Identifying Similar Thunderstorm Sequences for Airline Decision Support Using Optimal Transport Theory

Binshuai Wang
George Washington University, Washington, D.C, 20052, USA

James Pinto
National Center for Atmospheric Research, Boulder, Colorado, 80305, USA

Peng Wei
George Washington University, Washington, D.C, 20052, USA

We propose a new method to identify similar thunderstorm spatial-temporal sequences to support airline operations based on the optimal transport theory. Different from existing geometric methods, which often suffer from over-approximation of the covering geometric objects, our method models each thunderstorm as a probability distribution supported by the observed weather data. The core of our approach lies in measuring the similarity between thunderstorm sequences through the Wasserstein distance of their respective probability distributions. By setting different time weights and filter functions, this method can also incorporate the temporal features of the thunderstorms and consider the weather impact on key airspace/airport infrastructures. Furthermore, we apply a clustering algorithm within the probability distribution space of thunderstorms to categorize common patterns of archived thunderstorms in a given airspace region. We illustrate the effectiveness of this new method with our results with real-world weather data in the Dallas Fort Worth airspace.

I. Introduction

Understanding the impact of convective weather on airline operations remains a complex challenge. Thunderstorms moving near major hub airports frequently result in departure and arrival delays, flight cancellations, diversions, and airborne holdings. Although several meteorological models exist for forecasting short-term thunderstorm movements, it is crucial for airline operators to identify similar thunderstorms to better comprehend their impacts across different categories. This identification also offers operational references for dealing with similar historical thunderstorm events.

Several methods have been proposed for identifying/clustering similar thunderstorms in the literature. One common approach involves using convex polygons or ellipses to represent the thunderstorm-affected areas [1–4], with similarities defined by shapes' resemblances. However, these shape-based models often suffer from over-approximation and inconsistencies at different scales. Another prevalent method is the fractions skill score (FSS) [5], which involves dividing the space into a grid, assigning density values to each grid based on the number of thunderstorm points relative to the size of a local area, and then calculating similarity through density function differences. However, this approach is only doing well in snapshots, and can not be extended to thunderstorm sequences. Our approach, while sharing similarities with the FSS method, offers a distinct and innovative perspective on this challenge.

Distinct from the shape-based models prevalent in existing literature, our approach employs probability distribution-based models, initially proposed in the machine learning community [6, 7] for applications, such as image retrieval [7], face recognition [6], and hand-gesture recognition [8]. Compared to shape-based models, probability distribution-based models enjoy a strong capability to represent the data. It treats each dataset as a probability distribution and defines dataset similarity based on the similarity of these distributions. Common measurements for assessing probability distribution similarity include the Kullback–Leibler (KL) divergence, Euclidean distance, and the Wasserstein distance. While KL divergence is often used to measure disorders, the Wasserstein distance is preferable for quantifying differences in terms of energy.

We opt for the Wasserstein distance for two key reasons. Firstly, it is versatile enough to encompass both continuous and discrete distributions, unlike other measurements like KL divergence, which can falter when comparing a discrete distribution with a continuous one. Secondly, the Wasserstein distance satisfies critical metric properties like indiscernibility, non-negativity, symmetry, and triangle inequality, which the KL divergence does not satisfy. These make the Wasserstein distance a more reliable metric for our purpose of measuring similarity in probability distributions.

In our probability distribution-based model, each thunderstorm radar image (snapshot) or video (sequence) is treated as a comprehensive probability distribution, representing all the thunderstorm points within. The similarity between two thunderstorm sequences is quantified using the Wasserstein distance between them. To enhance the flexibility and applicability of this approach, we have integrated two key techniques:

Temporal Dimension Integration: by incorporating a time axis into the probability distribution space, each thunderstorm point is endowed with temporal information. This addition allows for a more holistic capture of temporal features, such as the direction of movement and duration of the thunderstorms, integrating these aspects seamlessly into the probability distributions.

Filter Functions: to tailor our model to various airline operational needs, we incorporate filter functions. These functions enable customization based on different operational considerations, integrating essential flight operation features, such as merging fixes, and airport runway orientations and gates. As a result, the distribution can be variably weighted to reflect airline operations in specific geolocations and time frames.

Apart from identifying similar thunderstorms, we also explore clustering thunderstorms with respect to the probability distribution space to understand the thunderstorm's patterns and categories. While numerous clustering algorithms exist, such as K-means, Gaussian Mixture Model (GMM), density-based spatial clustering of applications with noise (DBSCAN) [9], The Balance Iterative Reducing and Clustering using Hierarchies (BIRCH) [10], and Ordering Points To Identify the Clustering Structure (OPTICS) [11], we specifically select the OPTICS algorithm for clustering thunderstorms, based on two pivotal considerations: (1) The probability distribution space, structured with the Wasserstein distance, is a metric space rather than a Euclidean one. This distinction means certain properties, like means and inner products, are not well-defined in this space, making some algorithms, like K-means and GMM, fail for clustering; (2) The OPTICS algorithm can provide an interpretable visualization of high-dimensional data for decision-making, whereas the other algorithms can not provide such information.

The structure of our paper is organized as follows. Section 2 discusses related work and the background of the optimal transport theory and the Wasserstein distance. In section 3, we propose our method to identify similar thunderstorm sequences and two techniques, integrating temporal coordinates and filter functions. Meanwhile, we also show the computational complexity of the proposed algorithm and how to use the OPTICS algorithm to cluster the distributions. In section 4, we compare our method with the FSS method. Finally, in section 5, we show experimental results with snapshots and sequences comparison and clustering results based on the OPTICS algorithm.

