Wasserstein k-means++ for Cloud Regime Histogram Clustering

Matthew Staib and Stefanie Jegelka

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{mstaib, stefje}@mit.edu

Abstract

Much work has sought to discern the different types of cloud regimes, typically via Euclidean k-means clustering of histograms. However, these methods ignore the underlying similarity structure of cloud types. Wasserstein k-means clustering is a promising candidate for utilizing this structure during clustering, but existing algorithms do not scale well and lack the quality guarantees of the Euclidean case. We resolve this by generalizing k-means++ guarantees to the Wasserstein setting and providing a scalable minibatch algorithm for Wasserstein k-means. Our methods empirically perform well and lead to new, different cloud regime prototypes.

1 Motivation

Given the climatic importance of clouds, much recent work has focused on identifying and then analyzing the main cloud regimes [Jakob and Tselioudis, 2003; Rossow et al., 2005a,b; Williams and Tselioudis, 2007; Williams and Webb, 2009; Tselioudis et al., 2013; McDonald et al., 2016]. Once determined, these regimes are used in many settings, e.g., assessing general circulation models [Williams and Webb, 2009; Mason et al., 2015], and therefore accurately identifying these regimes is crucial to understanding the climate system.

The vast majority of work applies k-means clustering to joint histograms of cloud top pressure (PC) and optical depth (TAU) (henceforth PC-TAU histograms of "cloud types"), e.g. [Jakob and Tselioudis, 2003; Rossow et al., 2005a; Tselioudis et al., 2013]. Histograms are treated as vectors and compared via the Euclidean distance between them. This approach scales well to large datasets but ignores the latent structure of the data, in particular the similarity between different cloud types. Moreover, the clustering problem is solved via Lloyd's algorithm [Lloyd, 1982], which is empirically effective but gives no guarantees about the cluster quality.

Instead, we apply histogram clustering techniques based on Wasserstein distance [Villani, 2009], a metric between probability distributions (or histograms) that respects the underlying geometry of the space, in this case the similarity structure of cloud types. As illustrated in Figure 1, histograms with similar frequencies for similar cloud types are close in this metric, in contrast to Euclidean distance, which ignores cloud type similarity. We further 1) show that k-means++ seeding [Arthur and Vassilvitskii, 2007], which gives provably good cluster seedings in the Euclidean case, yields the same guarantee for the Wasserstein metric, 2) provide an efficient minibatch algorithm for Wasserstein k-means that scales to climate data, and 3) show histogram clustering can yield notably different cloud regimes than identified via Euclidean k-means.

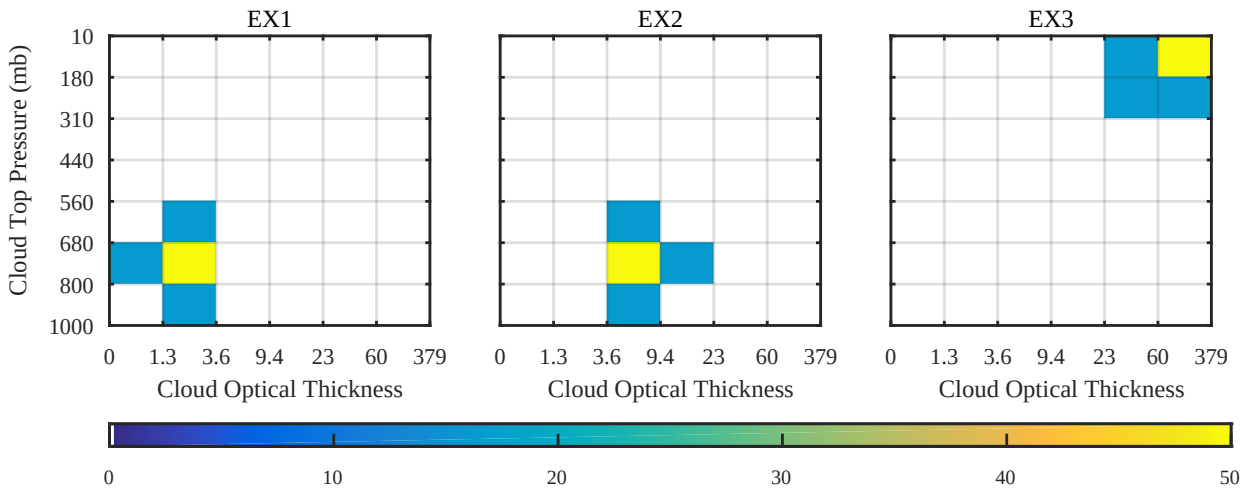

Figure 1: In Euclidean distance, EX1 is equally far from EX2 and EX3. In the specific Wasserstein distance defined in Section 4, EX3 is over 20 times farther from EX1 than EX2 is from EX1, because EX1 and EX2 are concentrated on similar PC-TAU cells.

