Supporting Information for "Cloud patterns have four interpretable dimensions"

DOI: 10.1002/

Martin Janssens[1], Jordi Vilà-Guerau de Arellano[1], Marten Scheffer[1], Coco Antonissen[2], A. Pier Siebesma[23], Franziska Glassmeier[2]

[1]Wageningen University & Research

[2]Delft University of Technology

[3]Royal Netherlands Meteorological Institute

Contents of this file

Corresponding author: M. Janssens, Departments of Meteorology & Air Quality and Aquatic Ecology & Water Quality Management, Wageningen University & Research, Wageningen, Lumen Building, Droevendaalsesteeg 3a, 6708 PB Wageningen, The Netherlands, (martin.janssens@wur.nl)

Introduction

This supplement contains further descriptions of the metrics that we use to characterise our cloud field pattern distribution (Text S1). Specifically, we elaborate upon details of and justify choices made in their computation. Code that evaluates these metrics given input scenes of cloud mask, cloud water path and cloud top height can be found in our accompanying GitHub repository (`https://github.com/martinjanssens/cloudmetrics`) and Figshare copy of this repository at the time of publication `https://figshare.com/projects/Cloud_field_organisation_description_with_metrics/86303`. The supplement also contains five figures, that quantify i) the sensitivity of our metric distribution to field resolution, object segmentation strategy and minimum cloud size, ii) the absolute Pearson correlation between all metrics, iii) the fraction of variance in each metric explained by every PC, iv) an estimate of the quality of our metric-based approach to approximating cloud field patterns and v) the sensitivity to free parameters of approximating principal components with a subset of metrics through sparse principal component analysis.

Text S1. - Details of metrics

Statistical moments of cloud field properties

We quantify several statistics of the extracted cloud field products. Some of these are straightforward computations that do not feature design choices (cloud fraction, total cloud water, standard deviation of cloud water over cloudy pixels). The other metrics require further qualification.

Mean and standard deviation of cloud top height ($\overline{\mathrm{CTH}}$ and St(CTH) respectively) i) explicitly ignore clouds higher than 5km, as cirrus wisps were found to disproportionately affect the results otherwise and ii) only consider cloudy pixels. Higher-order moments of these fields were small and are therefore not included.

Cloud water variance ratio R (CWP var. ratio) is directly adopted from Bretherton and Blossey (2017), but instead of being applied to the total, vertically integrated moisture field, it is here only applied to the cloud water:

$$R = \frac{Std\left(\overline{CWP_b} - \overline{CWP}\right)}{Std(CWP)} \tag{1}$$

In this relation, $\bar{\cdot}$ denotes a domain average and CWP_b indicates the cloud water contained in blocks of 16x16 pixels.

Object-based metrics

Object-based metrics follow from segmenting the cloud mask field into N_o objects according to their 4-connectivity. To avoid artefacts at the grid scale, we only consider objects with areas larger than 4 pixels. Each extracted object covers an area A_i, such that a typical length scale for that object is $l_i = \sqrt{A_i}$.

Mean object size is defined as $\frac{1}{N_o}\sum_i l_i$

Max object size is defined as $\max l_i$.

Mean perimeter is derived by extracting the perimeter of each object P_i and defining the mean perimeter $\overline{P} = \frac{1}{N_o}\sum_i P_i$.

The *Simple Convective Aggregation Index (SCAI)* (Tobin et al., 2012) is defined as:

$$\mathrm{SCAI} = \frac{N_o D_0}{N_{max}} \tag{2}$$

Where N_{max} is the number of pixels in a scene, $D_0 = \sqrt[N_p]{\prod_i^{N_p} d_i}$ is the geometric mean of Euclidian pairwise distance between all object centroids d_i and $N_p = N_o(N_o - 1)/2$.

The Convective Organisation Potential (COP) (White et al., 2018) is:

$$\mathrm{COP} = \frac{1}{N_p}\sum_{i=0}^{N_o}\sum_{j=i+1}^{N_o}\frac{l_i + l_j}{\sqrt{\pi} d_{ij}} \tag{3}$$

Where d_{ij} now explicitly represents the distance between two object centroids.