II. Related Work and Background

In this section, we present an overview of existing methodologies for identifying analogous weather conditions and provide a comprehensive background on the optimal transport theory. This exploration includes a critical analysis of previous approaches, highlighting their strengths and limitations in the context of weather pattern identification. Additionally, we delve into the foundational concepts of optimal transport theory.

A. Shape-based Methods to Identify Similar Patterns

Shape-based models are a prevalent choice for measuring the similarity of thunderstorm snapshots, as seen in various studies [2–4, 12]. These models primarily focus on the geometrical shapes formed by point clouds in thunderstorm data, such as triangles, circles, polygons, squares, or ellipses. The similarity between these point clouds is determined based on the resemblance of their shapes. The key advantages of shape-based models lie in their computational efficiency and intuitive interpretability.

However, these models encounter significant challenges due to the typically unstructured and complex patterns of thunderstorm points. This complexity makes it difficult for shape-based methods to provide consistent results across different scenarios. Moreover, they exhibit a heightened sensitivity to noise, particularly when the point clouds representing the data points are sparse.

B. Fractions Skill Score

The Fractions Skill Score (FSS) assesses the agreement between a forecast and an observed field by examining the fraction of grid points exceeding a specific threshold within a neighborhood around each point [13, 14]. A perfect forecast achieves an FSS of 1, indicating complete agreement, while a "no skill" forecast has a score of 0. The advantage of this approach is to evaluate the skill of spatial forecasts, especially when dealing with high-resolution models that provide detailed but potentially misplaced information.

C. Optimal Transport Theory and the Wasserstein Distance

In the space of probability distributions, consider two distributions, $\mu(x)$ and $\nu(y)$, in the same n dimension space. The core of the optimal transport problem lies in determining the most energy-efficient transportation plan to transform one distribution into the other, in accordance with a predefined cost function $c(x, y)$. This framework is versatile, accommodating a range of distribution types including continuous, discrete, and semi-continuous.

The cost function $c(x, y)$ can also extended to any non-negative function. However, a common choice for this function is $|x - y|^p$, which represents the distance as induced p norm (with $p \geq 1$) [15]. Although demonstrating the existence of a minimal transport cost for any given cost function can be challenging, it has been established that when the cost function is defined as $|x - y|^p$, the optimal transport problem is well-defined [15, 16]. Under these conditions, the existence of a minimal transport cost is guaranteed. This ensures the feasibility and applicability of the optimal transport theory in scenarios where such a cost function is employed.

Based on the existence of minimal transport cost, We can define a metric for any two probability distributions, which is referred as $p-$Wasserstein distance.

Definition $p-$Wasserstein distance [16].

$$\begin{aligned} W_p(\mu, \nu) &:= \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n} |x - y|^p d\gamma(x, y) \right)^{\frac{1}{p}} \\ \text{s.t.} &\int_{\mathbb{R}^n} \gamma(x, y) dy = \mu(x), \forall x \in \mathbb{R}^n, \\ &\int_{\mathbb{R}^n} \gamma(x, y) dx = \nu(y), \forall y \in \mathbb{R}^n, \end{aligned} \tag{1}$$

where $\Gamma(\mu, \nu)$ is the set of all possible couplings of μ and ν. A coupling, denoted as $\gamma(x, y)$ is a joint probability measure function on $\mathbb{R}^n \times \mathbb{R}^n$, with μ and ν serving as its marginals on the respective factors. These constraints ensure the conservation of probability density or mass.

The true value of the Wasserstein distance is tantamount to the minimal transport cost between any two given distributions. It is established that the Wasserstein distance qualifies as a true metric, adhering to the properties of indiscernibility, non-negativity, symmetry, and the triangle inequality [16]. As such, it serves as an essential and foundational concept in the study of distributional distances, providing a nuanced measure of how distant two given distributions are from each other.