2 Theoretical Background

Given a set of points $\{x^i\}_{i\in\mathcal{I}}$, metric k-means clustering seeks to find a set of centroids $\mathcal{C} = \{c^j\}_{j=1}^k$ in a convex set K (e.g. the probability simplex) minimizing

$$\phi(\mathcal{C}) = \sum_{i\in\mathcal{I}} \min_{j=1,\dots,k} d(x^i, c^j)^2. \tag{1}$$

Typically d is taken to be the Euclidean distance, $d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$. In this setting, Lloyd's algorithm [Lloyd, 1982], which alternates between assigning points x^i to the closest cluster centroid c^j and replacing c^j with the mean of the points assigned to it, converges to a local optimum but lacks other guarantees: in fact, finding an optimal set $\mathcal{C}$ of centroids is NP-hard [Aloise et al., 2009]. The k-means++ seeding algorithm alleviates this problem: this efficient, randomized algorithm produces an $O(\log k)$-optimal clustering in expectation [Arthur and Vassilvitskii, 2007]. This solution can then be fine-tuned by Lloyd's algorithm. This result has been extended to the case when $d(x,y)^2$ is replaced by a Bregman or total Jensen divergence [Sra et al., 2008; Nielsen and Nock, 2013]. Results for general metrics exist for other seeding algorithms, e.g. [Ahmadian et al., 2017], but these scale poorly and are hence impractical; to our knowledge, the general metric case has not yet been addressed for k-means++.

In contrast to Euclidean distance, Wasserstein distance between distributions μ and ν on points $\{y^i\}_{i=1}^n$ accounts for the "cost" C_{ij} of moving y^i to y^j. Viewing μ and ν as two piles of dirt, we can define a notion of distance between them: *how much* dirt must we move *how far* to transform one pile into the other, moving dirt as efficiently as possible? Formally, if $C_{ij} = g(y^i, y^j)^p$ for a distance metric g, the p-Wasserstein distance $W_p(\mu,\nu)$ is defined as the value of the linear program

$$\begin{array}{ll} \min & \langle C, T\rangle^{1/p} \equiv \min\left(\sum_{ij} C_{ij}T_{ij}\right)^{1/p} \\ \text{s.t.} & 1^T T = \mu,\ 1^T T^T = \nu,\ T \geq 0. \end{array} \tag{2}$$

Algorithm 1 Minibatch metric k-means

Input: point set X, parameter k
$\{c^j\}_{j=1}^k \leftarrow k$-means++Initialization$(X, k)$
$n_j \leftarrow 0$ for $j = 1, \ldots, k$ ▷ Cluster sizes
loop
  Draw $x^1, \ldots, x^m \sim X$
  $s_i \leftarrow \operatorname{argmin}_{j=1,\ldots,k} d(x^i, c^j)^2$ for $i = 1, \ldots, m$
  for $i = 1, \ldots, m$ **do**
    $j \leftarrow s_i$ ▷ Cluster index assigned for x^i
    $\gamma \leftarrow 1/n_j$
    $c_j \leftarrow \operatorname{proj}_K(c_j - \gamma \nabla_c [d(x^i, c^j)^2])$
    $n_j \leftarrow n_j + 1$
  end for
end loop

The joint distribution T is a "transport plan" that moves mass from μ to ν. A full discussion of Wasserstein distance and optimal transport is outside the scope of this paper; we refer the reader to [Villani, 2009; Santambrogio, 2015] for theoretical foundations, and [Pele and Werman, 2009; Cuturi, 2013; Genevay et al., 2016] for computing Wasserstein distance. In our clustering formulation, we use $d(x^i, c^j) = W_p(x^i, c^j)$.