Max RDF is the maximum value of the radial distribution function $\mathrm{RDF}(r)$ as proposed in Rasp, Selz, and Craig (2018):

$$\mathrm{RDF}(r) = \frac{1}{N_i}\sum_i \frac{\sum_{r \le r_i < r+dr} 1}{L\left(\pi\left(r + dr\right)^2 - r^2\right)} \tag{4}$$

Where r_i are pairwise distances from the i$^{\mathrm{th}}$ centroid to all other centroids, dr denotes the width of a radial annulus over which we sum such distances, L is the length of the scene's side, and N_i are the number of centroids that lie within a distance r_{max} from the domain edges. We only consider coordinates within a radius r_{max} from any original centroid. We

set $r_{max} = 20$ pixels, as in practice $\arg\max \mathrm{RDF}(r) < 20$ always, and use $dr = 1$ (the results are not sensitive to these parameters).

Degree variance of nearest-neighbour network representations of the scenes are quantified by constructing a Voronoi tessellation from the computed object centroids and measuring the variance in the degree (number of neighbours) distribution of the identified Voronoi cells.

I_{org} (Weger et al., 1992) is included in two flavours. The first is the original metric, which integrates the area under the curve defined by the NNCDF, the cumulative density function of nearest neighbour distances d_N between object centroids (y axis) and the corresponding Weibull distribution (x axis):

$$W = 1 - \exp\left(\frac{N_o}{L^2}\pi d_N^2\right) \tag{5}$$

If the object centroids are scattered as a Poisson point process, they should follow W exactly, resulting in $I_{org} = 0.5$. $I_{org} < 0.5$ if they are regularly spaced; if they appear in clusters, $I_{org} > 0.5$. As pointed out by Benner and Curry (1998), this overestimates the regularity of the cloud field, because in reality separate cloud objects are inhibited from forming within the area covered by another object. To account for this, we also include a second version of I_{org}, which we name I^*_{org}. This metric compares the cloud field NNCDF to an inhibition NNCDF, which is constructed by randomly scattering N_o objects throughout the scene, provided that they do not fall within the circular area of an object that has already been placed. The computer-generated random positions of this approach are less robust than the Weibull distribution (Weger et al., 1992), but we

find that repeating the computations 3 times does not impact the resulting I^*_{org} below the third significant digit.

Scale decomposition metrics

Size exponent b is computed by counting all cloud sizes N_c in bins of exponentially increasing width, and fitting the resulting cloud size distribution with a power law:

$$\log N_c \propto b \log l \tag{6}$$

The average coefficient of determination R^2 of fitting this relation to all scenes is good: 0.923. We also investigated a fit according to subcritical percolation theory that incorporates an exponential term. However, undersampling of large cloud structures make such fits quite unrealistic on a per-scene basis, even though the fit converges when sampling a large number of scenes at similar cloud fraction (not shown). It is therefore likely that these cloud fields obey the rules of subcritical percolation. Yet, the parameters of the corresponding fit cannot reliably be identified on a per scene basis.

The *box-counting dimension* D_f (fractal dim.) of each cloud mask field is derived by counting the number of square boxes N_c of dimension l_b that are neither fully cloud-free nor fully cloudy (i.e. boxes that contain cloud borders). D_f is then computed by least-squares fitting the following relation over a range of l_b:

$$\log N_c \propto D_f \log l_b \tag{7}$$

The average R^2 of this fit is 0.997, indicating an excellent goodness of fit.

The Spectral Length Scale (Spectral length) Λ is derived from the field's Fourier transform. Computing this value requires several design choices. First, the scenes are tilt-compensated by subtracting a scene's best-fit plane. Next, one would normally apply a radially symmetric window function to account for the scenes' aperiodicity. However, we find that the application of such a function occludes so much spatial information that our scenes are ordered much less coherently. Hence, we refrain from applying window functions. Next, we Fourier transform the scenes and construct their 1D PSD $S(k)$ by averaging the transform's power signals over shells of radial wavenumber k. The validity of this approach rests on the assumption that the satellite scenes are spatially isotropic, which they are often not. Yet, we find that on a scale from 0-1 (0 representing a 2D PSD where the power is equally distributed over the azimuthal direction and 1 representing the case where all power is concentrated in a single direction), the average anisotropy of all scenes is 0.104. We judge that this justifies the use of the 1D PSD. Finally, Λ is computed from the distribution's first moment, as suggested in Jonker, Duynkerke, and Cuijpers (1999):

$$\Lambda^{-a} = \frac{\int_0^{k_{Ny}} k^a S(k) dk}{\int_0^{k_{Ny}} S(k) dk}; \quad a \neq 0 \tag{8}$$

Where k_{Ny} is the Nyquist wavenumber and we choose to set $a = 1$.

