components of the GNSS radio signal. The polarimetric phase difference, $\Delta\Phi$, between H and V is related to the number of rain drops or ice crystals in the atmosphere (Tomás et al., 2018; Cardellach et al., 2019; Wang et al., 2022; Padullés et al., 2023) using $\Delta\Phi$ and has promising applications in weather model assimilation (Ruston and Healy, 2021; Wang et al., 2022; Hotta et al., 2024), climate monitoring (Cardellach et al., 2019; Gleisner et al., 2022), and atmospheric research (Turk et al., 2021; Padullés et al., 2023). Datasets from GNSS PRO contain data on refractivity and $\Delta\Phi$, both as functions of height. Unlike infrared instruments, PRO gives data even inside clouds with a higher vertical resolution than microwave (e.g., Turk et al., 2019).

Statistical correlations as a function of height between integrated water content (or water path) from CloudSat – a

NASA satellite mission to survey the vertical structure of clouds and their water content via a radar that launched on 28 April 2006 and ended on 20 December 2023 (NASA, 2024) – along the RO ray path and $\Delta\Phi$ were shown to be strong (Padullés et al., 2023). There are models for how a given thermodynamic state of the atmosphere will affect a propagating RO signal and cause a $\Delta\Phi$ (e.g., Padullés et al., 2023, and references therein). However, a precise formula is missing for how a measured $\Delta\Phi$ relates to thermodynamic atmospheric states. GNSS PRO is generally insensitive to non-dipolar and to spherically symmetric particles, such as aerosols and non-precipitating cloud droplets (Padullés et al., 2023; Hotta et al., 2024). CloudSat-based water content measurements tend to be more sensitive to these smaller particles and cloud tops (Ramon Padullés and F. Joseph Turk, private communication, 11 November 2024). $\Delta\Phi$ at a specific height could result from precipitation of liquid water, ice water, or both – and it may still be influenced by non-precipitating features, such as anisotropic ice crystals (Padullés et al., 2023). It remains an open question whether, or to what extent, differentiating between liquid and ice water precipitation – let alone non-precipitating hydrometeors – is possible using PRO data alone. Therefore, we explore if different vertical distributions of precipitation- or moisture-related variables – $\Delta\Phi$, liquid water path (LWP), ice water path (IWP), and water vapor pressure – are interrelated.

Cluster analysis is a family of methods used to collect and separate populations into different groups or “clusters” within a dataset based on some measure of similarity or hierarchy. One of the most popular clustering techniques is k -means clustering – a flexible, established unsupervised learning method that has been used to classify and analyze the different distributions of physical variables present in climatic and atmospheric datasets (e.g., Jakob and Tselioudis, 2003; Rossow et al., 2005; Yokoi et al., 2011; Wilks, 2019; Govender and Sivakumar, 2020; Nidzgorska-Lencewicz and Czarnecka, 2020).

This study uses cluster analysis to look at how the vertical shape of $\Delta\Phi$ along the RO ray correlates with that of other thermodynamic variables such as refractivity, water vapor pressure, liquid water path (LWP), and ice water path (IWP) along the ray as functions of height at given latitudes and longitudes. A k -means cluster analysis is performed to see if cluster centroids relate to physical phenomena across different variables, the variables being the aforementioned ones and a physically interpretable model for potential refractivity similar to the one introduced in, for example, Bean and Dutton (1966) and de la Torre Juárez et al. (2018). This analysis also looks at how the vertical integral of $\Delta\Phi$ relates to total column water vapor, and how well this confirms results and observations from prior studies. We explore if vertical profiles of $\Delta\Phi$ and refractivity can help to distinguish possible thermodynamic states and even the contributions from ice vs. liquid water precipitation. Through new statistical and

graphical analyses, this study hopes to help understand and quantify these relationships.

To this end, in Sect. 2, we describe the dataset; in Sect. 3, we outline how we classify different thermodynamic states from refractivity profiles alone and provide an overview of how we apply k -means clustering to different variables; in Sect. 4, we use our cluster analysis to search for a classification of disparate vertical structures and cross-correlate interpretations of clusters for different variables; and in Sect. 5, we summarize the aims and results of our study.

2 Data

The two datasets analyzed and used to train the data classification and potential refractivity model are Level 1C, 2A, and 2B Global Precipitation Measurement (GPM) satellite data from the NASA Goddard Space Flight Center and Level 2 radio occultations (ROs) and heavy precipitation (HP) data from the PAZ satellite (ROHP-PAZ) (Cardellach et al., 2019). From the former, we obtain the liquid water path (LWP, kg m^{-2}) and ice water path (IWP, kg m^{-2}) using emissivity principal component (EPC) profiling retrievals at each pixel across the scan of the GPM passive microwave (PMW) satellite radiometer. The EPC data have a spatial resolution of $0.1^\circ \times 0.1^\circ$, temporal resolution of 30 min, and 0.25 km height levels, as described in Appendix A of Turk et al. (2018) and in Table 1 and Sect. 3 of Turk et al. (2021). Meanwhile, ROHP-PAZ gives refractivity (N units) and $\Delta\Phi$ (mm), all as functions of height at different latitudes, longitudes, and times. Using refractivity and assuming temperature from NCEP, the direct method (Haji et al., 2002) is applied to derive the pressure (hPa) and water vapor pressure (hPa), by means of which the temperature (K) used in this study is derived from Eq. (1). For further details on how the aforementioned variables are retrieved from the datasets, we refer the reader to Turk et al. (2021) and the references therein for the GPM dataset and Cardellach et al. (2019) for the ROHP-PAZ dataset.

The direct method relies on the ancillary model refractivity agreeing with that of the RO and with its having the correct temperature distribution for that refractivity profile. While an agreement between the ancillary model and observation should yield reliable retrieved values, using the model temperature profile may introduce retrieval errors when the ancillary model and RO profile refractivities differ significantly (due to, e.g., collocation errors, RO bias, or limited model resolution). The direct method results in negative or otherwise unrealistic water vapor pressure values serving as quality-control flags.

The GPM and ROHP-PAZ profiles are matched across different latitudes, longitudes, and times whenever they coincided within a given spatiotemporal range. As described in Sect. 2 of Turk et al. (2021), the collocation criteria were that the GPM PMW satellite overpass had to occur within

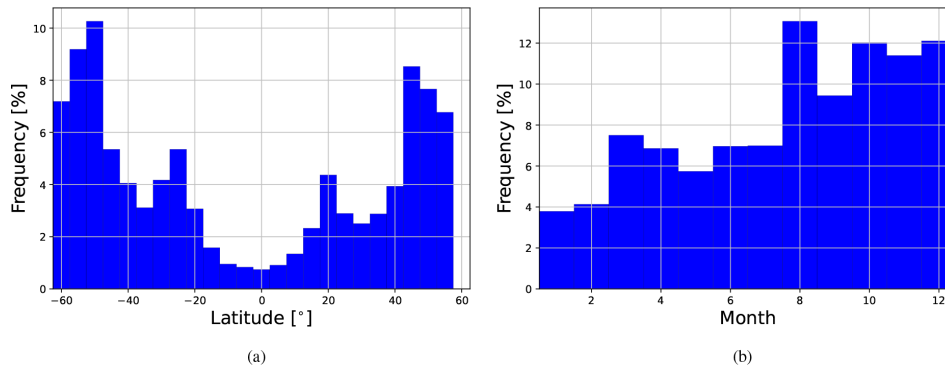


Figure 1. Sampling distributions for the collocations between the GPM and ROHP-PAZ datasets at different (a) latitudes and (b) months.

± 15 min of the ROHP-PAZ observation, and the ROHP-PAZ observation location had to fall within the PMW satellite's swath. For each ROHP-PAZ observation, the tangent point – the point closest to the Earth's surface along the ray path – was selected from the lowest-level RO. This gives

- 2362 coincidences from 26 July to 31 December 2018 (inclusive),
- 2943 coincidences from 1 March to 31 December 2019,¹ and
- 1401 coincidences from 1 January to 22 August 2020,

thereby yielding a total of 6706 coincidences from 26 July 2018 to 22 August 2020. At the latitude and longitude of each coincidence, the collocated data are interpolated onto a grid with equally spaced height intervals of 0.1 km.

Most of the aforementioned coincidences lie poleward of 40° N or S, as shown in Fig. 1a, enabling good statistics in those regions. There was a low number of coincidences in the tropics (within 15° of the Equator), which constrains our analysis in low-latitude regions. Furthermore, as Fig. 1b shows, there is also a slightly higher number of coincidences in the last 4 months of the year vs. the first 8 months, but this poses less of a problem as we do not assess seasonality.

Turk et al. (2021) computed LWP and IWP by integrating the condensed water content (kg m^{-3}) – estimated via EPC passive microwave precipitation profiling (Turk et al., 2018; Utsumi et al., 2020) – along each RO ray path in the ROHP dataset coinciding with GPM data. Integrating the condensed water content along the ray paths ensures that their values are related to $\Delta\Phi$, which is also computed by integrating along each ray path. As a first approximation, we partition the integrated water content into LWP and IWP based on whether the

retrieved or model temperature is above or below 273 K, respectively (Turk et al., 2021). Since non-precipitating supercooled water is not expected to be asymmetric, it should induce little to no polarimetric phase difference (Padullés et al., 2023; Hotta et al., 2024). Hence, this approach misclassifies some supercooled water as ice, in which case this misclassification would predict $\Delta\Phi$ associated with ice when no $\Delta\Phi$ is measured. Finally, as with $\Delta\Phi$, the values of the LWP and IWP at a given latitude, longitude, and height are given according to where the lowest-level tangent point for the given ray path lies.

To compute the total column water and ice paths from the aforementioned data for each profile, the water and ice paths are integrated, respectively, from 1 to 10 km only if a profile has data at 1 and 10 km.² For computing the vertical integral of $\Delta\Phi$, since the error associated with this variable in the ROHP-PAZ dataset is roughly ± 2 mm at each height, $\Delta\Phi$ is integrated from 2.5 to 10 km after rounding $\Delta\Phi$ to the nearest multiple of 2 mm if data exist from 2 to 2.5 km and at 10 km. The latter condition ensures that the endpoints of the integral are correct, and we exclude faulty retrievals which tend to deteriorate near the bottom of the profiles before the data become corrupted or missing.

Finally, for computing the total column water vapor, the water vapor pressure is integrated from 2.5 to 10 km, excluding profiles that feature no data at 2.5 or 10 km, negative water vapor pressure values, or unrealistically high water vapor pressure values – these situations are unphysical and likely result from ancillary model or retrieval errors. The profiles with unrealistically high water vapor pressure values were identified by running an initial k -means cluster analysis with $k = 8$ – as explained later in Sect. 3.2 and 3.3 – but on every water vapor pressure profile in the dataset. For the dataset,

¹Technical issues with the processing of ROHP-PAZ retrievals were encountered in January and February 2019. Although these issues have since been resolved in the currently available ROHP-PAZ dataset, the collocated dataset described in Turk et al. (2021) and analyzed in this study was created before then.