Wasserstein distance has been applied to a limited extent to histogram clustering [Li and Wang, 2008; Ye et al., 2017]. The main computational challenge is computing the centroid, i.e., the Wasserstein barycenter of the measures in one cluster, in place of the Euclidean mean. Reasonably efficient barycenter algorithms exist [Cuturi and Doucet, 2014; Ye et al., 2017; Staib et al., 2017] but scaling to large datasets remains an active research area.

3 Theory and Algorithm

We sample initial cluster centroids via k-means++ seeding where we replace the Euclidean by Wasserstein distance. Then we fine-tune the seeding with a stochastic minibatch k-means algorithm suitable for large scale climate data. Our Theorem 3.1 states an approximation guarantee for our method; the seeding guarantee is proved by building on results from Nielsen and Sun [2017, Theorem 2]:

Theorem 3.1. *Suppose centroids $\mathcal{C}$ are chosen via k-means++ seeding applied to any metric d (e.g. $d = W_p$). Then the objective function $\phi(\mathcal{C})$ satisfies*

$$\mathbb{E}[\phi(\mathcal{C})] \leq 8(\ln k + 2) \min_{\mathcal{C}^*} \phi(\mathcal{C}^*). \tag{3}$$

Proof. For any metric d, by squaring the triangle inequality we have:

$$\begin{aligned} d(x, y)^2 &\leq (d(x, z) + d(z, y))^2 \\ &= d(x, z)^2 + d(z, y)^2 + 2d(x, z)d(z, y). \end{aligned}$$

By the arithmetic mean-geometric mean inequality,

$$2d(x, z)d(z, y) \leq d(x, z)^2 + d(z, y)^2.$$

Combining these, it follows that

$$d(x,y)^2 \leq 2(d(x,z)^2 + d(z,y)^2),$$

i.e. in the language of Nielsen and Sun [2017], d^2 satisfies the 2-approximate triangle inequality. The result then follows from [Nielsen and Sun, 2017, Theorem 2]. □

Once an initial seeding is selected, Lloyd's algorithm can be applied to fine-tune the clustering, and can only improve the objective value. However, updating the centroids requires expensive full passes over the dataset.

A more scalable alternative is a variant of online or minibatch gradient descent applied to Problem (1). In particular, we generalize an algorithm from Sculley [2010] to the Wasserstein case. The result is our algorithmic contribution: Algorithm 1 enjoys the guarantees of Theorem 3.1, and efficiently fine-tunes the clusters without many expensive passes over the entire dataset. In particular, for W_p distances, we can compute the required gradients $\nabla_c[d(x^i, c^j)^2]$ via linear programming and the chain rule, and project efficiently onto the simplex K [Held et al., 1974; Michelot, 1986; Duchi et al., 2008]. Note that we accomplish this without ever needing to compute a Wasserstein barycenter, in contrast to past work on histogram clustering.

4 Experiments

Experimental setup. We applied our clustering framework to PC-TAU histograms from the International Satellite Cloud Climatology Project (ISCCP) [Rossow and Dueñas, 2004]. We focused specifically on data from the tropical region within 15° of the equator as in [Rossow et al., 2005a], in 3 hour increments from 1994-2009.

Wasserstein distances depend on a "ground" distance metric g between points: we built the ground metric g by mapping the cloud top pressure and optical depth pairs to an equally-spaced grid in $\mathbb{R}^2$ and using Euclidean distance. An extra "no cloud" state is added with constant distance $0.5D$ to each other state as in [Pele and Werman, 2009], where D is the maximum distance otherwise. We ran Algorithm 1 for 20 iterations with minibatch sizes of $m = 1000$. Gradients $\nabla_c[d(x^i, c^j)^2]$ were computed using Gurobi [Gurobi Optimization, 2016], and each outer iteration took about 10 seconds on a modern 8-core desktop computer. The initial k-means++ seeding was approximated using the algorithm from [Bachem et al., 2016], with 2000 burn-in steps.

Both Euclidean and Wasserstein-based clustering were tested. Prior work had carefully determined the number of clusters k by analyzing correlations between cluster centroids [Rossow et al., 2005a,b; Williams and Tselioudis, 2007; Tselioudis et al., 2013]. In the Euclidean case, we chose $k = 6$ to match [Rossow et al., 2005a]. In the Wasserstein case, we instead analyzed the minimum W_p distance between cluster centroids, seeking a balance between a low objective value and spread out centroids.