We compute *Wavelet-based Organisation Indices (WOIs)* following Brune, Kapp, and Friederichs (2018). These metrics are based on the domain-averaged, squared coefficients of the 2D stationary wavelet transform (SWT) of each scene's cloud water path (CWP) field, E_{CWP}. We use the Haar wavelet as our basis. E_{CWP} contains a scale decomposition

over three (horizontal, vertical, diagonal) directions, with each scale representing a power of 2 that exactly fits the 512 pixel field. Using E_{CWP}, we derive the metrics proposed by (Brune et al., 2018):

$$WOI_1 = \frac{\overline{E^l_{CWP}}}{\overline{E_{CWP}}} \tag{9}$$

$$WOI_2 = \frac{\overline{E_{CWP}}}{N_c} \tag{10}$$

$$WOI_3 = \frac{1}{3}\sqrt{\sum_d \left(\frac{E^l_{CWP_d} - \overline{E^l_{CWP}}}{\overline{E^l_{CWP}}}\right)^2 + \left(\frac{E^s_{CWP_d} - \overline{E^s_{CWP}}}{\overline{E^s_{CWP}}}\right)^2} \tag{11}$$

Where $\cdot^l$ and $\cdot^s$ indicate total energy contained in the large scales (resolution $2^1 - 2^5$) and small scales (resolution $2^6 - 2^9$) respectively, $\bar{\cdot}$ indicates averaging over all three directions and N_c is the number of cloudy pixels in a scene. These metrics measure the fraction of cloud water contained in the scene's large scales (WOI_1), the average cloud water in cloudy pixels (WOI_2) and the anisotropy in the spectrum's three directions (WOI_3). Since WOI_1 and WOI_2 are almost exact mirrors of R (eq. 1) and cloud water variance in cloudy pixels respectively, respectively, we choose to only include WOI_3 in our analysis.

Our simple *Clear Sky* metric extracts the scene's largest rectangular area spanned by the horizontal and vertical lines drawn through any cloud-free pixel whose ends are the first cloudy pixel encountered along those lines. This rectangle is normalised by the domain size, to arrive at a fraction that represents the largest, contiguous, clear sky area.

References

Benner, T. C., & Curry, J. A. (1998). Characteristics of small tropical cumulus clouds and their impact on the environment. *Journal of Geophysical Research: Atmospheres*,

103(D22), 28753–28767.

Bretherton, C., & Blossey, P. (2017). Understanding mesoscale aggregation of shallow cumulus convection using large-eddy simulation. *Journal of Advances in Modeling Earth Systems*, *9*(8), 2798–2821.

Brueck, M., Hohenegger, C., & Stevens, B. (2020). Mesoscale marine tropical precipitation varies independently from the spatial arrangement of its convective cells. *Quarterly Journal of the Royal Meteorological Society*, *146*(728), 1391–1402.

Brune, S., Kapp, F., & Friederichs, P. (2018). A wavelet-based analysis of convective organization in ICON large-eddy simulations. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2812–2829.

Denby, L. (2020). Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophysical Research Letters*, *47*(1), e2019GL085190.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, *80*(2), 977–1002.

Glassmeier, F., & Feingold, G. (2017). Network approach to patterns in stratocumulus clouds. *Proceedings of the National Academy of Sciences*, *114*(40), 10578–10583.

Jonker, H. J., Duynkerke, P. G., & Cuijpers, J. W. (1999). Mesoscale fluctuations in scalars generated by boundary layer convection. *Journal of the atmospheric sciences*, *56*(5), 801–808.

Neggers, R., Griewank, P., & Heus, T. (2019). Power-law scaling in the internal variability of cumulus cloud size distributions due to subsampling and spatial organization.

Journal of the Atmospheric Sciences, *76*(6), 1489–1503.