²While requiring path data down to 1 km may seem too stringent, requiring this only ends up excluding six profiles at most for both water and ice paths or under 0.1 % of all the profiles in the dataset. Thus, it is not too stringent unless these six profiles happen to be rather extreme cases. The analysis at hand aims at finding general trends and associations rather than atypical cases.

only one cluster contained profiles with unrealistically large water vapor pressure values – at least above 250 hPa at some height in all profiles – while the other clusters contained profiles with water vapor pressure values below 150 hPa. Hence, all profiles in the anomalous cluster were excluded from analyses that relied on water vapor pressure, particularly in the final water vapor pressure cluster analysis.

The number of profiles where these conditions were not met are as follows:

- for total column water vapor, 33 profiles (0.49 % of all profiles in the dataset);
- for total column water path, 1 profile (0.01 %);
- for total column ice path, 6 profile (0.09 %);
- for total column water plus ice path, 6 profiles (0.09 %); and
- for the vertical integral of $\Delta\Phi$, 923 profiles (13.76 %).

For all cases, the integration is implemented in Python using the composite trapezoidal rule (Atkinson, 1988).

3 Methods

The PRO observables are $\Delta\Phi$ and refractivity as functions of height, latitude, and longitude. Hence, this study explores how far one can get with PRO observables while remaining as independent from externally derived weather analyses as possible. To this end, we develop a model for potential refractivity as a function of height assuming a constant lapse rate (which can be non-adiabatic), hydrostatic balance, and a constant water vapor mixing ratio.

3.1 Potential refractivity in a polytropic atmosphere

A first classification criterion organizes profiles based on the differences between observed refractivity profiles and those expected for polytropic atmospheres where air can expand and compress with adiabatic and non-adiabatic heat transfer. If an air parcel moving vertically through the atmosphere follows a polytropic process – a *polytropic* atmosphere – and the ideal gas law holds, then p/ρ^m and therefore $p^{1-m}T^m$ are constant, where m is the polytropic index of the atmosphere.

We define $K \doteq p^{1-m}(z)T^m(z) = p(z_0)^{1-m}T(z_0)^m$ for some reference height z_0 . In hydrostatic balance, we have $\frac{\partial p}{\partial z} = -\rho g$, and polytropy also implies that $\frac{\partial p}{\partial z} = \frac{\partial(K\rho^m)}{\partial z} = mK\rho^{m-1}\frac{\partial\rho}{\partial z}$. Balancing these two equations necessitates that $-\rho g = mK\rho^{m-1}\frac{\partial\rho}{\partial z}$, and after multiplying both sides by $\frac{m-1}{mK\rho}$, one gets

$$g \frac{1-m}{mK} = \frac{\partial\rho^{m-2}}{\partial z}.$$

At constant $m \neq 0, 1$, the solution is

$$\rho^{m-1}(z) = \rho(z_0) + g \frac{1-m}{m}(z-z_0). \tag{2}$$

At $m = 0$, the pressure is constant and cannot satisfy hydrostatic equilibrium unless $\rho = 0$, while at $m = 1$, the density decays exponentially, typical of an isothermal atmosphere. When $m = \gamma$, where $\gamma = 5/3$ is the adiabatic index, the change in temperature incurred by air parcels moving vertically in this atmosphere follows an adiabatic process – an *adiabatic* atmosphere.

Using Eq. (2) for the vertical profile of an ideal gas, where $\frac{p(z)}{\rho(z)} = RT(z)$, and by polytropy again, $p(z) = K\rho^m(z) = K\rho(z)\rho^{m-1}(z)$ implies that

$$\begin{aligned} RT(z) &= \frac{p(z)}{\rho(z)} = K\rho^{m-1}(z) \\ &= K \left[\rho(z_0) + g \frac{1-m}{m}(z-z_0) \right]. \end{aligned}$$

This shows that an ideal gas atmosphere in hydrostatic equilibrium and with constant polytropic index $m \neq 0$ with height has a linear temperature profile $T(z) = \hat{T}(z_0) - \hat{\Gamma}(z-z_0)$, where $\hat{T}(z_0) = \frac{K\rho(z_0)}{R}$ and $\hat{\Gamma} = -Kg \frac{m-1}{mR}$. When $m = 1$, the solution holds with $\hat{\Gamma} = 0$ and a constant temperature with height. At constant m and R , $\hat{\Gamma} = -\frac{\partial T}{\partial z}$, and hence, the lapse rate is constant.

When including water vapor processes, one can characterize the temperature profiles in a polytropic atmosphere as (1) a completely dry atmosphere or (2) an *unsaturated* moist atmosphere. Additionally, one can approximate temperature via a linear relationship with height for (3) a *saturated* moist atmospheric layer where the expansion and contraction of air are reversible or (4) an atmosphere in which water that condenses in an air parcel is instantaneously removed via precipitation – a *pseudoadiabatic* atmosphere (e.g., Emanuel, 1994). The lapse rate, Γ , is nearly constant and called a dry adiabatic lapse rate in the first case, a moist–unsaturated adiabatic lapse rate in the second, a reversible moist adiabatic lapse rate in the third, and a pseudoadiabatic lapse rate in the fourth. The temperature profile is precisely linear with height for only the first case and close to linear for the others. Each of the four thermodynamic cases above would be represented by a different conservation law (Emanuel, 1994) – dry adiabatic (for 1), moist adiabatic (for 2 and 3), and pseudoequivalent potential temperatures (for 4) – and, by analogy, via a different type of potential refractivity profile. These conserved quantities can be used to define different types of potential refractivity, \hat{N} , based on fitting data to physical laws describing adiabatic and pseudoadiabatic processes (e.g., de la Torre Juárez et al., 2018).

\hat{N} is derived here for an atmosphere with the following properties: (1) Eq. (1), (2) the ideal gas law, (3) a linear temperature profile with height representative of a polytropic atmosphere, (4) a constant specific humidity representative of a

subsaturated atmosphere, and (5) in hydrostatic equilibrium. Deviations between the measured refractivity N and the fit to the model \widehat{N} signal the presence of changes in mixing ratio, precipitation, or non-equilibrium physics (e.g., gravity waves or turbulence). From the above assumptions, we derive in Appendix A the model for \widehat{N} :

$$\widehat{N}(z) = \frac{N(z_0)}{[1 - c_1(z - z_0)]^2} \times \{(1 - c_2)[1 - c_1(z - z_0)]^{c_0} + c_2\}, \quad (3)$$

where $c_0 = \frac{g}{R\widehat{\Gamma}} + 1 = \frac{2m-1}{K(m-1)}$, $c_1 = \frac{\widehat{\Gamma}}{T_0} = \frac{g(m-1)}{m\rho(z_0)}$, and $c_2 = \frac{k_2\widehat{e}_0}{N(z_0)T_0^2}$ are coefficients which must be fit to a given refractivity profile and provide information about the polytropic index. $\widehat{T}_0 \doteq T(z_0)$ and $\widehat{e}_0 \doteq e(z_0)$ are the temperature and water vapor pressure, respectively, at reference height z_0 . In particular, for $m = 1$, \widehat{N} has an exponential relationship with z (e.g., Bean and Dutton, 1966). The fit coefficients $\mathbf{c} = (c_0, c_1, c_2)$ are defined in terms of the following physical parameters: the acceleration due to gravity on Earth $g = 9.81 \text{ m s}^{-2}$, specific gas constant of dry air $R = 287.05 \text{ J kg}^{-1} \text{ K}^{-1}$, mean tropospheric lapse rate $\widehat{\Gamma}$ (in K km^{-1}), and constants k_1 and k_2 defined in Sect. 1.

The lapse rate can change with height across moist and dry sections, e.g., in the transition between the boundary layer and the free atmosphere (von Engel et al., 2005; Ao et al., 2012), in the transition from the mid-troposphere to the tropical tropopause layer (TTL) (Fueglistaler et al., 2009), or when clouds are present in a real atmosphere (e.g., Peng et al., 2006; Mascio et al., 2021). Based on these three examples, the fits for \widehat{N} are made in an altitude range that is likely to have a constant lapse rate under the five assumed properties from the last subsection. We set the lower limit of the fit at $z_0 = 2.5 \text{ km}$, which is mostly above the boundary layer. Furthermore, we set the upper limit to ensure that the fit remains below changes in sign in the lapse rate, such as those caused by gravity waves, stratospheric intrusions, and thermal inversions. Specifically, the upper limit to the fit is 200 m below the lowest height above 5 km where $\left| \frac{\partial T}{\partial z} \right|$ is minimized – which defines a local minimum or a maximum in temperature, as expected for either the cold point tropopause, large gravity waves, or the bottom of the TTL (Fueglistaler et al., 2009) – and where the temperature is within 10 K of the minimum temperature below 25 km, i.e., within 10 K of the temperature at the cold-point tropopause. Second-order central differences are used to estimate $\frac{\partial T}{\partial z}$ across all of the heights for each given profile (Atkinson, 1988). For this purpose, we use the RO-derived temperature to estimate T to establish a rigorous criterion across all of the profiles and ensure that the fit \widehat{N} is consistently being used where it would be expected to hold, especially for accurate clustering in $N - \widehat{N}$.

More details on the numerical fitting of Eq. (3) are given in Appendix B.

3.2 Time series k-means clustering

Across the profiles in the merged dataset described above, we apply k -means clustering with $k = 8$ clusters for each of the following variables:

- RO-measured variables – $\Delta\Phi$ (2.5 to 10 km), $N - \widehat{N}$ (2.5 to 8 km), and the three fit coefficients for \widehat{N} (the vector \mathbf{c});
- variables from ancillary data – RO plus model-derived water vapor pressure (e , 2.5 to 10 km) and GPM plus RO ray-path-computed liquid water path (LWP, 1 to 10 km), ice water path (IWP, 1 to 10 km), and total (liquid plus ice) water path (TWP, 1 to 10 km).

In all cases aside from the clustering for coefficients (in which case we use standard k -means clustering with the standard Euclidean distance), a variation in naive k -means clustering called time series k -means with dynamic time warping (DTW) (Izakian et al., 2015) is applied. As with naive k -means, the dataset is partitioned into k clusters, but instead of measuring the distances between profiles using the Euclidean distance, DTW is used. The numerical procedure for running k -means clustering on the aforementioned variables is described in Appendix C.

We introduce quality-control criteria for each of the variables informed by how k -means clustering detected outliers and other physical considerations. For instance, for the $\Delta\Phi$ clustering, we excluded $\Delta\Phi$ profiles where the retrieval for $\Delta\Phi$ cut off above 2 km, and for the e clustering, we excluded e profiles with unphysically high values, identified using the same clustering technique described in Sect. 2 for excluding e profiles used in the total column water vapor calculations. Including these faulty profiles affects the accuracy when we compare the shapes of the $\Delta\Phi$ profiles for clustering and compute the integral of $\Delta\Phi$. We found that faulty retrievals tend to deteriorate near the bottom of the profiles before the data become corrupted or missing. The percent of profiles excluded ranged from 0.01 % (for LWP clustering) to 13.76 % (for $\Delta\Phi$ clustering). See Appendix D for more details on the precise quality-control criteria used for each clustering variable.