Results. First, we applied Algorithm 1 to the standard Euclidean setting, producing cluster centroids (weather states) as shown in Figure 2. We essentially reproduce the same weather states as in [Rossow et al., 2005a] for the same tropical region.

We then clustered with respect to W_p distance, for $p \in \{1, 2\}$. Qualitatively, $p = 2$ led to centroids that are more spread out, as W_2 induces a lower penalty for moving mass between very close points. Hence, we focus on $p = 1$ in this paper. Table 3 shows the minimum W_1 distances

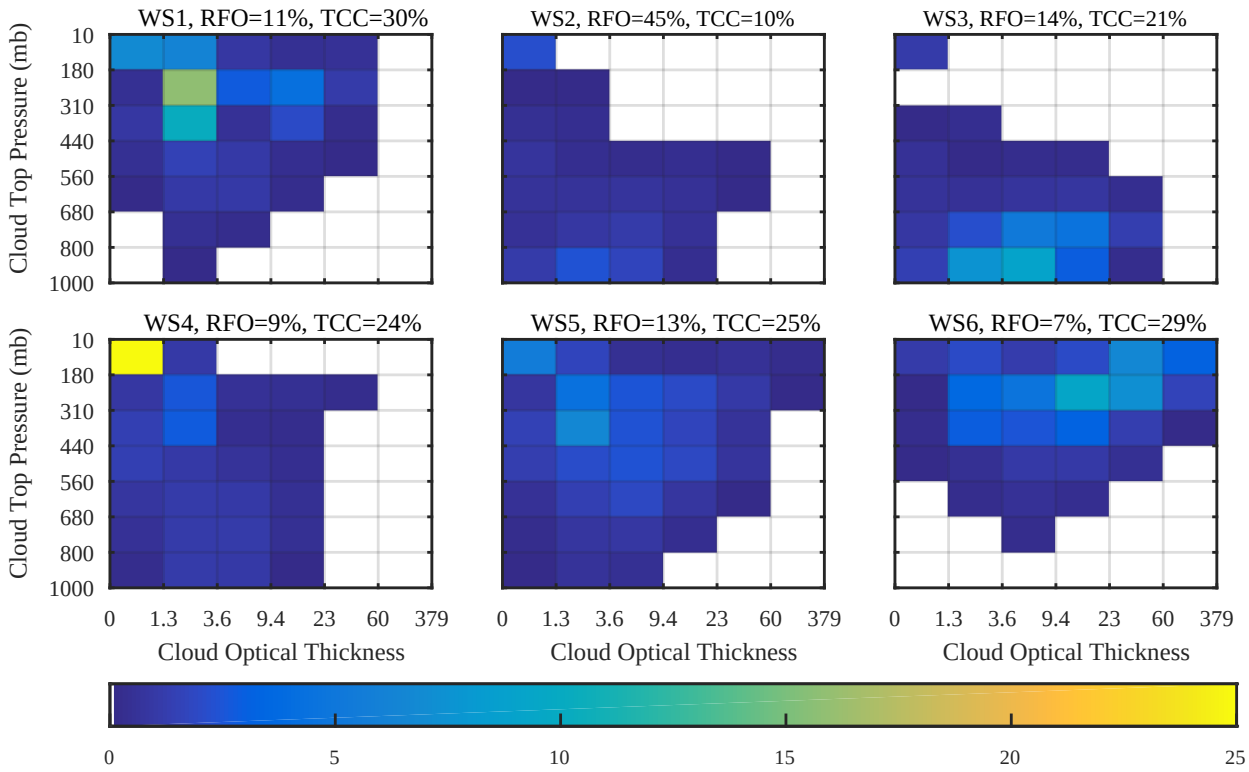

Figure 2: Weather states (cluster centroids) produced by Algorithm 1 applied to Euclidean distance. Note the similarity to those from [Rossow et al., 2005a]. RFO is relative frequency of occurence; TCC is total cloud cover.

k	4	5	6	7	8
$W_1(c^i, c^j)^2$	0.283	0.244	0.168	0.141	0.174
$\phi(\mathcal{C})/\lvert\mathcal{I}\rvert$	0.107	0.098	0.086	0.078	0.074

Figure 3: Minimum squared Wasserstein distance $W_1(c^i, c^j)^2$ between cluster centroids and the scaled k-means objective value $\phi(\mathcal{C})$, as the number of clusters k varies. Note that the nearest distance drops considerably from $k = 5$ to $k = 6$.