Rasp, S., Selz, T., & Craig, G. C. (2018). Variability and clustering of midlatitude summertime convection: Testing the Craig and Cohen theory in a convection-permitting ensemble with stochastic boundary layer perturbations. *Journal of the Atmospheric Sciences*, *75*(2), 691–706.

Tobin, I., Bony, S., & Roca, R. (2012). Observational evidence for relationships between the degree of aggregation of deep convection, water vapor, surface fluxes, and radiation. *Journal of Climate*, *25*(20), 6885–6904.

van Laar, T. W. (2019). *Spatial patterns in shallow cumulus cloud populations over a heterogeneous surface* (Doctoral dissertation, University of Cologne). Retrieved from `http://kups.ub.uni-koeln.de/id/eprint/10221`

Weger, R., Lee, J., Zhu, T., & Welch, R. (1992). Clustering, randomness and regularity in cloud fields: 1. theoretical considerations. *Journal of Geophysical Research: Atmospheres*, *97*(D18), 20519–20536.

White, B., Buchanan, A., Birch, C., Stier, P., & Pearson, K. (2018). Quantifying the effects of horizontal grid length and parameterized convection on the degree of convective organization using a metric of the potential for convective interaction. *Journal of the Atmospheric Sciences*, *75*(2), 425–450.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*(2), 265–286.

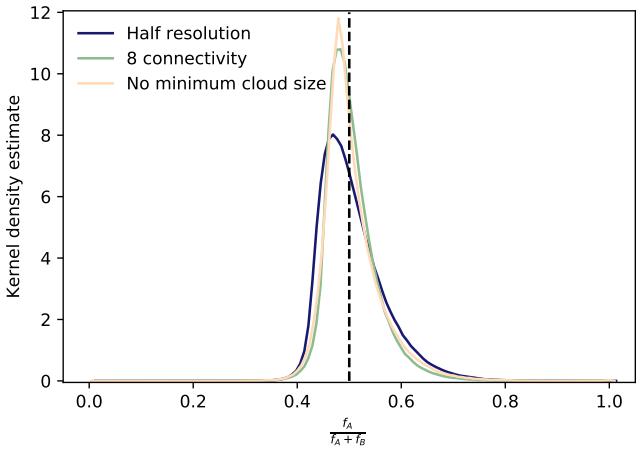

Figure S1. Gaussian kernel density estimates of the ratio $D = \frac{f_A}{f_A - f_B}$, which is constructed from high-dimensional kernel density estimates of the reference metric distribution used in the main text (f_A) and three separately perturbed metric distributions (f_B). An identical distribution to the original would yield a Dirac pulse centred at 0.5 (dashed line); deviations from this line quantify the contrast between the original and perturbed distributions. Sensitivities are quantified with respect to i) scenes that are downsampled to half the original resolution (most sensitive), ii) object segmentation based on 8-connectivity rather than 4-connectivity and iii) not including a lower bound to the minimum cloud size that is considered an object (least sensitive). All perturbed distributions are narrow and have an expected value around 0.5, indicating the robustness of the distribution presented in the main text. Furthermore, the visual relation between metrics is largely unaffected (not shown).