3.3 Dynamic time warping

DTW is a technique originating in time series analysis that measures the similarity between two signals which are functions of time (or some analogous variable – in this case, height) by finding an optimal alignment between the two signals by “warping” the sample points of each signal such that the measurements in each signal are matched to their nearest point(s) in the other signal as measured by the Euclidean norm, regardless of the times at which each point was measured (Müller, 2007). We still assume that the start and end points match in each case, that the ordering of measurements (with time) within each profile stay the same, and that each

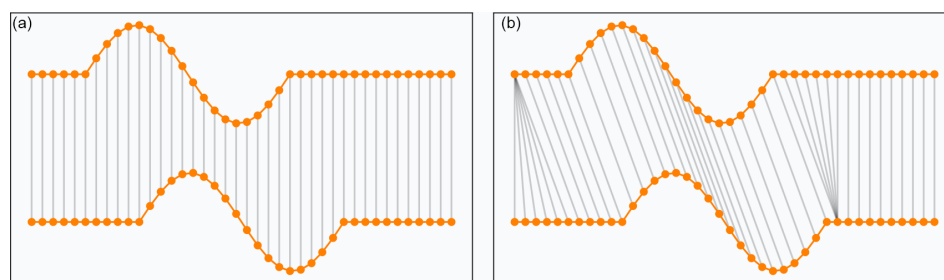


Figure 2. A visual comparison (Tavenard, 2021) showing the difference between Euclidean distance (a) and the DTW measure (b). Time series are shifted vertically in the visualization, but assume that the y axis values match. Thus, between the two time series shown, the Euclidean distance would be nonzero but the DTW measure would be zero.

point in one signal is matched to at least one point in the other. This ensures the following:

1. For cases of missing or uneven data points within a given profile, we can still compare the rough shape of this profile with others.
2. For translations in sampling (e.g., when two measurements are out of phase or when recorded heights are imprecise), DTW can make up for this by shifting the heights at which measurements are taken when comparing two profiles.

See Tavenard (2021) or Müller (2007) for more details on how DTW is calculated.

Figure 2 features an intuitive visualization of how DTW works when comparing time series. The featured example is taken from Tavenard (2021) and shows two signals consisting of horizontal lines combined with one period of a sinusoid. Note how DTW matches the patterns and overall shape of each time series, which intuitively should result in a more sound similarity assessment than when using the Euclidean distance, since the latter matches timestamps (or heights for this study) regardless of when they were sampled.

4 Results and analysis

Clustering provides an initial classification for the types of atmospheric profiles that can occur across the dataset by looking at the centroids in different clustering variables. Figure 3 shows the results of *k*-means cluster analysis when applied to the following variables in the dataset: (a) $N - \hat{N}$, (b) $\Delta\Phi$, (c) IWP, (d) LWP, (e) TWP, and (f) water vapor pressure. The plots in Fig. 3 show the eight clustering centroids for each variable. Each centroid is an average profile representing the general shape and magnitude of the indicated variable for profiles within its cluster.

A second step in the analysis uses frequency histograms of different cluster groups to summarize the relationships between clusters. Tables 1, 2, and 3 feature frequency histograms that compare clustering in different variables: $N - \hat{N}$

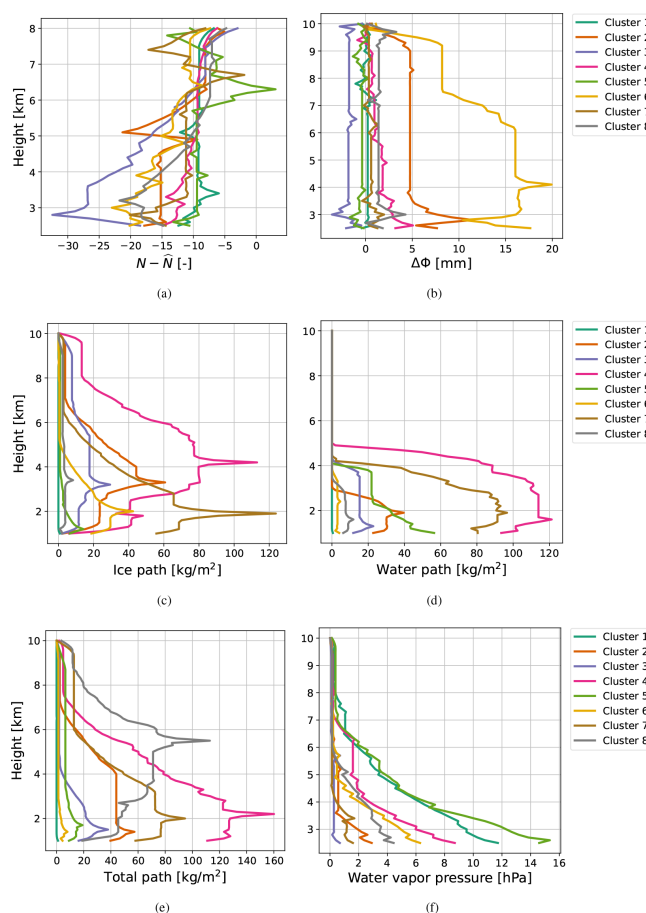


Figure 3. Cluster analysis centroids computed by applying time series *k*-means clustering across profiles in (a) $N - \hat{N}$, (b) $\Delta\Phi$, (c) IWP, (d) LWP, (e) TWP, and (f) water vapor pressure.

with water vapor pressure, the \hat{N} coefficients c with water vapor pressure, and $\Delta\Phi$ against the path variables (LWP, IWP, and TWP), respectively. These tables look for patterns in the ability of $N - \hat{N}$ to predict different distributions of vertical water vapor pressure and $\Delta\Phi$ to predict different types of water path profiles across the vertical profiles in the dataset. Percentages in the topmost row and leftmost column reflect

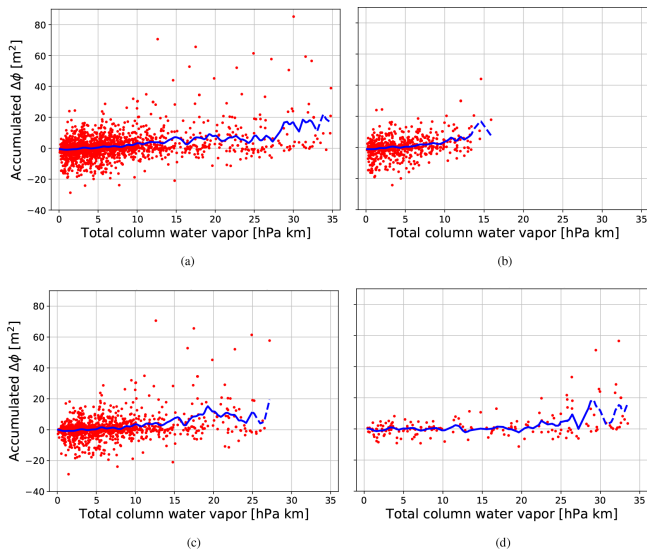


Figure 4. Moving averages (blue) of accumulated $\Delta\Phi$ vs. accumulated water vapor pressure over scatter plots (red) across (a) all latitudes, (b) $|\text{latitude}| > 50^\circ$, (c) $|\text{latitude}| \geq 20^\circ$ and $|\text{latitude}| \leq 50^\circ$, and (d) $|\text{latitude}| < 20^\circ$. Dashed portions of the moving averages correspond to where each bin had less than 34 data points.

the total number of profiles that meet the clustering requirements for each specified clustering variable. Bolded black and bolded grey indicate the maximum and minimum percentages within each respective row. Note that since profiles were excluded from the cluster analyses for certain variables, the weighted averages for each column or row will not always add up as expected from the law of total probability.

4.1 Total column $\Delta\Phi$ and total column water vapor

Bretherton et al. (2004) showed an exponentially increasing relationship between precipitation and total column relative humidity over the tropics. Later studies (Muller et al., 2009; Holloway and Neelin, 2010; Emmenegger et al., 2022) demonstrate a similar and related positive relationship between precipitation and total column water vapor (TCWV) in the tropics, where under a certain TCWV value, precipitation is generally near-zero in a given profile, and above a “pickup” threshold in TCWV, precipitation *may* become non-negligible and increase exponentially. To evaluate the statistical representativity of our dataset, we tested the validity of using the magnitude of $\Delta\Phi$ as a proxy for the magnitude of precipitation by looking for a monotonic relationship – and in particular, the precipitation pickup pattern (Holloway and Neelin, 2010) – between TCWV and the total column of the PRO observable $\Delta\Phi$.

Figure 4 presents scatter plots of accumulated $\Delta\Phi$ vs. TCWV for all profiles in the dataset at (a) all latitudes, (b) upper midlatitudes (above 50°), (c) subtropics and midlatitudes (between 20° and 50°), and (d) tropics (below 20°), with over-

laid moving averages. These moving averages were done using the `filter1d` tool Generic Mapping Tools (`gmt`) Version 6.3 (Wessel et al., 2019). Averaging was done with a Gaussian filter of width 2 hPa km (option `-Fg2`) excluding outputs where the input data have a gap exceeding 0.2 (option `-L0.2`) and including ends of the time series in the output (option `-E`).

The sparse statistics and high variability across higher-moisture profiles within the dataset make it difficult to filter out outlier profiles that could significantly bias the moving averages. Thus, Fig. 4 shows insufficient data for higher-moisture cases to replicate the precipitation pickup pattern with much fidelity, as shown by the dashed lines in Fig. 4.³ Nonetheless, after averaging, a positive correlation between accumulated $\Delta\Phi$ and TCWV was found across *all* latitudes ($r_p = 0.940$) and for the three latitudinal ranges separately (see Table 4).

The strength of the correlation between accumulated $\Delta\Phi$ and TCWV also depends on which data the correlation analyses ran. The correlation coefficients in Table 4 indicate a low positive correlation between accumulated $\Delta\Phi$ and TCWV in the raw dataset – i.e., in the individual profiles. After applying the Gaussian filter with results in Fig. 4 and running correlation analyses on the filtered data, we find a high positive correlation between the same two quantities in Table 4. This suggests that, on average, there is a global positive relationship between the total column $\Delta\Phi$ and water vapor pressure, but this relationship is weak across individual profiles. Hence, when classifying individual profiles, accumulated $\Delta\Phi$ does not appear to be a good proxy for precipitation on a single profile; Sect. 4.2 gives a more useful way to predict water vapor pressure profiles using RO observables.

Similarly, Fig. 4b and d show that, even when using running means, the limited data and high variability across individual profiles only weakly suggest a threshold at which TCWV begins to induce precipitation – i.e., the critical level at which the precipitation pickup starts. This threshold appears to be notably lower in the upper midlatitudes than in the tropics: the accumulated $\Delta\Phi$ moving averages reach similar magnitudes at approximately 12–13 hPa km in high latitudes vs. 25–26 hPa km in the tropics. However, particularly for tropical profiles, significantly more data are needed to robustly confirm how accurately polarimetry can capture the precipitation pickup pattern from accumulated $\Delta\Phi$ averages.

4.2 $N - \hat{N}$ and water vapor pressure

We represent the deviations of N from a profile with the properties outlined in Sect. 3.1 by looking at overlaid graphs of N and \hat{N} as functions of height and by plotting $N - \hat{N}$ as a function of height. Figure 5 shows two examples – Fig. 5a and b – where $N - \hat{N}$ does not correlate strongly with $\Delta\Phi$,

³A total of 34 counts per bin was chosen as a consistent threshold for all plots in Fig. 4 to show where the density of data falls below a given reference value.