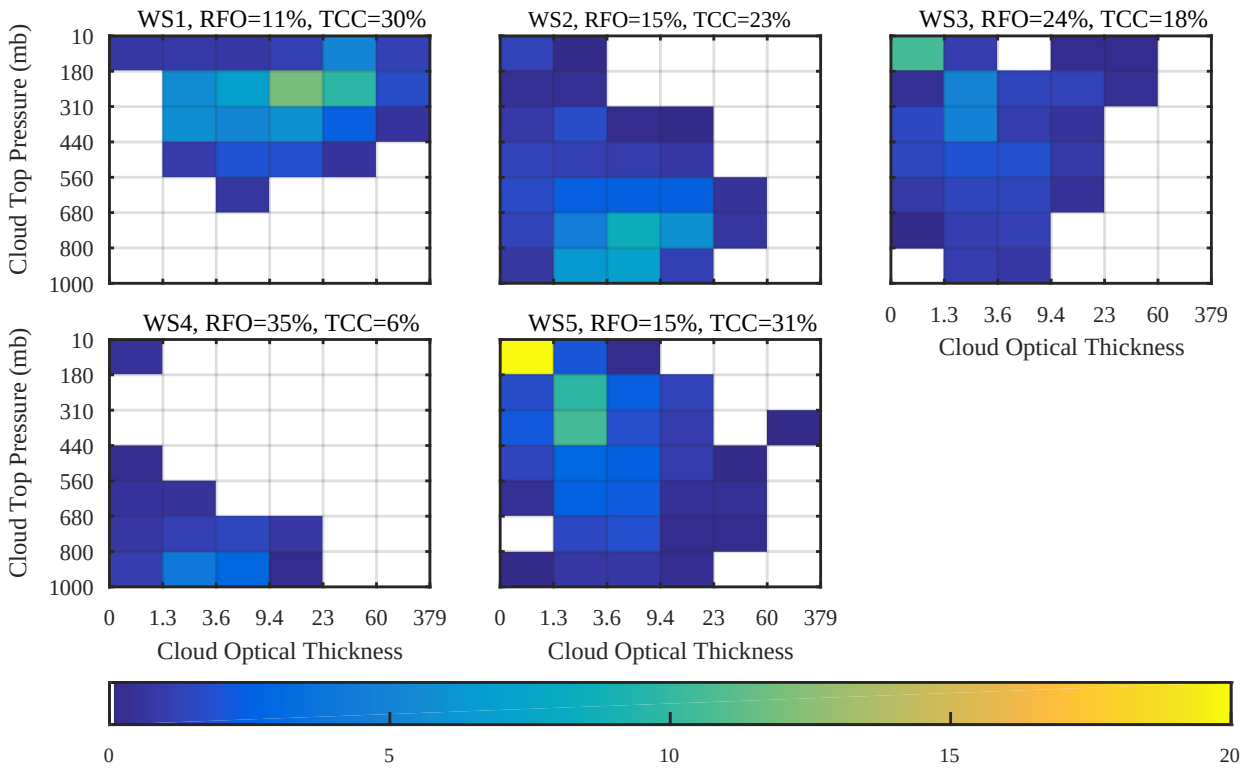

Figure 4: Weather states from Algorithm 1 applied to W_1 distance. RFO is relative frequency of occurence; TCC is total cloud cover.

between cluster centroids, together with estimates of $\phi(\mathcal{C})$. There is a notable dropoff in minimum distance after $k = 5$ without great improvement in the objective, so 5 clusters were chosen.

The resulting $k = 5$ weather states (WS) are shown in Figure 4. For each point in the tropical region, we give in Figure 5 a visual breakdown of how frequently that point belongs to each weather state (c.f. [Rossow et al., 2005a, Figure 2]). There are clear correspondences between the Euclidean-derived weather states and the Wasserstein ones. Note that Euclidean WS1, WS4, and WS5 split into Wasserstein WS3 and WS5. These Euclidean weather states are more muddled, having very similar total cloud cover and concentration (under g); in contrast, their Wasserstein counterparts have similar concentration, but notably different total cloud cover.