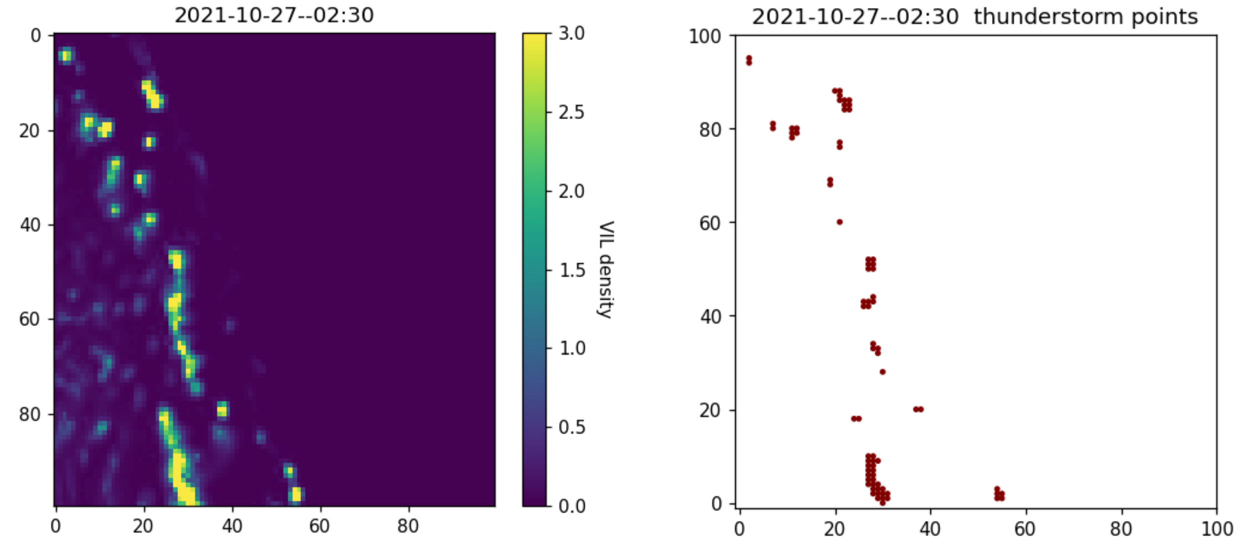
III. Methodology

A. The Probability Distribution Representation of Thunderstorms

In this section, we demonstrate the application of the Wasserstein distance in identifying similar thunderstorms using real-world weather radar data. We first describe the probability distribution representation of thunderstorms, supported by intensity data points. The intensity of thunderstorms can be estimated by the Vertically Integrated Liquid (VIL) density [17]. The VIL density is calculated by the following formula:

$$\text{VIL density} = \frac{\text{VIL}(kg/m^2)}{\text{Echo top}(m)} \times 1000, \tag{2}$$

where VIL represents the total mass of precipitation in the clouds, derived from the reflectivity readings captured by weather radar. The Echo top denotes the cloud's height at a local area, as detected by radar. Studies have shown that the VIL density has a strong positive correlation with the size of hail and the intensity of thunderstorms [17]. Specifically, a VIL density exceeding $3kg \cdot m^{-3}$ typically indicates a high probability of thunderstorm occurrence in that region [17]. While VIL density is our primary parameter, reflectivity is another viable measurement for assessing thunderstorm intensity, with a threshold value of 35 dBZ. For the sake of simplicity and clarity in our ensuing analysis, we will focus on using VIL density as our main indicator.



(a) The VIL density of a thunderstorm. The lightness indicates the intensity of the thunderstorm.

(b) The binarized thunderstorm points of the thunderstorm.

Fig. 1 VIL density value and thunderstorm points of a thunderstorm at 02:30-10/27/2021.

To facilitate efficient data storage and computation, we employ a binarization process for the thunderstorm data. In this approach, we only consider pixels with a Vertically Integrated Liquid (VIL) density of 3 or higher as valid indicators of thunderstorm activity. These selected pixels, representing valid thunderstorm points, are then used as the supports for our probability distribution, with each point assigned an equal probability mass. For instance, if a thunderstorm contains n valid points, each point is assigned a probability mass $\frac{1}{n}$.

To illustrate this process, Figure 1 presents a comparative visualization. Subfigure (a) displays the raw data, showcasing the VIL density values across a thunderstorm. Subfigure (b), on the other hand, reveals the resulting probability distribution post-binarization.

B. Similarity of Thunderstorms Defined by the Wasserstein Distance

After modeling thunderstorms with the probability distributions, we choose the Wasserstein distances of probability distributions to define the similarity of thunderstorms. Our approach focuses on the binarized representation of thunderstorm points, which simplifies the analysis to discrete cases only. This means that, instead of dealing with the complexities of continuous distributions, our analysis is confined to the optimal transport problem in a discrete setting. In this discrete framework, the Wasserstein distance provides a meaningful and computationally feasible measure of similarity, as it quantifies the 'transportation cost' of transforming one binarized thunderstorm distribution to another.

To delve into the specifics, let μ and ν be source and target distributions, separately. Suppose μ has m_1 support points and ν has m_2 support points. The coordinates of these support points are denoted by $\{x_i\}$ for μ and $\{y_j\}$ for ν, where $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$. Correspondingly, $\{a_i\}$ and $\{b_j\}$ represent the assigned probability masses for each support point in μ and ν, respectively.

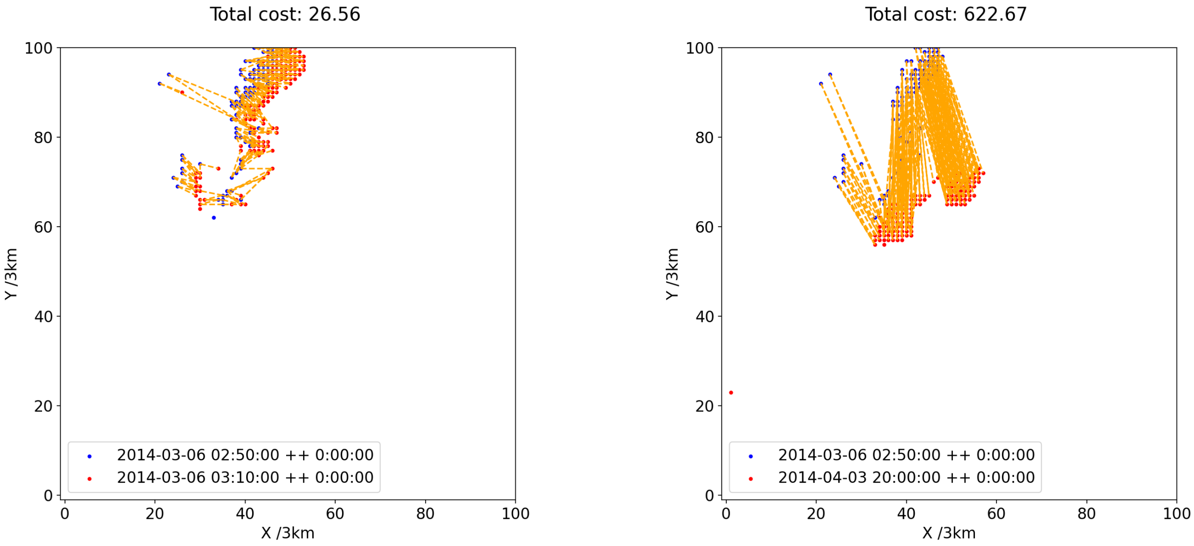
The cost of transporting mass from point x_i in μ to point y_j in ν is represented by C_{ij}, which is calculated using the cost function $c(x_i, y_j) = |x_i - y_j|^p$.