Table 1. Percent of profiles in each e cluster (column) for each $N - \hat{N}$ cluster (row). Cluster numbers are ordered from smallest (most negative/zero) to largest (most positive) values by comparing their corresponding centroids in Fig. 3. Overbars and underbars indicate percentages that fall above or below, respectively, 1.5 times the weighted standard deviation (SD) from the mean percentage for each given row; the SD is weighted by the percentage of $N - \hat{N}$ corresponding to each case (6.66 %, 5.34 %, 7.06 %, etc.). Bolded black and bolded grey indicate the maximum and minimum percentages for each row, respectively.

		$N - \hat{N}$: most negative			...	most positive				
e : driest ... wettest	$N - \hat{N}$: →	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	<u>0.46%</u>	<u>1.73%</u>	<u>0.66%</u>	14.45%	2.88%	27.63%	31.66%	23.98%
	19.63%	2	8.56%	10.12%	<u>0.66%</u>	35.45%	8.23%	23.05%	14.81%	23.53%
	12.87%	8	14.35%	<u>24.57%</u>	<u>2.18%</u>	21.91%	12.76%	11.12%	7.15%	28.51%
	10.89%	6	<u>30.09%</u>	32.66%	7.42%	15.91%	<u>29.22%</u>	4.90%	3.91%	14.93%
	7.27%	4	30.56%	20.52%	19.43%	6.09%	<u>24.28%</u>	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%

Table 2. Percent of profiles in each e cluster (column) for each c cluster (row). e cluster numbers are ordered roughly from smallest to largest values by comparing their corresponding e centroids in Fig. 3f, while the c cluster numbers are merely listed in numerically increasing order (arbitrarily). Bolding, coloring, and over- and underbars are for each row as in Table 1.

e : driest ... wettest	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%	<u>2.75%</u>	40.00%	34.67%	3.75%
	20.53%	7	9.34%	29.13%	29.49%	18.98%	8.06%	<u>0.00%</u>	35.91%	8.85%
	19.63%	2	19.78%	16.57%	15.21%	27.23%	22.59%	<u>0.00%</u>	15.79%	16.09%
	12.87%	8	15.73%	6.12%	9.31%	16.05%	<u>22.40%</u>	40.00%	<u>6.50%</u>	<u>21.18%</u>
	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%

whereas Fig. 5c highlights a profile in which a small bump in $N - \hat{N}$ and in water vapor correlates with a large $\Delta\Phi$. This supports the interpretation of $\Delta\Phi$ as being caused by an ice cloud. Figure 5a and b instead demonstrate the ability of the deviation from potential refractivity $N - \hat{N}$ to predict moisture distributions, even when $\Delta\Phi$ shows little to no correlation with these moisture changes as a function of height. For example, the profile in Fig. 5b shows negligible $\Delta\Phi$, suggesting that the water vapor profile likely indicates ice-crystal-free clouds from approximately 7.5 km down to nearly 5.5 km (see, e.g., the method in Peng et al., 2006).

Hence, Fig. 5b shows that differences in N from \hat{N} tend to correspond to altitudinal excursions from a near-exponential water vapor pressure as expected for a constant c_2 in Eq. (3). Table 1 verifies this by measuring the frequency with which different $N - \hat{N}$ clusters agree with specific e clusters; their centroids are shown in Fig. 3a and f, respectively. For example, Cluster 1 for $N - \hat{N}$ is the most flat and occurs most frequently to – and correlates most strongly with – Clusters

3 and 7 for e , the latter of which corresponds to profiles with little to no moisture. Conversely, Cluster 6 for $N - \hat{N}$ correlates well with the highest-moisture profiles in Clusters 1 and 5 for e and contains almost none of the low- or no-moisture profiles (Clusters 3, 7, 2, and 8 for e).

The $N - \hat{N}$ centroids in Fig. 3a tend to deviate from a constant value, primarily in the negative direction for $N - \hat{N}$ clusters associated with higher moisture, e.g., $N - \hat{N}$ Cluster 6. This indicates that $N < \hat{N}$ within a profile correlates with the presence of moisture, as a higher specific humidity generally increases refractivity (Friehe et al., 1975; Takamura et al., 1984; also see Eq. 1). Hence, because the potential refractivity \hat{N} is fit across both moist and dry regions of a profile, the background measured refractivity N in regions without moisture may fall below the vertically representative \hat{N} .

On the other hand, as shown in Table 1, Cluster 3 for $N - \hat{N}$ features larger values of $|N - \hat{N}|$ than Cluster 6 for $N - \hat{N}$ yet does not correlate with profiles that have a higher water

Table 3. Percent of profiles in each cluster for the column variable listed – liquid water path (LWP), ice water path (IWP), and liquid plus ice water path (TWP) – for each $\Delta\Phi$ cluster indicated by the row. Cluster numbers are ordered from smallest (most negative/zero) to largest (most positive) value by comparing their corresponding centroids in Fig. 3. Bolding, coloring, and over- and underbars are for each row as in Table 1.

		$\Delta\Phi$: most negative/zero			...	most positive					
LWP: driest	$\Delta\Phi$:	→	5.20%	25.32%	34.62%	23.45%	3.87%	5.74%	1.56%	0.24%	
	LWP:	↓	3	5	1	7	8	4	2	6	
		89.14%	96.35%	95.70%	95.25%	91.22%	66.07%	<u>54.52%</u>	<u>17.78%</u>	<u>7.14%</u>	
		5.34%	6	3.32%	3.28%	3.15%	6.05%	11.61%	<u>15.96%</u>	6.67%	0.00%
		2.85%	8	0.33%	0.75%	1.25%	1.99%	12.05%	15.96%	7.78%	28.57%
		1.51%	3	0.00%	0.20%	0.15%	0.44%	8.04%	11.14%	22.22%	7.14%
		0.81%	2	0.00%	0.00%	0.20%	0.22%	2.23%	2.11%	<u>27.78%</u>	28.57%
		0.28%	5	0.00%	0.00%	0.00%	0.07%	0.00%	0.30%	<u>14.44%</u>	<u>14.29%</u>
		0.04%	7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	<u>2.22%</u>	<u>7.14%</u>
		0.03%	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.11%	<u>7.14%</u>
...											
IWP: driest	$\Delta\Phi$:	→	5.20%	25.32%	34.62%	23.45%	3.87%	5.74%	1.56%	0.24%	
	IWP:	↓	3	5	1	7	8	4	2	6	
		85.78%	94.68%	94.26%	93.91%	85.03%	63.39%	<u>29.82%</u>	<u>3.33%</u>	<u>0.00%</u>	
		7.31%	8	2.99%	4.78%	4.80%	9.88%	<u>18.75%</u>	<u>19.58%</u>	3.33%	7.14%
		3.75%	5	1.99%	0.75%	0.85%	4.06%	8.93%	<u>32.23%</u>	10.00%	14.29%
		1.72%	3	0.33%	0.14%	0.35%	0.88%	7.14%	13.25%	21.11%	7.14%
		0.75%	6	0.00%	0.00%	0.05%	0.07%	0.00%	4.22%	31.11%	14.29%
		0.49%	2	0.00%	0.00%	0.05%	0.00%	1.34%	0.90%	<u>22.22%</u>	28.57%
		0.13%	4	0.00%	0.00%	0.00%	0.07%	0.45%	0.00%	<u>7.78%</u>	<u>0.00%</u>
		0.07%	7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.11%	<u>28.57%</u>
...											
TWP: driest	$\Delta\Phi$:	→	5.20%	25.32%	34.62%	23.45%	3.87%	5.74%	1.56%	0.24%	
	TWP:	↓	3	5	1	7	8	4	2	6	
		84.55%	1	93.02%	92.96%	93.26%	84.14%	60.71%	<u>28.92%</u>	<u>3.33%</u>	<u>0.00%</u>
		8.46%	6	4.98%	5.81%	5.09%	12.09%	15.63%	<u>24.40%</u>	3.33%	0.00%
		3.99%	5	1.99%	1.09%	1.15%	2.88%	17.41%	<u>29.82%</u>	12.22%	14.29%
		1.79%	3	0.00%	0.07%	0.45%	0.66%	5.36%	14.46%	<u>25.56%</u>	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%	<u>14.29%</u>
		0.09%	8	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	4.44%	<u>7.14%</u>
		0.07%	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	<u>3.33%</u>	<u>14.29%</u>
...											
wettest	$\Delta\Phi$:	→	5.20%	25.32%	34.62%	23.45%	3.87%	5.74%	1.56%	0.24%	
	LWP:	↓	3	5	1	7	8	4	2	6	
		89.14%	96.35%	95.70%	95.25%	91.22%	66.07%	<u>54.52%</u>	<u>17.78%</u>	<u>7.14%</u>	
		5.34%	6	3.32%	3.28%	3.15%	6.05%	11.61%	<u>15.96%</u>	6.67%	0.00%
		2.85%	8	0.33%	0.75%	1.25%	1.99%	12.05%	15.96%	7.78%	28.57%
		1.51%	3	0.00%	0.20%	0.15%	0.44%	8.04%	11.14%	22.22%	7.14%
		0.81%	2	0.00%	0.00%	0.20%	0.22%	2.23%	2.11%	<u>27.78%</u>	28.57%
		0.28%	5	0.00%	0.00%	0.00%	0.07%	0.00%	0.30%	<u>14.44%</u>	<u>14.29%</u>
		0.04%	7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	<u>2.22%</u>	<u>7.14%</u>
		0.03%	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.11%	<u>7.14%</u>

vapor pressure (i.e., Clusters 1 and 5 for e). The examples in Fig. 5 also demonstrate this; in particular, Fig. 5c features a profile with a notably higher value of e than the one in Fig. 5b yet exhibits smaller values of $|N - \hat{N}|$ overall. This suggests that the actual *magnitude* of deviations of N from \hat{N} does not necessarily correspond to the magnitude of water vapor pressure. Nonetheless, the clustering indicates a weak inverse relationship between $N - \hat{N}$ and e – the upper-left and bottom-right corners of Table 1 consist mostly of bolded grey

values, while the bottom-left and upper-right corners consist mostly of bolded black ones.

The aforementioned observation raises two possible hypotheses for why the relationship between the magnitudes of $N - \hat{N}$ and e is not more direct. Firstly, it is possible that the relationship between e and $N - \hat{N}$ is between the derivatives of one or both. Furthermore, \hat{N} is fit to most of the troposphere down to 2.5 km. Hence, the difference between the measured refractivity, N , and the potential refractivity model, \hat{N} , is most pronounced when there are concentrated

Table 4. Pearson’s correlation coefficient (r), Spearman’s rank correlation coefficient (ρ), and Kendall’s rank correlation coefficient (τ) for all pairs of the accumulated $\Delta\Phi$ vs. total column water vapor across varying latitudinal ranges for the raw dataset and the moving averages. Each correlation coefficient has a p value below 10^{-9} , indicating a high statistical significance for all coefficients.