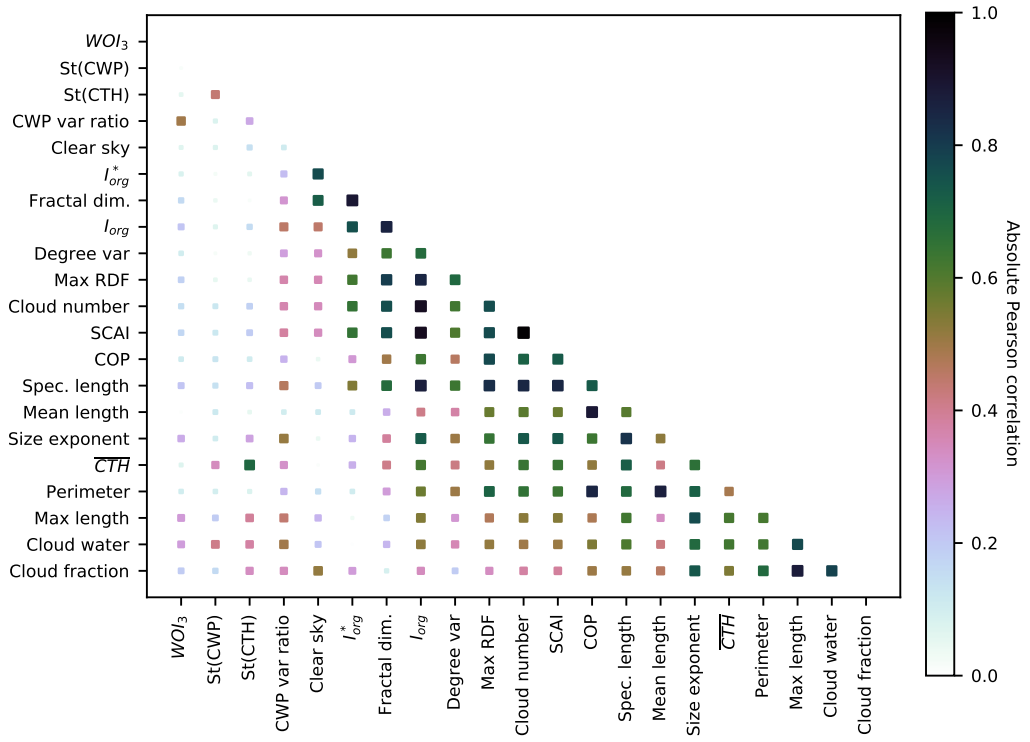

Figure S2. Standardised metric correlation matrix, with squares sized and coloured according to absolute Pearson correlation between a metric pair. Many metrics closely correlate, indicating that their cumulative information can be captured by a smaller number of effective indicators. Several closely correlating metrics follow well-known relationships, e.g. perimeter and mean length (any combination yields approximately constant fractal dim.), or cloud number and numerous aggregation metrics (this relation is similar for deep convective organisation (Brueck et al., 2020)). Others follow rather trivial ones, e.g. max length and cloud fraction, or the Spectral Length and size exponent. Several strong correlations are at first sight not trivial. For instance, I_{org} (both versions) and Fractal dim. are highly similar (up to a factor -1). Hence, highly concentrated shallow cloud clusters in rather empty scenes (high I^*_{org}) tend towards "lines" (low fractal dimension, approaching 1 from above); $I^*_{org} = 0.5$ and fractal dim.=2 both indicate random scattering of points. Finally, while some effort has been invested in contrasting and improving aggregation/clustering measures (e.g. SCAI, I_{org} and max RDF (van Laar, 2019)), these are extremely similar. Instead, shifting focus to metrics that are comparatively *uncorrelated* might be more more fruitful to further develop our understanding of shallow cloud field organisation.

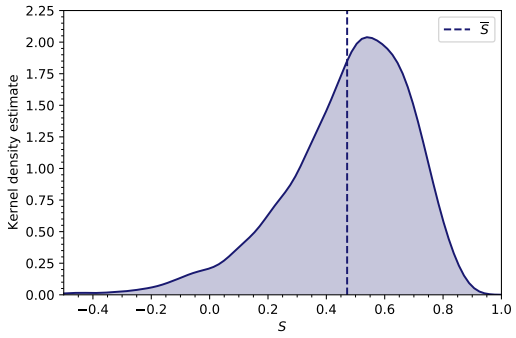

Figure S3. Gaussian kernel density estimate of $S = 1 - \frac{\|x_{a_i} - x_{n_i}\|_2}{\|x_{a_i} - x_{r_i}\|_2}$ compiled from $0 \leq i < 3951$ scenes, where x_{a_i} is the vector of the metrics for an "anchor scene", x_{n_i} are the metrics of a "neighbour scene" that overlaps with half the area of the anchor scene and x_{r_i} are the average metrics of 100 randomly sampled scenes. S measures how much the metrics minimise the Euclidian distance between an anchor and its half-overlapping scene, relative to the average Euclidian distance to a randomly sampled scene. If $S = 0$, the metrics estimate that a half-overlapping scene is equally similarly organised as a randomly sampled scene; if $S = 1$, the anchor and half-overlapping scene are estimated to be identically organised. Since half-overlapping scenes share numerous spatial features, they should usually be more similarly organised than random scenes ($S > 0$) - a feature we expect the metrics to capture. As 96% of the distribution exceeds $S = 0$, this inspires confidence in this ability. The dashed line indicates the mean, $\overline{S} = 0.47$. While this lies significantly below 1, we expect the desired upper bound of S to also lie below 1, since half-overlapping scenes are (by visual inspection) rarely *identically* organised. Estimating this bound requires knowing how far a typical pattern extends beyond a scene's boundaries; this demands a better characterisation of the relation between the measurement scale ("scene") and the true scale of a pattern. However, even without an explicit upper bound on $S < 1$, this distribution shows that our metrics on average come closer to that bound than to being random. Proficiency of a cloud field description can also be assessed by comparing S across approaches. A version of S already served as cost function for a machine-learned pattern description (Denby, 2020). One could also compile statistics on how similar humans find half-overlapping scenes compared to random scene pairs. Comparing both resulting S to our metrics could more objectively assess which approach to pattern description (human, metrics or machine) is best.