With this setup, the task of computing the Wasserstein distance between these two discrete probability distributions transforms into a linear optimization problem as follows,

$$
\begin{aligned}
W_p(\mu,\nu)^p := \min_{P_{ij}}. \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} C_{ij}P_{ij} & \\
\text{s.t.} \sum_{j=1}^{m_2} P_{ij} = a_i, i = 1, ..., m_1, & \\
\sum_{i=1}^{m_1} P_{ij} = b_j, j = 1, ..., m_2, & \\
P_{ij} \geq 0, i = 1, ..., m_1, j = 1, ..., m_2, &
\end{aligned}
\tag{3}
$$

where P_{ij}s are the decision variables, representing the probability mass transported from point x_i to point y_j. Given that our data is binarized, the probability mass a_i, $(1 \leq i \leq m_1)$, simplifies to $\frac{1}{m_1}$ and similarly, b_i, $(1 \leq i \leq m_2)$, simplifies to $\frac{1}{m_2}$. For practical computation, we primarily focus on using orders $p = 1$ and $p = 2$.

Figure 2 illustrates this concept effectively. It shows that when two thunderstorm points have similar distributions, the cost of transporting the probability mass between them is lower compared to points with more disparate distributions. This visual representation underscores the utility of the Wasserstein distance in evaluating the similarity of thunderstorm patterns, effectively capturing the cost implications of differences in the spatial distribution of the binarized fields.



(a) Wasserstein distance computed for a comparison of binarized points from two cases given by blue and red dots. A relatively smaller cost, approximately 26, is incurred in the process of transporting the source points to the target points. This lower cost indicates a higher degree of similarity between the source and target distributions

(b) Wasserstein distance computed for a comparison of binarized points from two cases given by blue and red dots. A significantly larger cost, estimated to be around 622, is necessitated for the transportation of the source points to the target points. This higher cost suggests a lower degree of similarity between the source and target distributions.

Fig. 2 The source distribution is depicted in blue, while the target distribution is represented in red. To clearly demonstrate the optimal transport mappings, we use orange dashed lines. These lines effectively illustrate the connections between corresponding points in the source and target distributions.

A simplified understanding of this linear optimization problem can be outlined as follows. The objective function aims to minimize the total energy cost, which is conceptualized as the product of the transported probability mass and the distance over which it is transported. This model operates under a few key constraints:

Inflow Balance: This set of constraints ensures that the total mass transported from a specific point equals the mass initially present at that point.

Outflow Balance: These constraints ensure that the total mass received at a given point is equivalent to the mass required at that point.

Non-Negativity: The final set of constraints confirms that the concept of "mass" remains non-negative.

Together, these constraints form the backbone of the linear optimization problem, guiding the solution towards a feasible and realistic transport plan that reflects the actual dynamics of mass transportation.

In the following, we would like to show two techniques, temporal dimension integration and filter functions, to enable a more tailored and accurate representation of probability distributions in real-world scenarios.

C. Temporal Dimension Integration

The first technique involves enriching the probability distribution with temporal information. To achieve this, we extend the spatial coordinates of each data point by appending time-dimensional coordinates. This integration ensures that the coordinates encapsulate both spatial and temporal information, offering a more comprehensive representation of the data.

Given the inherent independence of time from spatial dimensions, we introduce a time weight, denoted as w to the model. This weight is crucial in balancing the influence of spatial and temporal information within the probability distribution. By adjusting w, we can fine-tune the model to either emphasize the spatial aspects or give more weight to the temporal dynamics, depending on the specific requirements of the analysis. This flexibility allows for a more nuanced and accurate representation of phenomena where both space and time play integral roles, such as in the study of evolving weather patterns.

To illustrate the role of the time weight w in our model, suppose that s is a binarized probability distribution generated by a thunderstorm and let this distribution be represented as

$$s = \{(t_1, x_1), ..., (t_m, x_m)\} \tag{4}$$

where m is the number of thunderstorm points. Here, x_i, $(1 \leq i \leq m)$, are the spatial coordinates of valid thunderstorm points, and t_i, $(1 \leq i \leq m)$, correspond to time information of these coordinates. Assume the time weight w is a non-negative number, constrainted within the range $[0, +\infty)$.

Then, the time-weighted probability distribution can be defined as follows,

$$s^w := \{(w \cdot t_1, x_1), ..., (w \cdot t_m, x_m)\}. \tag{5}$$

The introduction of the time weight parameter offers the flexibility to adjust the temporal emphasis within the same model:

(1) when w is close to 0, $w \cdot t_i$ $(1 \leq i \leq m)$ approaches 0, effectively reducing the cost along the time axis. Consequently, the Wasserstein distance in this case predominantly reflects spatial similarities.

(2) when w goes to infinity, $w \cdot t_i$ $(1 \leq i \leq m)$ tends towards infinity, substantially increasing the cost along the time axis. Thus, in this situation, the Wasserstein distance becomes more reflective of temporal similarities.