Latitudinal range: \rightarrow Correlation coefficient: \downarrow	All	$> 50^\circ$	$\geq 20^\circ$ and $\leq 50^\circ$	$< 20^\circ$
Pearson’s r_p (raw dataset)	0.332	0.315	0.349	0.375
Spearman’s ρ_s (raw dataset)	0.216	0.223	0.206	0.287
Kendall’s τ_k (raw dataset)	0.147	0.151	0.139	0.194
Pearson’s r_p (moving averages)	0.940	0.901	0.921	0.708
Spearman’s ρ_s (moving averages)	0.971	0.964	0.947	0.683
Kendall’s τ_k (moving averages)	0.864	0.847	0.803	0.508

moisture anomalies within narrow bands of the troposphere. Conversely, the sensitivity of the derivatives of N and \hat{N} with respect to height suggests that there could be cases where a profile is moist, yet the model \hat{N} still closely matches the observed N . This can happen when a moist–unsaturated adiabatic lapse rate (Emanuel, 1994) holds throughout most of the profile. In such cases, $N - \hat{N}$ could be close to zero, even when the water vapor pressure remains elevated, provided that the water vapor pressure gradients remain small. As an example, the centroid for $N - \hat{N}$ Cluster 7 is relatively flat (Fig. 3a), but Table 1 shows that e Clusters 4 and 6 – both moderately high-moisture cases (Fig. 3f) – are the most commonly represented e clusters in $N - \hat{N}$ Cluster 7.

4.3 \hat{N} model coefficients and cluster groups

The \hat{N} coefficients $\mathbf{c} = (c_0, c_1, c_2)$ tend to only exhibit 2 degrees of freedom across the profiles in the dataset. Figure 6 shows how projecting the \mathbf{c} clusters onto the $(\hat{\Gamma}, \hat{T}_0)$ plane leads to a clear partitioning across different \mathbf{c} clusters. This suggests that the dominant clusters for \hat{e} (and therefore e) in the dataset are related to changes in $\hat{\Gamma}$ and \hat{T}_0 . Note that changes in $\hat{\Gamma}$ and \hat{T}_0 are related to changes in the polytropic index m and therewith the underlying heat transfer thermodynamics.

The clustering across \mathbf{c} was generally able to partition the physical and nonphysical fits. Clusters 2, 3, 7, and 8 for \mathbf{c} feature physical values of \hat{T}_0 and $\hat{\Gamma}$, while the other clusters feature nonphysically extreme values of \hat{T}_0 (mainly Cluster 6), $\hat{\Gamma}$ (Clusters 1 and 4), or both (Cluster 5). Such nonphysical fits indicate where the assumed physics is not reflective of the actual physics in those profiles. Sometimes, we observed that faulty retrievals fell within these clusters with unphysical profiles, suggesting (and perhaps identifying) retrieval issues – e.g., those discussed in Sect. 2 – rather than physical phenomena.

Figure 6 shows a moderately negative linear correlation between \hat{T}_0 and $\hat{\Gamma}$ for the fits which feature physically realistic values of $\hat{\Gamma}$. Between \hat{T}_0 and $\hat{\Gamma}$ across all latitudes for $\hat{\Gamma} > 0.1$, 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 . Each correlation coefficient has a p value below machine epsilon (i.e., at least below 2.22×10^{-16}), thereby showing the statistical significance of this negative correlation. This correlation reflects that the moist adiabatic lapse rate has a negative relationship with temperature for profiles with sufficient moisture. Since the moist adiabatic lapse rate approaches the dry adiabatic lapse rate for temperatures roughly below 230 K, a higher lapse rate can be observed for colder profiles.

Figure 7 features two scatter plots that show how the fit coefficient vector \mathbf{c} relates to the LWP and IWP as a function of height. The fit values of $\hat{\Gamma}$ and \hat{T}_0 generally do not correlate with path clusters. However, when $\hat{\Gamma} > 10^{-1} \text{ K km}^{-1}$ and $\hat{T}_0 > 280 \text{ K}$ for a given profile, that profile has little to no precipitation, as shown by the near-uniformity of Cluster 1 (turquoise) for either LWP or IWP in that region, as indicated by Fig. 7a and b, respectively. That is to say that \mathbf{c} is not too informative in confirming the presence of precipitation; however, \mathbf{c} can sometimes rule out the presence of moisture and thereby precipitation. Similar yet weaker relationships between \mathbf{c} and particular precipitation regimes can also be seen across other ranges of $\hat{\Gamma}$ and \hat{T}_0 in Fig. 7; e.g., $\hat{\Gamma} > 10 \text{ K km}^{-1}$ tends to also correlate with low- or no-moisture cases.

As the aforementioned relationship between $N - \hat{N}$ and e suggests, \mathbf{c} also exhibits an apparent relationship with e . Figure 6 suggests that Cluster 2 for \mathbf{c} tends to contain profiles where \hat{e} is near-zero. This tends to correspond to cases when e is too low for there to be precipitation: as confirmed in Fig. 4, when the water vapor pressure is too low, precipitation cannot form. The relationship between \mathbf{c} and e may be analyzed more precisely by looking at Table 2, which demonstrates the predictive power in using \mathbf{c} clusters to predict representative water vapor pressure profiles, i.e., the centroids for e clusters shown in Fig. 3f.

4.4 $\Delta\Phi$ and both liquid and ice water paths

Table 3 explores the correlation of PRO $\Delta\Phi$ profiles with precipitation in a given profile. Clusters with large $\Delta\Phi$ tend to correlate with those of large LWP or IWP, and inversely,

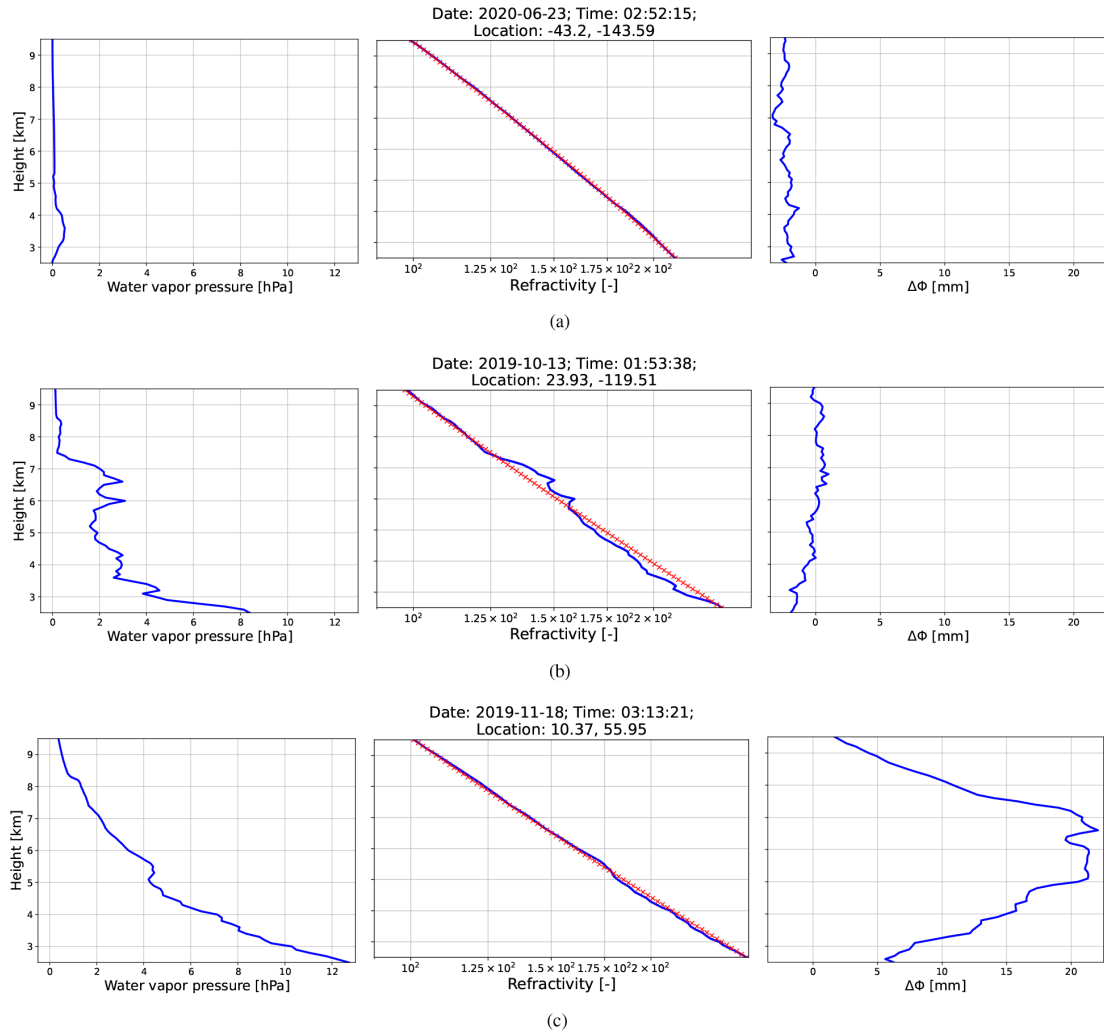


Figure 5. Three examples of thermodynamic profiles at various times and locations with (a) low moisture and no apparent precipitation, (b) some moisture but no apparent precipitation, and (c) high moisture and high precipitation. For each, we show the height on the y axis and the following on the x axes: e (left), N in blue and \bar{N} in red (center), and $\Delta\Phi$ (right). The date format is year-month-day.

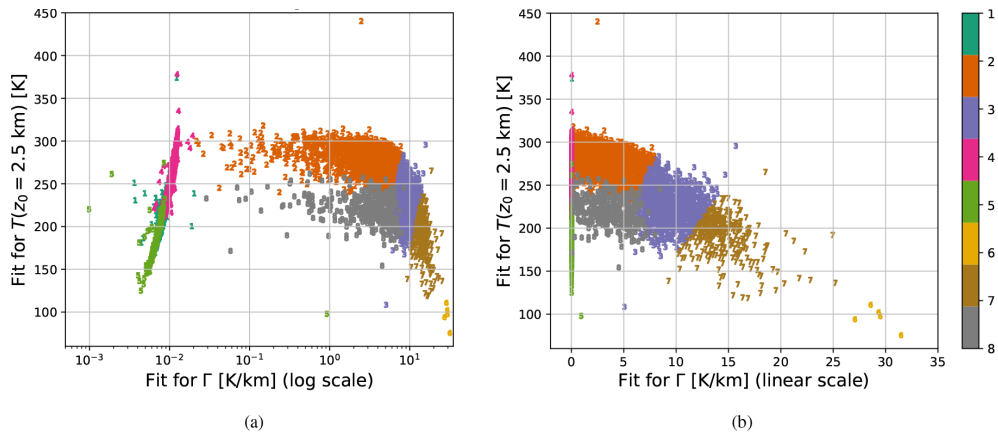


Figure 6. Scatter plots of the best-fit values of \hat{T}_0 vs. $\hat{\Gamma}$ across all latitudes in the dataset using (a) logarithmic scaling and (b) linear scaling in $\hat{\Gamma}$ to make the separation in $\hat{\Gamma}$ more apparent for small values of $\hat{\Gamma}$. The colors and symbols correspond to the associated \hat{N} coefficient (c) clusters for each point, as indicated by the color bars on the right-hand sides.