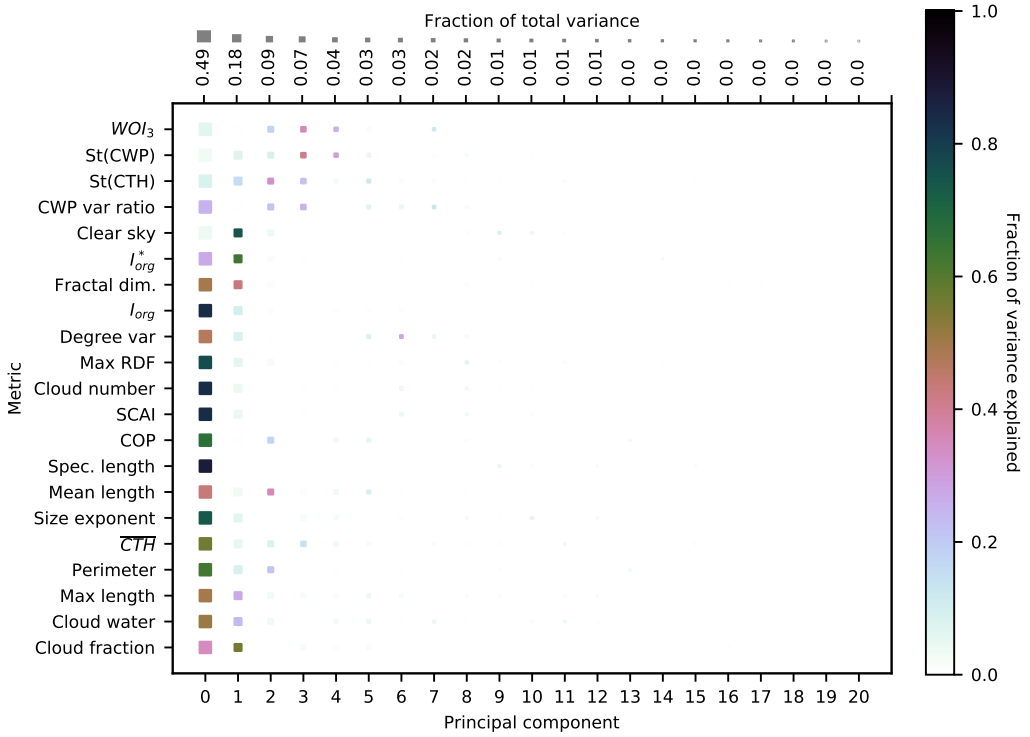

Figure S4. Fraction of variance (colour) in each metric (vertical axis) explained by each PC (horizontal axis). Sizes of squares are scaled by the total dataset's explained variance fraction in each PC (top horizontal axis). 17/21 metrics have more than 70% of their variance captured by the first two PCs; the remaining 4/21 metrics reach this threshold after four PCs.