We illustrate the impacts of different weights in the following Fig 3.

D. Filter Functions

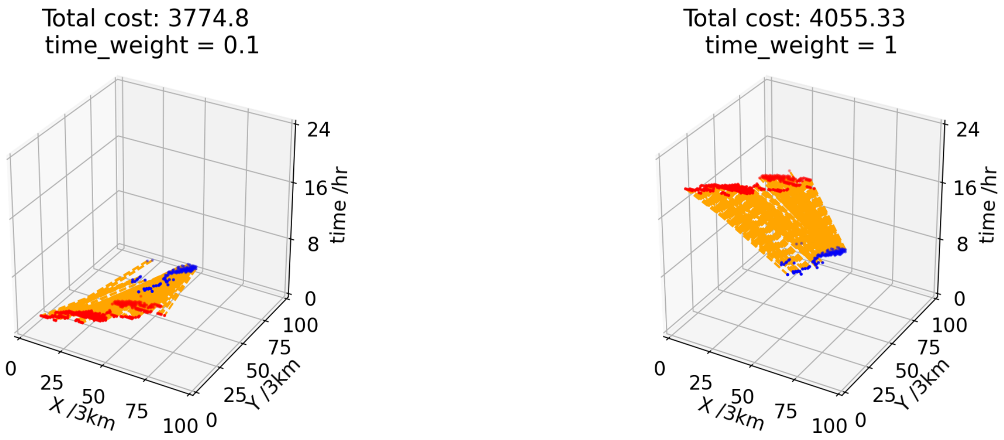
In addition to including the temporal coordinates, our approach also focuses on integrating vital airport infrastructures and flight trajectories with thunderstorm data points. This integration is facilitated through the application of filter function techniques. The essence of this method is to employ a kernel function or a weight function, tailored to assign varying weights to different points based on their significance.

To illustrate this, consider a filter function designed such that locations near airport infrastructures or flight trajectories are assigned higher positive values. This is because these areas are of greater importance due to their operational significance. Conversely, areas farther from these critical points receive lower positive values, reflecting their relative lack of immediate impact on key infrastructures or flight paths.

This filter function thus enables the creation of a customized probability distribution. This is achieved through a coordinate-wise multiplication of the original probability distribution and the filter function. For example:

Suppose that s is a discrete probability distribution generated by a thunderstorm and let f denote the filter function, defined as:

$$f : \mathbb{R}^2 \to \mathbb{R}, \tag{6}$$



(a) A small time weight makes Wasserstein distance reflect the spatial similarity.

(b) A large time weight makes Wasserstein distance reflect the temporal similarity.

Fig. 3 The change of time weight will lead to different temporal and spatial combinations. The orange line segments show the distances between the best-matched pixels. The time of source and target distribution are 02:50-03/06/2014 and 20:00-04/03/2014, respectively.

$$s = \{(t_1, x_1), ..., (t_m, x_m)\}, \tag{7}$$

where m is the number of thunderstorm points. The customized probability distribution s^f, is defined as follows:

$$s^f = \{f(x_1) \cdot (t_1, x_1), ..., f(x_m) \cdot (t_1, x_m)\}. \tag{8}$$

As demonstrated in Figure 4, the result of the coordinate-wise multiplication using the filter function is visually apparent. In this figure, the thunderstorm pixels of Figure 4 (a) in the cross area of Figure 4 (b), which represent regions of high importance such as near airport infrastructures or flight paths, retain their brightness in Figure 4 (c). Conversely, pixels in less critical areas become darker. This change in brightness across different regions aligns with the intended effect of the filter functions, emphasizing the significance of certain areas over others in the context of our analysis. This visual representation effectively illustrates how the filter function prioritizes certain regions within the probability distribution, in accordance with their operational importance.

E. Complexity of Computation

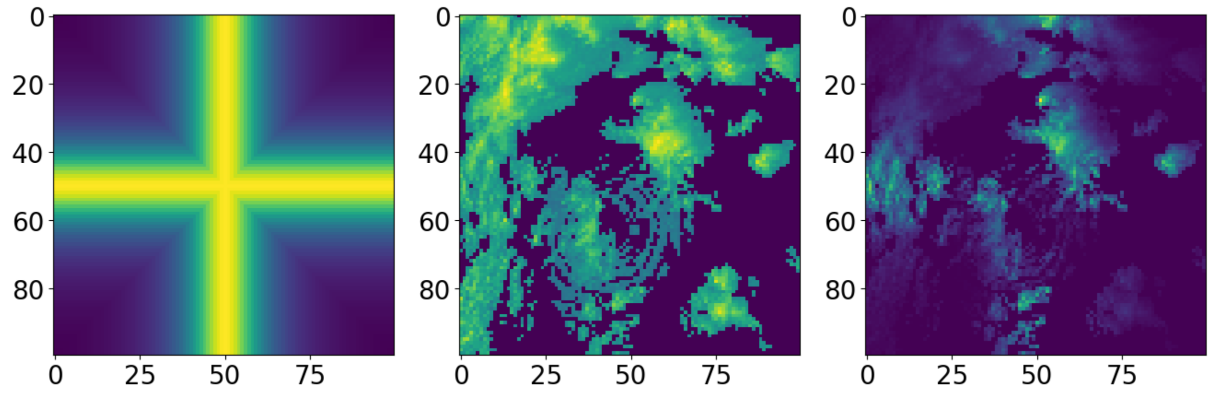
In the practical implementation of the optimal transport problem, particularly for computing the Wasserstein distance for discrete distributions, several methodologies are commonly employed. These include linear programming methods like the simplex method [15], as well as entropy regularization-based convex optimization techniques, such as Sinkhorn's algorithm [18, 19]. For our purposes, we utilize Sinkhorn's algorithm, renowned for its polynomial complexity and effectiveness as an inner-point method.