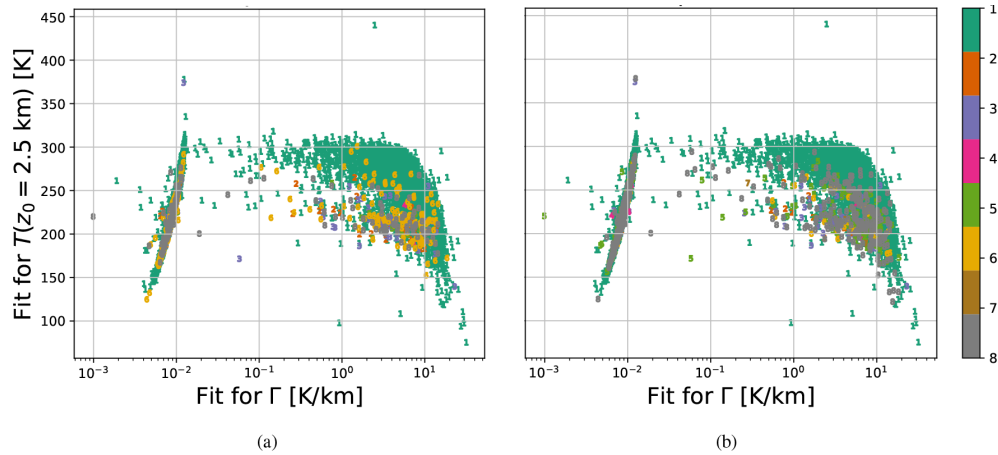


Figure 7. (a) LWP clusters and (b) IWP clusters over \hat{T}_0 vs. $\hat{\Gamma}$ across all profiles and latitudes.

those with small $\Delta\Phi$ also relate to profiles with little to no LWP or IWP. This should already be expected, as prior studies (e.g., Cardellach et al., 2019; Wang et al., 2022; Padullés et al., 2023) already indicate relationships between $\Delta\Phi$ and both water precipitation and ice.

Despite how Clusters 2 and 6 for $\Delta\Phi$ feature large values of $\Delta\Phi$ (> 4 mm) quite deep into the atmosphere – up to around 9 km according to their respective centroids – Table 3 shows that ice precipitation is not necessarily deep for those cases. In particular, Clusters 2 and 6 for $\Delta\Phi$ both correlate well with Clusters 2 and 6 for IWP, but the centroids for the latter two drop to zero near 7 and 5 km, respectively. This could be because $\Delta\Phi$ across different heights need not correspond one-to-one with the LWP or the IWP at those heights and also because LWP and IWP do not necessarily signal precipitation right at the time they are measured.

Even though the height of a particular onset or peak in $\Delta\Phi$ might not correlate with onsets or peaks, respectively, in the path cluster centroids, the shapes of the $\Delta\Phi$ and total path cluster centroids appear to correlate in both precipitation and non-precipitation cases, as demonstrated in Table 3. This consistency in shape but not in height is a property of the DTW measure used for the clustering. Hence, the lack of height correlations in our clusters does not contradict the model predictions of Padullés et al. (2023) since their model directly matches features in $\Delta\Phi$ and precipitation as a function of height.

5 Conclusions

In summary, *k*-means clustering has been used to evaluate its ability to identify different types of correlations between the vertical distributions of precipitation- and moisture-related variables. Our work shows the application and physical interpretability of using an unsaturated polytropic potential refractivity fit, \hat{N} , when there is a linear temperature pro-

file with height, which is expected in a polytropic atmosphere. Deviations from \hat{N} relate to the presence of water vapor pressure anomalies at given latitudes, longitudes, and times (Sect. 4.2). In particular, Table 1 demonstrates a visibly strong yet non-monotonic relationship between the shapes and amplitudes of $N - \hat{N}$ vs. *e*. For instance, the moderately negative Cluster 6 for $N - \hat{N}$ corresponds well with very moist profiles, yet the more negative clusters for $N - \hat{N}$ correspond to only moderately moist profiles. Inversely, the mostly flat Cluster 1 for $N - \hat{N}$ corresponds to profiles with little to no moisture (Clusters 3 and 7 for *e*), yet the most positive Cluster 5 for $N - \hat{N}$ corresponds to profiles with low to moderate moisture. This can be explained by how the deviation of *N* from \hat{N} will be muted if \hat{N} has been fit to a profile which is moist overall, and thereby $|N - \hat{N}|$ will be largest when the moisture is high and relatively localized (e.g., in the presence of clouds).

\hat{N} coefficient (*c*) clusters can flag physical vs. nonphysical values of observed and derived variables (Sect. 4.3, Fig. 6). As shown in Fig. 6, Clusters 5–7 for *c* generally correspond to temperature values that are far too low, indicating either a problem with the data from the retrievals or a profile which does not satisfy the physical assumptions made in deriving \hat{N} (see Sect. 3.1). Inversely, the values of *c* for a given profile can identify when a profile has no moisture or precipitation with very high accuracy – as shown in Fig. 7, profiles with $\hat{\Gamma} > 10^{-1} \text{ K km}^{-1}$ and $\hat{T}_0 > 280 \text{ K}$ have little to no precipitation. Related correlations between different *c* and *e* clusters are also shown in Table 2, where we see that different clusters for *c* correspond to profiles with low, medium, and high water vapor pressure throughout.

Similarly, vertical distributions of $\Delta\Phi$ are found to correlate to specific vertical profiles of liquid and ice precipitation. In particular, the amplitude and shape of $\Delta\Phi$ centroids correlate with the amplitudes and shapes of LWP and IWP centroids, respectively (Sect. 4.4, Table 3). This correlation persists across low and high levels of LWP, IWP, and both

combined, thereby demonstrating a strong one-to-one relationship between $\Delta\Phi$ and water path.

In conclusion, the clustering centroids (i.e., “representative” profiles) correlate with the general magnitude of a variable for a given profile and also the general shape of that variable as a function of height. The latter is especially evident for variables that correlate with water content: $\Delta\Phi$ and the path variables. As a demonstration of how the centroids capture the magnitude of profiles in their associated clusters, consider the ice water path (IWP) clusters shown in Fig. 3c: Clusters 4 and 7 for IWP both correspond to higher-than-average ice content in their respective profiles, and a similar comparison can be drawn between Clusters 2 and 6 for IWP. Relatedly, as a demonstration of how the centroids capture the shape, consider the liquid water path (LWP) clusters shown in Fig. 3d: Clusters 2 and 5 for LWP both correspond to non-negligible water precipitation, but Cluster 5 features profiles with deeper precipitation than those in Cluster 2. Thus, clustering in the manner introduced in this study confirms its value as a tool for the quality control of profiles and automates the classification of physical phenomena found across large datasets.

Appendix A: Derivation of \hat{N}

Combining the equation for hydrostatic equilibrium and the ideal gas law, we have

$$p(z) = p(z_0) \exp\left(-\frac{g}{R} \int_{z_0}^z \frac{ds}{T(s)}\right), \quad (\text{A1})$$

where $g = 9.8 \text{ g m}^{-2}$ is the acceleration due to gravity on Earth and $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$ is the specific gas constant for dry air. In a polytropic atmosphere, $T(z) = T(z_0) - \Gamma(z - z_0)$ for a lapse rate Γ to be determined by a fit to the data together with $T(z_0)$. The integral in Eq. (A1) for this temperature profile can be computed as

$$-\frac{g}{R} \int_{z_0}^z \frac{ds}{T(s)} = \frac{g}{R\Gamma} \log\left[1 - \frac{\Gamma(z - z_0)}{T(z_0)}\right],$$

which in turn implies that (e.g., Dutton, 1976)

$$p(z) = p(z_0) \left[1 - \frac{\Gamma}{T(z_0)}(z - z_0)\right]^{\frac{g}{R\Gamma}}. \quad (\text{A2})$$

Substituting Eq. (A2) and $T(z) = T(z_0) - \Gamma(z - z_0)$ into Eq. (1) and putting a hat on N since \hat{N} is the idealized model, we have

$$\hat{N}(z) = \frac{k_1 p(z_0)}{T(z_0) \left[1 - \frac{\Gamma(z - z_0)}{T(z_0)}\right]^2} \times \left\{ \left[1 - \frac{\Gamma(z - z_0)}{T(z_0)}\right]^{\frac{g}{R\Gamma} + 1} + \frac{k_2 e}{k_1 p(z_0) T(z_0)} \right\}. \quad (\text{A3})$$

While $p(z_0)$ might not be available directly in a typical PRO profile (which only contains refractivity and $\Delta\Phi$), there will be data for $N(z_0)$. Hence, we solve for $p(z_0)$ in terms of $T(z_0)$ and $N(z_0)$ to constrain the number of fitting parameters. Rewriting Eq. (1) at $z = z_0$, we have

$$k_1 p(z_0) = T(z_0) \left[N(z_0) - \frac{k_2 e}{T(z_0)^2} \right]. \quad (\text{A4})$$

Substituting Eq. (A4) into Eq. (A3) for a constant, representative, \hat{e}_0 , and rewriting in terms of the fit coefficients c_0 , c_1 , and c_2 , leads to Eq. (3).

Appendix B: Numerical fitting procedure for \hat{N}

Once Eq. (3) has been fit to a given profile, we can use c_0 to solve for $\hat{\Gamma}$, then use this and c_1 to solve for \hat{T}_0 , and finally use c_2 and \hat{T}_0 to solve for the representative value of \hat{e} . To do this fitting routine in practice, since k_2 , \hat{e} , and $N \geq 0$, we impose the constraint $c_2 \geq 0$ and use the curve-fitting utility `optimize.least_squares` from the SciPy package (version 1.7.3) in Python 3.7.4 with the initial conditions $c_0 = 4.5$, $c_1 = 0.01 \text{ m}^{-1}$, and $c_2 = 0$ to fit Eq. (3) to each N profile for all cases. For reasons which are generally internal to the default `optimize.least_squares` algorithm, either the nonlinear fitting procedure did not always converge within the preset maximum number of iterations, 10 000, with prescribed error tolerances `ftol = xtol = 10-12`, or the profile in question was missing too much data for the model coefficients to be uniquely determinable – this only occurred in 5 profiles out of the 6706 in the dataset or 0.07%. The latter could have occurred either because there were not enough data overall or because there were no refractivity data at $z_0 = 2.5 \text{ km}$.

Appendix C: Clustering algorithm for k -means

For the numerical implementation of time series k -means clustering, we use version 0.6.2 of the Python package `tslearn`, which provides machine learning tools for the analysis of time series data and builds on the `scikit-learn`, `scipy`, and `numpy` libraries (Tavenard et al., 2020). To run time series k -means clustering for all variables in the dataset, we use `tslearn.clustering.TimeSeriesKMeans` with $k = 8$ clusters, the DTW metric, and a maximum of 30 iterations of the algorithm, and we fix the random state to 0 to ensure that the cluster labels stay consistent upon each run.

There are ways to estimate the most “statistically meaningful” number of clusters for a given dataset, even when not using the Euclidean metric – e.g., the average silhouette method (Rousseeuw, 1987) or the gap statistic method (Tibshirani et al., 2001) – which could give different numbers

of clusters for each variable. However, to keep a consistent number of clusters for each variable and to give some semblance of the same hierarchy in magnitude across clustering in each variable, this study uses the same number of clusters for all variables and defers to using a number that is possibly too large rather than too small.