5 Discussion

We propose Wasserstein histogram clustering as a way to leverage prior knowledge about similarity and geometry in learning from climate datasets. We demonstrate that Wasserstein k-means++ clustering is achievable at large scale and with provable guarantees. Applying these techniques to cloud regimes yields different inferred weather states than Euclidean clustering.

For determining cloud regimes, we still need a principled way to select the ground distance metric between cloud types, perhaps via metric learning. Further in-depth analysis of these new, different weather states is needed, and of cloud regimes beyond the tropics considered here. More generally, identifying new geometry-aware clustering tasks in climate science is fertile ground for future work.

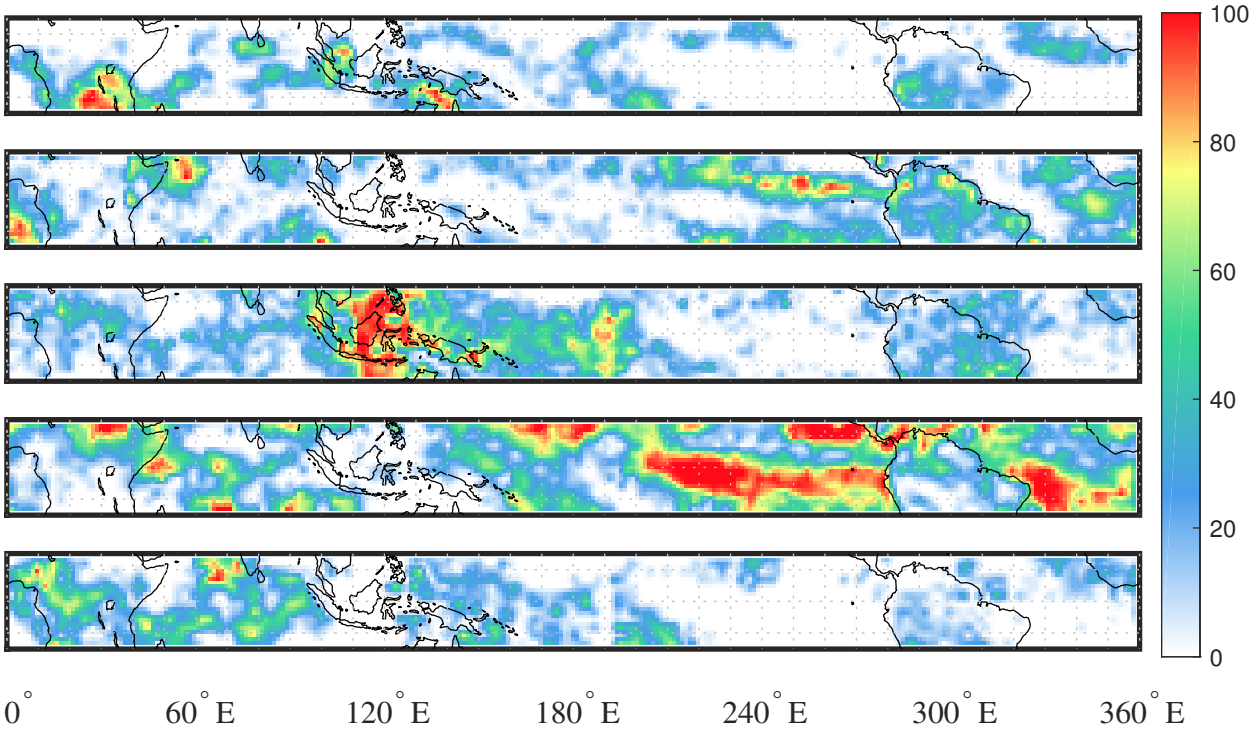

Figure 5: Heatmaps for weather states 1 (top) through 5 (bottom). On the heatmap for one weather state, each point is colored according to how often it belongs to that state.

Acknowledgments

This research was conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a, and also supported by NSF CAREER award 1553284.

References

Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better Guarantees for k-Means and Euclidean k-Median by Primal-Dual Algorithms. In *Foundations of Computer Science*, 2017.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Mach Learn*, 75(2):245–248, May 2009. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-009-5103-0.

David Arthur and Sergei Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-89871-624-5.

Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Approximate K-Means++ in Sublinear Time. In *Thirtieth AAAI Conference on Artificial Intelligence*, February 2016.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

Marco Cuturi and Arnaud Doucet. Fast Computation of Wasserstein Barycenters. In *Proceedings of The 31st International Conference on Machine Learning*, pages 685–693, 2014.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279. ACM, 2008.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

Inc. Gurobi Optimization. Gurobi Optimizer Reference Manual. 2016.

Michael Held, Philip Wolfe, and Harlan P. Crowder. Validation of subgradient optimization. *Mathematical Programming*, 6(1):62–88, December 1974. ISSN 0025-5610, 1436-4646. doi: 10.1007/BF01580223.

Christian Jakob and George Tselioudis. Objective identification of cloud regimes in the Tropical Western Pacific. *Geophys. Res. Lett.*, 30(21):2082, November 2003. ISSN 1944-8007. doi: 10.1029/2003GL018367.

Jia Li and James Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans Pattern Anal Mach Intell*, 30(6):985–1002, June 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.70847.

S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, March 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489.

Shannon Mason, Jennifer K. Fletcher, John M. Haynes, Charmaine Franklin, Alain Protat, and Christian Jakob. A Hybrid Cloud Regime Methodology Used to Evaluate Southern Ocean Cloud and Shortwave Radiation Errors in ACCESS. *J. Climate*, 28(15):6001–6018, April 2015. ISSN 0894-8755. doi: 10.1175/JCLI-D-14-00846.1.

Adrian J. McDonald, John J. Cassano, Ben Jolly, Simon Parsons, and Alex Schuddeboom. An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states. *J. Geophys. Res. Atmos.*, 121(21):2016JD025199, November 2016. ISSN 2169-8996. doi: 10.1002/2016JD025199.

C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\propto$n. *J Optim Theory Appl*, 50(1):195–200, July 1986. ISSN 0022-3239, 1573-2878. doi: 10.1007/BF00938486.

Frank Nielsen and Richard Nock. Total Jensen divergences: Definition, Properties and k-Means++ Clustering. *arXiv:1309.7109 [cs, math]*, September 2013.

Frank Nielsen and Ke Sun. Clustering in Hilbert simplex geometry. *arXiv:1704.00454 [cs]*, April 2017.

O. Pele and M. Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, September 2009. doi: 10.1109/ICCV.2009.5459199.

W. B. Rossow and E. N. Dueñas. The International Satellite Cloud Climatology Project (ISCCP) Web Site: An Online Resource for Research. *Bull. Amer. Meteor. Soc.*, 85(2):167–172, February 2004. ISSN 0003-0007. doi: 10.1175/BAMS-85-2-167.

William B. Rossow, George Tselioudis, Allyson Polak, and Christian Jakob. Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures. *Geophys. Res. Lett.*, 32(21):L21812, November 2005a. ISSN 1944-8007. doi: 10.1029/2005GL024584.

William B. Rossow, Yuanchong Zhang, and Junhong Wang. A Statistical Model of Cloud Vertical Structure Based on Reconciling Cloud Layer Amounts Inferred from Satellites and Radiosonde Humidity Profiles. *J. Climate*, 18(17):3587–3605, September 2005b. ISSN 0894-8755. doi: 10.1175/JCLI3479.1.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2.

D. Sculley. Web-scale K-means Clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772862.

Suvrit Sra, Stefanie Jegelka, and Arindam Banerjee. Approximation algorithms for Bregman clustering, co-clustering and tensor clustering. Technical report, 2008.

Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel Streaming Wasserstein Barycenters. In *Neural Information Processing Systems*, 2017.

George Tselioudis, William Rossow, Yuanchong Zhang, and Dimitra Konsta. Global Weather States and Their Properties from Passive and Active Satellite Cloud Retrievals. *J. Climate*, 26(19): 7734–7746, May 2013. ISSN 0894-8755. doi: 10.1175/JCLI-D-13-00024.1.

Cédric Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

K. D. Williams and G. Tselioudis. GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Clim Dyn*, 29(2-3):231–250, August 2007. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-007-0232-2.

K. D. Williams and M. J. Webb. A quantitative performance assessment of cloud regimes in climate models. *Clim Dyn*, 33(1):141–157, July 2009. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-008-0443-1.

J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support. *IEEE Trans. Signal Process.*, 65(9):2317–2332, May 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2659647.