Sinkhorn's algorithm, in particular, offers advantages in terms of computational efficiency. It can be further optimized through approximation algorithms and parallel computing approaches, thereby reducing the overall computational complexity [15]. This makes it an ideal choice for handling large datasets and complex probability distributions.

The following framework shows how to identify similar thunderstorms in a historical thunderstorm dataset.

F. Clustering Thunderstorms with the OPTICS Algorithm

While probability distributions defined with the Wasserstein distance can be regarded as a metric space, they cannot be extended to a vector space due to the lack of associativity inherent in probability distributions. This characteristic precludes the use of popular vector space-based clustering algorithms like K-means or Gaussian Mixture Models (GMM) for clustering these distributions. However, with the availability of a Wasserstein distance matrix, we can effectively utilize distance-based clustering algorithms, such as OPTICS (Ordering Points To Identify the Clustering Structure)



(a) A filter function.

(b) A real weather radar image.

(c) The coordinate-wisely product of the filter function and the real weather radar image.

Fig. 4 A filter function, and a real weather radar image and their product. The filter function, when applied to the radar image, modulates the intensity of each pixel based on its location and relevance. Areas deemed of greater importance in the context of the study—such as regions near critical infrastructures or key flight paths—are highlighted by lighter pixels, reflecting higher values from the filter function. Conversely, areas of lesser significance are rendered as darker pixels, indicating lower values.

Algorithm 1: Algorithm for Identifying the Most Similar Thunderstorm based on the Wasserstein Distance.

```
Input: A given discrete probability distribution, $s_0$, and a dataset $S = \{s_i\}(1 \le i \le |S|)$, a filter function $f$ and
    a time weight $w_0$ based on airline operation consideration.
Output: An index of the most similar thunderstorm.
Based on the filter function $f$ and time weight $w_0$, generate the weighted probability distribution $s_0^g$ and a set of
  probability distributions $\{s_i^g\}(1 \le i \le |S|)$.
Initialize $d_{min}$ to a large positive number and $idx$ to 0.
for $i \leftarrow 1$ to $|S|$ do
    Compute the Wasserstein distance $d_i = W(s_i^g, s_0^g)$
    if $d_i \le d_{min}$ then
        Update the minimum distance $d_{min} = d_i$
        Update the index of the most similar thunderstorm $idx = i$
    end
end
return $idx$
```

[11]. A key advantage of OPTICS is its proficiency in identifying clusters of varying densities in real-world datasets, a task often challenging for many algorithms. Additionally, OPTICS demonstrates a lower sensitivity to input parameters compared to algorithms like DBSCAN.

The core methodology of OPTICS involves constructing a minimum spanning tree based on the Wasserstein distances between points. From this tree, a reachability plot is derived. The algorithm begins by selecting an arbitrary point as the starting point and then identifying its neighborhood points based on their Wasserstein distances. For each of these points, OPTICS calculates the core distance and the reachability distance. The algorithm then processes the point with the smallest reachability distance, subsequently updating the reachability distances of its neighbors. This procedure continues iteratively until all points in the dataset are processed.

In the resulting reachability plot, clusters manifest as valleys, marked by low reachability distances between neighboring points within the same cluster. The clear visibility of these valleys in the plot facilitates a more straightforward determination of the number of clusters, making OPTICS particularly effective for datasets where cluster densities vary significantly.

IV. Comparison with Fractions Skill Score (FSS)

One significant advantage of our proposed method, when contrasted with the traditional FSS [14] methods that rely on discretization in grids, is its reliance solely on the distances between points, eliminating the need for discretization. This characteristic becomes particularly advantageous when dealing with large ambient spaces, as our method tends to be more efficient than FSS methods under such circumstances.

Furthermore, unlike FSS methods, our approach offers an additional benefit: it provides explicit transportation plans for the points. This aspect is not just a theoretical advantage but has practical implications in terms of understanding and managing the spatial dynamics of the elements being studied. It allows for a more nuanced and operationally relevant interpretation of data, particularly in applications where the movement or transition of points is of interest.

It is also important to recognize that these are fundamentally different approaches. To illustrate the distinction, consider the following counterexample involving three thunderstorm snapshots, $\{T_1, T_2, T_3\}$, each containing a single point located at different positions $(0, 0)$, $(10, 0)$, $(20, 0)$, respectively, and all with the same mass. Assume the length of the neighborhood is h.

In the scenario where $h \leq 5$, for the FSS case, we have, $FSS(T_1, T_2) = FSS(T_1, T_3) = \sqrt{2}$, which indicates T_2 and T_3 have the same level of similarity with respect to T_1.

n contrast, using the optimal transport method with the Wasserstein distance, we obtain $W_2(T_1, T_2) = 10$ and $W_2(T_1, T_3) = 20$. This indicates T_2 is more similar than T_3 with respect to T_1.

This example demonstrates that while the FSS method is more reflective of the local neighbor's distribution, the optimal transport method provides insights into both local and global aspects. This distinction highlights the broader applicability and nuanced perspective offered by the optimal transport method in analyzing thunderstorm data.