Appendix D: Quality-control criteria for clustering

Profiles are excluded from each cluster according to the quality-control criteria listed below:

- For $\Delta\Phi$ (923 profiles excluded or 13.76 %), files are excluded by the same criteria used for the vertical integral of $\Delta\Phi$.
- For c_0 , c_1 , and c_2 (5 profiles excluded or 0.07 %), the fit for \hat{N} must converge, i.e., the algorithm for computing the best-fit coefficients c_0 , c_1 , and c_2 must converge, which means that there must be refractivity data at 2.5 km, there must be enough refractivity data between 2.5 km and the estimated lapse-rate tropopause (the latter of which was explained earlier), and the fit must converge within 10 000 iterations for tolerance conditions $ftol = xtol = 10^{-12}$.
- For $N - \hat{N}$ (223 profiles excluded or 3.33 %), along with the same criteria related to \hat{N} used for the coefficient clusters, cases where the tropopause is below 8.2 km are skipped and three files from clustering for $N - \hat{N}$ are taken out manually and excluded. These three files contained unphysically large values of N ($N > 600$) and likely indicate an issue with retrieving the refractivity for the RO dataset.
- For water vapor pressure (33 profiles excluded or 0.49 %), files are excluded by the same criteria used for the total column water vapor described in Sect. 2. It should be noted that the three files with unphysically large values of N that were manually excluded from clustering for $N - \hat{N}$ also had unphysically large water vapor pressure values. Hence, these profiles were also excluded from the clustering for water vapor pressure, thereby showing that at least some of the erroneous water vapor pressure retrievals were caused by issues with the retrieved RO refractivity.
- For LWP (1 profile excluded or 0.01 %), files without LWP data from 1 to 10 km are excluded.
- For IWP (6 profiles excluded or 0.09 %), files without IWP data from 1 to 10 km are excluded.
- For TWP (6 profiles excluded or 0.09 %), files without LWP or IWP data from 1 to 10 km are excluded.

Code and data availability. The datasets associated with this study have been uploaded to the Jet Propulsion Laboratory’s GENESIS (Global Environmental and Earth Science Information System) site: https://genesis.jpl.nasa.gov/ftp/paz_pol/ (Padullés et al., 2020). Further ROHP data are available at <https://paz.ice.csic.es> (Cardellach et al., 2020). GPM Level 1C, 2A, and 2B PMW radiometer data are openly available via the Precipitation Processing System (PPS) at NASA Goddard Space Flight Center: <https://pps.gsfc.nasa.gov/> (NASA, 2025).

Author contributions. Conceptualization: JK, MTJ, KNW. Data curation: JT, KNW, RP. Formal analysis: JK, MTJ, TK. Funding acquisition: MTJ. Investigation: all. Methodology: JK, MTJ, TK, KNW. Project administration: MTJ. Resources: MTJ, KNW. Software: all. Supervision: MTJ, TK, JT. Validation: all. Visualization: JK, MTJ, TK. Writing (original draft preparation): JK, MTJ, TK. Writing (review and editing): JK, MTJ, TK, JT, RP.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Special issue statement. This article is part of the special issue “Observing atmosphere and climate with occultation techniques – results from the OPAC-IROWG 2022 workshop”. It is a result of the International Workshop on Occultations for Probing Atmosphere and Climate, Leibnitz, Austria, 8–14 September 2022.

Acknowledgements. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under the JPL Visiting Student Research Program, with financial support from NASA – as detailed under “Financial support” below – and with a stipend and teaching fellowship from the Yale Graduate School of Arts and Sciences. The authors would like to thank Joe Turk for collecting and preparing the GPM dataset; Kuo-Nung Wang and Ramon Padullés for preparing and managing the ROHP-PAZ dataset and collocations between the GPM and ROHP-PAZ datasets; and various technical support staff at the Jet Propulsion Laboratory for their tireless help with data, equipment, and account access. We would also like to thank Chi O. Ao for helping to manage and prepare the data and software resources used in this study.

Additionally, the authors thank Hui Shao and Andreas Richter, the handling editors, for their time and effort in supervising the review process for this paper. The authors are also grateful to the two anonymous referees for their detailed, constructive feedback on the initial submission, as well as to the Copernicus Publications editorial team and support staff for their diligence in ensuring that the final paper is well-presented and adheres to journal requirements.

Financial support. This research has been supported by the National Aeronautics and Space Administration (under program NH19ZDA001N-GNSS, contract no. 80NM0018D0004).

Review statement. This paper was edited by Hui Shao and reviewed by two anonymous referees.

References

- Ao, C. O., Waliser, D. E., Chan, S. K., Li, J.-L., Tian, B., Xie, F., and Mannucci, A. J.: Planetary boundary layer heights from GPS radio occultation refractivity and humidity profiles, *J. Geophys. Res.-Atmos.*, 117, D16117, <https://doi.org/10.1029/2012JD017598>, 2012.
- Atkinson, K. E.: *An Introduction to Numerical Analysis*, Wiley, New York, ISBN 978-0471624899, 1988.
- Bean, B. and Dutton, E.: Radio Meteorology, no. 92 in (National Bureau of Standards), U.S. Govt. Print. Off., <https://api.semanticscholar.org/CorpusID:124549052> (last access: 28 April 2024), 1966.
- Biondi, R., Randel, W. J., Ho, S.-P., Neubert, T., and Syndergaard, S.: Thermal structure of intense convective clouds derived from GPS radio occultations, *Atmos. Chem. Phys.*, 12, 5309–5318, <https://doi.org/10.5194/acp-12-5309-2012>, 2012.
- Bretherton, C. S., Peters, M. E., and Back, L. E.: Relationships between Water Vapor Path and Precipitation over the Tropical Oceans, *J. Climate*, 17, 1517–1528, [https://doi.org/10.1175/1520-0442\(2004\)017<1517:rbwvpa>2.0.co;2](https://doi.org/10.1175/1520-0442(2004)017<1517:rbwvpa>2.0.co;2), 2004.
- Cardellach, E., Oliveras, S., Rius, A., Tomás, S., Ao, C. O., Franklin, G. W., Iijima, B. A., Kuang, D., Meehan, T. K., Padullés, R., de la Torre Juárez, M., Turk, F. J., Hunt, D. C., Schreiner, W. S., Sokolovskiy, S. V., Hove, T. V., Weiss, J. P., Yoon, Y., Zeng, Z., Clapp, J., Xia-Serafino, W., and Cerezo, F.: Sensing Heavy Precipitation With GNSS Polarimetric Radio Occultations, *Geophys. Res. Lett.*, 46, 1024–1031, <https://doi.org/10.1029/2018gl080412>, 2019.
- Cardellach, E., Padullés, R., and Oliveras, S.: Radio Occultation and Heavy Precipitation with PAZ (ROHP-PAZ), <https://paz.ice.csic.es/dataAcces.php?idi=EN> (last access: 29 December 2023), 2020.
- de la Torre Juárez, M., Padullés, R., Turk, F. J., and Cardellach, E.: Signatures of Heavy Precipitation on the Thermodynamics of Clouds Seen From Satellite: Changes Observed in Temperature Lapse Rates and Missed by Weather Analyses, *J. Geophys. Res.-Atmos.*, 123, 13033–13045, <https://doi.org/10.1029/2017JD028170>, 2018.
- Dutton, J.: *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion*, McGraw-Hill, ISBN 9780070184077, <https://books.google.com/books?id=9CxRAAAAMAAJ> (last access: 28 April 2024), 1976.
- Emanuel, K.: *Atmospheric Convection*, Oxford University Press, ISBN 9780195066302, <https://books.google.com/books?id=VdaBBHEGAcMC> (last access: 28 April 2024), 1994.
- Emmenegger, T., Kuo, Y.-H., Xie, S., Zhang, C., Tao, C., and Neelin, J. D.: Evaluating Tropical Precipitation Relations in CMIP6 Models with ARM Data, *J. Climate*, 35, 6343–6360, <https://doi.org/10.1175/JCLI-D-21-0386.1>, 2022.
- Friehe, C. A., Rue, J. C. L., Champagne, F. H., Gibson, C. H., and Dreyer, G. F.: Effects of temperature and humidity fluctuations on the optical refractive index in the marine boundary layer, *J. Opt. Soc. Am.*, 65, 1502–1511, <https://doi.org/10.1364/JOSA.65.001502>, 1975.
- Fueglistaler, S., Dessler, A. E., Dunkerton, T. J., Folkins, I., Fu, Q., and Mote, P. W.: Tropical tropopause layer, *Rev. Geophys.*, 47, RG1004, <https://doi.org/10.1029/2008RG000267>, 2009.
- Gleisner, H., Ringer, M. A., and Healy, S. B.: Monitoring global climate change using GNSS radio occultation, *Clim. Atmos. Sci.*, 5, 6, <https://doi.org/10.1038/s41612-022-00229-7>, 2022.
- Govender, P. and Sivakumar, V.: Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), *Atmos. Pollut. Res.*, 11, 40–56, <https://doi.org/10.1016/j.apr.2019.09.009>, 2020.
- Hajj, G., Kursinski, E., Romans, L., Bertiger, W., and Leroy, S.: A technical description of atmospheric sounding by GPS occultation, *J. Atmos. Solar-Terrest. Phys.*, 64, 451–469, [https://doi.org/10.1016/S1364-6826\(01\)00114-6](https://doi.org/10.1016/S1364-6826(01)00114-6), 2002.
- Holloway, C. E. and Neelin, J. D.: Temporal Relations of Column Water Vapor and Tropical Precipitation, *J. Atmos. Sci.*, 67, 1091–1105, <https://doi.org/10.1175/2009JAS3284.1>, 2010.
- Hotta, D., Lonitz, K., and Healy, S.: Forward operator for polarimetric radio occultation measurements, *Atmos. Meas. Tech.*, 17, 1075–1089, <https://doi.org/10.5194/amt-17-1075-2024>, 2024.
- Huang, Y., Leroy, S. S., and Anderson, J. G.: Determining Longwave Forcing and Feedback Using Infrared Spectra and GNSS Radio Occultation, *J. Climate*, 23, 6027–6035, <https://doi.org/10.1175/2010jcli3588.1>, 2010.
- Izakian, H., Pedrycz, W., and Jamal, I.: Fuzzy clustering of time series data using dynamic time warping distance, *Eng. Appl. Artif. Intell.*, 39, 235–244, <https://doi.org/10.1016/j.engappai.2014.12.015>, 2015.
- Jakob, C. and Tselioudis, G.: Objective identification of cloud regimes in the Tropical Western Pacific, *Geophys. Res. Lett.*, 30, 2082, <https://doi.org/10.1029/2003GL018367>, 2003.
- Johnston, B. R., Xie, F., and Liu, C.: The Effects of Deep Convection on Regional Temperature Structure in the Tropical Upper Troposphere and Lower Stratosphere, *J. Geophys. Res.-Atmos.*, 123, 1585–1603, <https://doi.org/10.1002/2017JD027120>, 2018.
- Johnston, B. R., Xie, F., and Liu, C.: Relationships between Extratropical Precipitation Systems and UTLS Temperatures and Tropopause Height from GPM and GPS-RO, *Atmosphere*, 13, 196, <https://doi.org/10.3390/atmos13020196>, 2022.
- Kim, J. and Son, S.-W.: Tropical Cold-Point Tropopause: Climatology, Seasonal Cycle, and Intraseasonal Variability Derived from COSMIC GPS Radio Occultation Measurements, *J. Climate*, 25, 5343–5360, <https://doi.org/10.1175/jcli-d-11-00554.1>, 2012.
- Kliore, A., Cain, D. L., Fjeldbo, G., Seidel, B. L., and Rasool, S. I.: Preliminary Results on the Atmospheres of Io and Jupiter from the Pioneer 10 S-Band Occultation Experiment, *Science*, 183, 323–324, <https://doi.org/10.1126/science.183.4122.323>, 1974.
- Kuo, Y.-H., Sokolovskiy, S., Anthes, R., and Vandenberghe, F.: Assimilation of GPS Radio Occultation Data for Numerical Weather Prediction, *Terrestrial, Atmos. Ocean. Sci.*, 11, 157–186, [https://doi.org/10.3319/TAO.2000.11.1.157\(COSMIC\)](https://doi.org/10.3319/TAO.2000.11.1.157(COSMIC)), 2001.