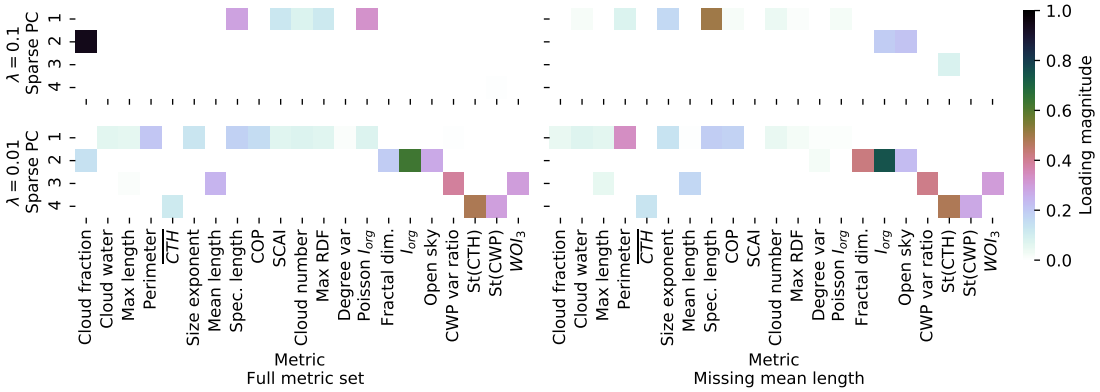

Figure S5. Sensitivity of Sparse Principal Component Analysis (SPCA, Zou et al., 2006). SPCA encourages sparsity in the weighting of metrics that form each of the four main, approximate PCs ("loadings") by casting the PCA as a regression problem, whose cost function contains at least i) a least squares error term of the PCA fit and ii) a penalty (in the L_0 or L_1 norm) on the magnitude of the regression coefficients (the loadings). This penalty is weighted by a regularisation parameter λ. We solve the resulting non-convex optimisation problem using the approach developed by Erichson et al. (2020) and refer to that paper for further details. This figure shows the optimal sparsity structure in the loadings identified by SPCA under four combinations of two free parameters: The magnitude of the sparsity penalty λ (top row vs bottom row) and the omission of a single, seemingly redundant, metric (SCAI, left column vs right column). Unfortunately, the optimal sparsity structure i) is rather sensitive and ii) reacts relatively unpredictably to changes in these free parameters. This is true also when other metrics are excluded, when a different sparsity-inducing algorithm is used or when the sparsity penalty is in the L_0 norm, rather than the L_1 norm as displayed here. These considerations curb SPCA's utility for metric selection and prevent us from recommending its use.

Table S1. Metrics quantified for initial analysis. Selection for paper is guaranteed by meeting either criteria 1 or 2, and separately meeting criterion 3, as presented in section 2.2. This excludes the lower portion of the table. Two metrics in the table's middle section meet the criteria, but are still excluded: WOI_1, WOI_2 (see Text S1). Metrics annotated with (*) are not included in the coded library.

Metric	**Criterion 1** *Unique*	**Criterion 2** *Recurrent/recent*	**Criterion 3** *Interpretable*
Cloud fraction	No	Yes	Yes
Cloud water	Yes	Yes	Yes
Max length	No	Yes	Yes
Perimeter	No	Yes	Yes
$\overline{CTH}$	Yes	Yes	Yes
Size exponent	Yes	Yes	Yes
Mean length	No	Yes	Yes
Spectral length scale	No	Yes	Yes
COP	No	Yes	Yes
SCAI	No	Yes	Yes
Cloud number	No	Yes	Yes
Max RDF	No	Yes	Yes
Degree var.	Yes	Yes	Yes
I_{org}	No	Yes	Yes
Fractal dimension	Yes	Yes	Yes
I^*_{org}	Yes	No	Yes
Open sky	Yes	No	Yes
CWP var ratio	Yes	Yes	Yes
St(CTH)	Yes	Yes	Yes
St(CWP)	Yes	Yes	Yes
WOI_3	Yes	Yes	Yes
WOI_1	No	Yes	Yes
WOI_2	No	Yes	Yes
Multifractality index (*)	Yes	Yes	No
Multifractal intermittency (*)	Yes	Yes	No
Object eccentricity	No	No	Yes
Covariance-based orientation	No	No	Yes
Raw moment-based orientation	No	No	Yes
b_{org} in small clouds (Neggers et al., 2019)	Yes	Yes	No
Skewness/kurtosis of CTH, CWP	Yes	Yes	No
Geometric mean nearest neighbour distance	No	No	Yes
Variance of CTH, CWP in largest cloud	No	No	Yes
1D PSD slope	No	No	Yes
Variance in azimuthal PSD	No	No	Yes
Aboav-Wearie fit (Glassmeier & Feingold, 2017)	Yes	Yes	No