V. Experimental Results

To validate our proposed method, we focused on a rectangular area extending 150 km from Dallas Fort Worth International Airport (DFW). Our thunderstorm image dataset was constructed using data from the High-Resolution Rapid Refresh (HRRR) database [20]. The HRRR is a high-fidelity, real-time atmospheric model that operates with a 3-km resolution with radar data assimilated every 15 minutes. This model has been consistently producing weather forecasts since 2014, offering a comprehensive and long-term dataset.

HRRR produces a variety of feature data. For our study, we selected the subhourly fields from the 2D surface level data as our primary source. Using this data, we calculated the ratio of Vertically Integrated Liquid (VIL) to echo top on a pixel-wise basis, thereby generating a VIL density image for each timestamp.

Given that thunderstorms in the DFW area predominantly occur between March and October, our data collection was restricted to this timeframe. As a result, the dataset encompasses 196,992 snapshots, providing a rich and detailed basis for our analysis and the validation of our method.

A. Thunderstorm Snapshots Comparison

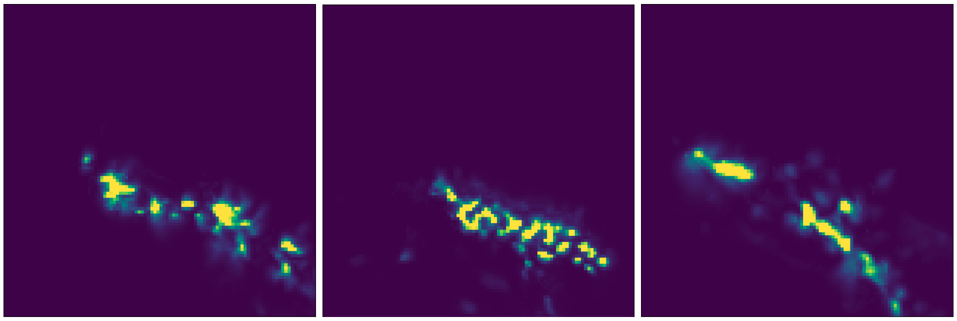
If the objective is to identify thunderstorms based solely on spatial similarity, our analysis can be confined to comparing individual snapshots. Utilizing this approach, we have identified snapshots from our dataset that are similar to a given source thunderstorm. These similar snapshots are depicted in Figure 5.

In Figure 5, subfigure (a) showcases the source thunderstorm, captured at 01:30 on 05/03/2022. Subfigures (b) and (c) represent thunderstorms that are similar to the source, captured at 06:15 on 06/12/2021 and 15:45 on 08/30/2019, respectively. The similarity of these thunderstorms is determined based on the W_2 (Wasserstein) distance, emphasizing their spatial characteristics. This comparative visualization underscores the effectiveness of our method in identifying spatially similar thunderstorms from a historical dataset.

B. Thunderstorm Sequences Comparison

If the user wants to seek both spatial and temporal similarity, then we will set w properly, and compare distributions of the sequence of snapshots. Assume the given thunderstorm evolves as indicated in Fig 6.

With setting time weight $w = 0.1$, we can find a similar thunderstorm as Fig 7. Notice that some snapshots in the sequence are not quite similar to each other, but the overall sequence and the timeline are similar to each other.

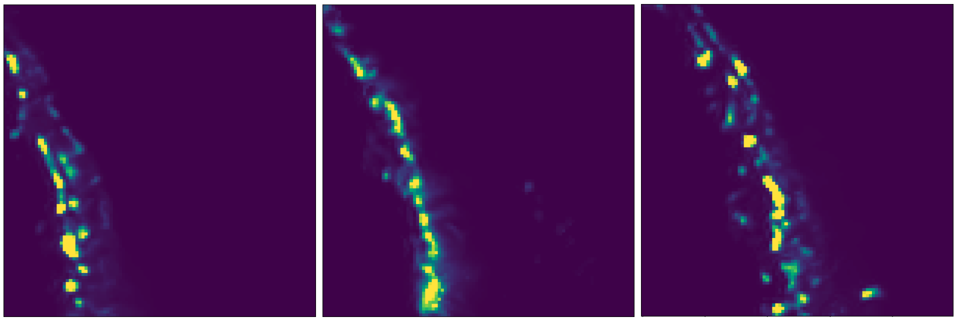


(a)

(b)

(c)

Fig. 5 Similar thunderstorm scenarios based on W_2 distance. Subfigure (a) is the source thunderstorm at 01:30-05/03/2022, and subfigures (b) and (c) are similar thunderstorms at 06:15-06/12/2021 and 15:45-08/30/2019.

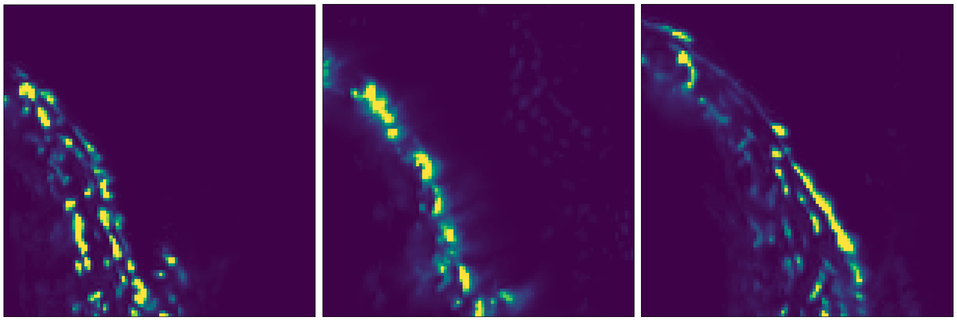


(a) 20:15

(b) 20:45

(c) 21:15

Fig. 6 Snapshots of the source thunderstorm sequence with time, which happened on 10/10/2021.