- Kursinski, E. R., Hajj, G. A., Schofield, J. T., Linfield, R. P., and Hardy, K. R.: Observing Earth's atmosphere with radio occultation measurements using the Global Positioning System, *J. Geophys. Res.-Atmos.*, 102, 23429–23465, <https://doi.org/10.1029/97JD01569>, 1997.
- Lackner, B. C., Steiner, A. K., Hegerl, G. C., and Kirchengast, G.: Atmospheric Climate Change Detection by Radio Occultation Data Using a Fingerprinting Method, *J. Climate*, 24, 5275–5291, <https://doi.org/10.1175/2011jcli3966.1>, 2011.
- Mascio, J., Leroy, S. S., d'Entremont, R. P., Connor, T., and Kursinski, E. R.: Using Radio Occultation to Detect Clouds in the Middle and Upper Troposphere, *J. Atmos. Ocean. Technol.*, 38, 1847–1858, <https://doi.org/10.1175/JTECH-D-21-0022.1>, 2021.
- Müller, M.: Dynamic Time Warping, Information Retrieval for Music and Motion, Springer Berlin Heidelberg, Berlin, Heidelberg, 69–84, https://doi.org/10.1007/978-3-540-74048-3_4, 2007.
- Muller, C. J., Back, L. E., O'Gorman, P. A., and Emanuel, K. A.: A model for the relationship between tropical precipitation and column water vapor, *Geophys. Res. Lett.*, 36, L16804, <https://doi.org/10.1029/2009GL039667>, 2009.
- NASA: CloudSat, <https://eosps.nasa.gov/missions/cloudsat> (last access: 28 April 2024), 2024.
- NASA: Precipitation Processing System (PPS), NASA, <https://pps.gsfc.nasa.gov/> (last access: 29 December 2023), 2025.
- Nidzgorska-Lencewicz, J. and Czarnecka, M.: Thermal Inversion and Particulate Matter Concentration in Wrocław in Winter Season, *Atmosphere*, 11, 1351, <https://doi.org/10.3390/atmos11121351>, 2020.
- Nishida, M., Shimizu, A., Tsuda, T., Rocken, C., and Ware, R. H.: Seasonal and Longitudinal Variations in the Tropical Tropopause Observed with the GPS Occultation Technique (GPS/MET), *J. Meteorol. Soc. JPN II*, 78, 691–700, https://doi.org/10.2151/jmsj1965.78.6_691, 2000.
- Padullés, R., Ao, C. O., Turk, F. J., de la Torre Juárez, M., Iijima, B., Wang, K. N., and Cardellach, E.: PAZ calibrated polarimetric products, NASA [data set], https://genesis.jpl.nasa.gov/ftp/paz_pol/ (last access: 29 December 2023), 2021.
- Padullés, R., Cardellach, E., and Turk, F. J.: On the global relationship between polarimetric radio occultation differential phase shift and ice water content, *Atmos. Chem. Phys.*, 23, 2199–2214, <https://doi.org/10.5194/acp-23-2199-2023>, 2023.
- Peng, G., de la Torre-Juárez, M., Farley, R., and Wessel, J.: Impacts of upper tropospheric clouds on GPS radio refractivity, in: 2006 IEEE Aerospace Conference, 6 pp., <https://doi.org/10.1109/AERO.2006.1655899>, 2006.
- Randel, W. J., Wu, F., and Ríos, W. R.: Thermal variability of the tropical tropopause region derived from GPS/MET observations, *J. Geophys. Res.*, 108, 4024, <https://doi.org/10.1029/2002JD002595>, 2003.
- Rossow, W. B., Tselioudis, G., Polak, A., and Jakob, C.: Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures, *Geophys. Res. Lett.*, 32, L21812, <https://doi.org/10.1029/2005GL024584>, 2005.
- Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Mathe.*, 20, 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7), 1987.
- Ruston, B. and Healy, S.: Forecast Impact of FORMOSAT-7/COSMIC-2 GNSS Radio Occultation Measurements, *Atmos. Sci. Lett.*, 22, e1019, <https://doi.org/10.1002/asl.1019>, 2021.
- Scherllin-Pirscher, B., Steiner, A. K., Anthes, R. A., Alexander, M. J., Alexander, S. P., Biondi, R., Birner, T., Kim, J., Randel, W. J., Son, S.-W., Tsuda, T., and Zeng, Z.: Tropical Temperature Variability in the UTLS: New Insights from GPS Radio Occultation Observations, *J. Climate*, 34, 2813–2838, <https://doi.org/10.1175/jcli-d-20-0385.1>, 2021.
- Schmidt, T., Wickert, J., Beyerle, G., and Reigber, C.: Tropical tropopause parameters derived from GPS radio occultation measurements with CHAMP, *J. Geophys. Res.-Atmos.*, 109, D13105, <https://doi.org/10.1029/2004jd004566>, 2004.
- Smith, E. K. and Weintraub, S.: The Constants in the Equation for Atmospheric Refractive Index at Radio Frequencies, *Proc. IRE*, 41, 1035–1037, <https://doi.org/10.1109/JRPROC.1953.274297>, 1953.
- Son, S.-W., Lim, Y., Yoo, C., Hendon, H. H., and Kim, J.: Stratospheric Control of the Madden–Julian Oscillation, *J. Climate*, 30, 1909–1922, <https://doi.org/10.1175/jcli-d-16-0620.1>, 2017.
- Takamura, T., Tanaka, M., and Nakajima, T.: Effects of Atmospheric Humidity on the Refractive Index and the Size Distribution of Aerosols as Estimated from Light Scattering Measurements, *J. Meteorol. Soc. JPN II*, 62, 573–582, https://doi.org/10.2151/jmsj1965.62.3_573, 1984.
- Tavenard, R.: An introduction to Dynamic Time Warping, <https://rtavenar.github.io/blog/dtw.html> (last access: 28 April 2024), 2021.
- Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., and Woods, E.: Tslearn, A Machine Learning Toolkit for Time Series Data, *J. Mach. Learn. Res.*, 21, 1–6, <http://jmlr.org/papers/v21/20-091.html> (last access: 28 April 2024), 2020.
- Tibshirani, R., Walther, G., and Hastie, T.: Estimating the Number of Clusters in a Data Set Via the Gap Statistic, *J. Roy. Stat. Soc. B*, 63, 411–423, <https://doi.org/10.1111/1467-9868.00293>, 2001.
- Tomás, S., Padullés, R., and Cardellach, E.: Separability of Systematic Effects in Polarimetric GNSS Radio Occultations for Precipitation Sensing, *IEEE T. Geosci. Remote Sens.*, 56, 4633–4649, <https://doi.org/10.1109/tgrs.2018.2831600>, 2018.
- Turk, F. J., Haddad, Z. S., Kirstetter, P.-E., You, Y., and Ringerud, S. E.: An observationally based method for stratifying a priori passive microwave observations in a Bayesian-based precipitation retrieval framework, *Q. J. Roy. Meteor. Soc.*, 144, 145–164, <https://doi.org/10.1002/qj.3203>, 2018.
- Turk, F. J., Padullés, R., Ao, C. O., Juárez, M. d. I. T., Wang, K.-N., Franklin, G. W., Lowe, S. T., Hristova-Veleva, S. M., Fetzer, E. J., Cardellach, E., Kuo, Y.-H., and Neelin, J. D.: Benefits of a Closely-Spaced Satellite Constellation of Atmospheric Polarimetric Radio Occultation Measurements, *Remote Sens.*, 11, 2399, <https://doi.org/10.3390/rs11202399>, 2019.
- Turk, F. J., Padullés, R., Cardellach, E., Ao, C. O., Wang, K.-N., Morabito, D. D., de la Torre Juárez, M., Oyola, M., Hristova-Veleva, S., and Neelin, J. D.: Interpretation of the Precipitation Structure Contained in Polarimetric Radio Occultation Profiles Using Passive Microwave Satellite Observations, *J. Atmos. Ocean. Technol.*, 38, 1727–1745, <https://doi.org/10.1175/JTECH-D-21-0044.1>, 2021.
- Utsumi, N., Turk, F. J., Haddad, Z. S., Kirstetter, P.-E., and Kim, H.: Evaluation of precipitation vertical profiles estimated by GPM-era satellite-based passive microwave retrievals, *J. Hydrometeorol.*, 22, 95–112, <https://doi.org/10.1175/JHM-D-20-0160.1>, 2020.

- von Engel, A., Teixeira, J., Wickert, J., and Buehler, S. A.: Using CHAMP radio occultation data to determine the top altitude of the Planetary Boundary Layer, *Geophys. Res. Lett.*, 32, L06815, <https://doi.org/10.1029/2004GL022168>, 2005.
- Wang, K.-N., Ao, C. O., Padullés, R., Turk, F. J., de la Torre Juárez, M., and Cardellach, E.: The Effects of Heavy Precipitation on Polarimetric Radio Occultation (PRO) Bending Angle Observations, *J. Atmos. Ocean. Technol.*, 39, 149–161, <https://doi.org/10.1175/jtech-d-21-0032.1>, 2022.
- Wee, T.-K., Anthes, R. A., Hunt, D. C., Schreiner, W. S., and Kuo, Y.-H.: Atmospheric GNSS RO 1D-Var in Use at UCAR: Description and Validation, *Remote Sens.*, 14, 5614, <https://doi.org/10.3390/rs14215614>, 2022.
- Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., and Tian, D.: The Generic Mapping Tools Version 6, *Geochem. Geophys. Geosyst.*, 20, 5556–5564, <https://doi.org/10.1029/2019gc008515>, 2019.
- Wilks, D. S.: *Statistical methods in the atmospheric sciences*, Elsevier Science Publishing, Philadelphia, PA, 4 edn., Elsevier, <https://doi.org/10.1016/c2017-0-03921-6>, 2019.
- Xian, T. and Fu, Y.: Characteristics of tropopause-penetrating convection determined by TRMM and COSMIC GPS radio occultation measurements, *J. Geophys. Res.-Atmos.*, 120, 7006–7024, <https://doi.org/10.1002/2014jd022633>, 2015.
- Yokoi, S., Takayabu, Y. N., Nishii, K., Nakamura, H., Endo, H., Ichikawa, H., Inoue, T., Kimoto, M., Kosaka, Y., Miyasaka, T., Oshima, K., Sato, N., Tsushima, Y., and Watanabe, M.: Application of Cluster Analysis to Climate Model Performance Metrics, *J. Appl. Meteorol. Climatol.*, 50, 1666–1675, <https://doi.org/10.1175/2011JAMC2643.1>, 2011.