(a) 21:15

(b) 21:45

(c) 22:15

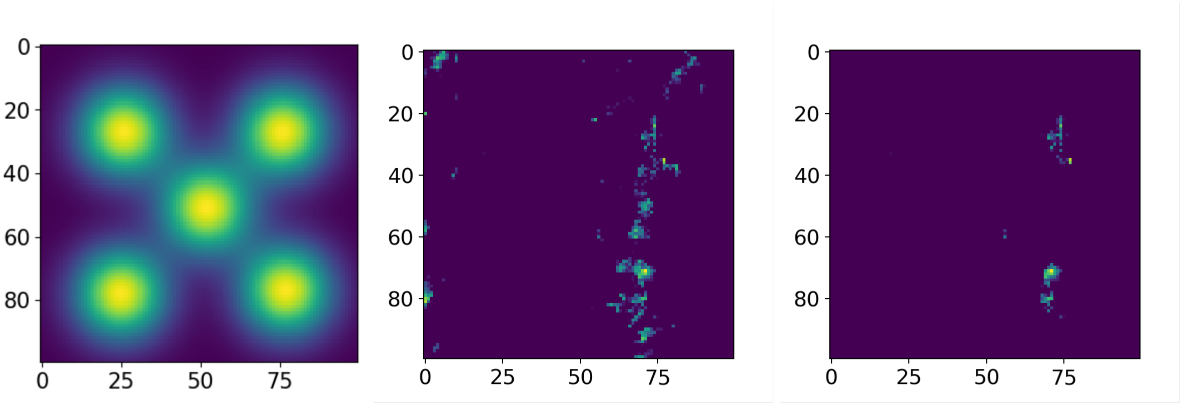
Fig. 7 Snapshots of a similar thunderstorm with time, computed by W_2, which happened on 04/26/2016.

C. Filtered Snapshots Comparison

We evaluate the efficacy of filter functions in the context of comparing thunderstorm snapshots for Dallas Fort Worth Airport (DFW). We identify key geographical locations for this analysis: the coordinates of four corner gates — Bowie (N33°32.15', W97°49.28'), Bonham (N33°32.25', W96°14.05'), Cedar Creek (N32°11.14', W96°13.09'), Glen Rose (N32°09.58' W97°52.66') — along with the center of the airport (N32°53.40', W97°02.40'). These locations are selected due to their strategic importance in the DFW airspace. We only consider the spatial coordinates for simplicity.

To construct the filter functions, we utilize a summation of five Gaussian partial density functions. The centers of these Gaussian functions correspond to the coordinates of the aforementioned important locations. We choose a standard deviation (σ) of 10 km to define the effective regions around these points, ensuring that the filter functions are concentrated around these areas of interest.

The resultant filter function is illustrated in Figure 8. This visualization provides insight into how the filter functions emphasize the selected key locations, with Gaussian peaks centered around each. The application of these filter functions in our analysis allows for a targeted examination of thunderstorm activity in relation to these critical areas within the DFW airport.



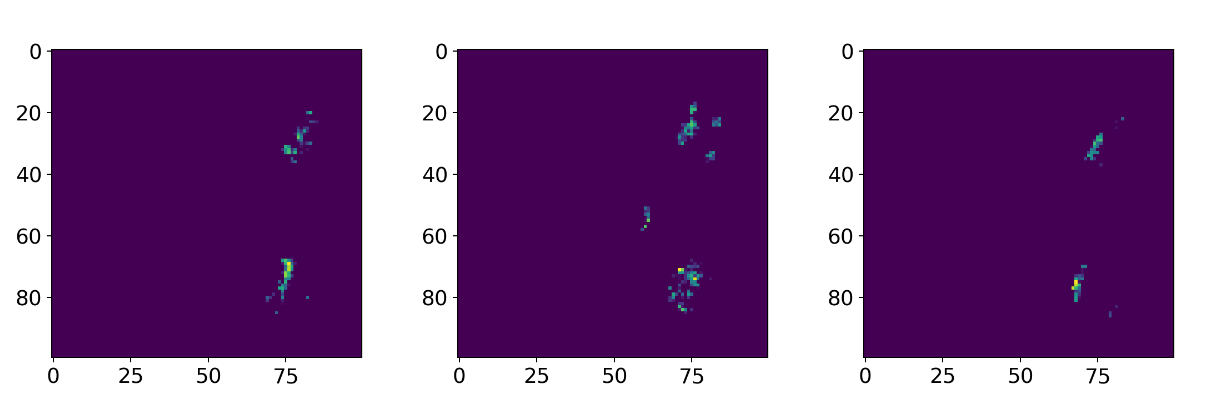
(a) The constructed filter function.

(b) A source thunderstorm snapshot.

(c) A filtered thunderstorm snapshot.

Fig. 8 The filter function at DFW and filtered image. The time is at 08:00-03/21/2022.

We retrieve the filtered snapshot in the dataset, compared with the source filtered image. These similar snapshots are depicted in Figure 9. The similarity of these thunderstorms is determined by their filter snapshots based on the W_2 (Wasserstein) distance, emphasizing the key geolocations. This comparative visualization underscores the effectiveness of our method in identifying spatially similar thunderstorms from a historical dataset that are operationally relevant.



(a)

(b)

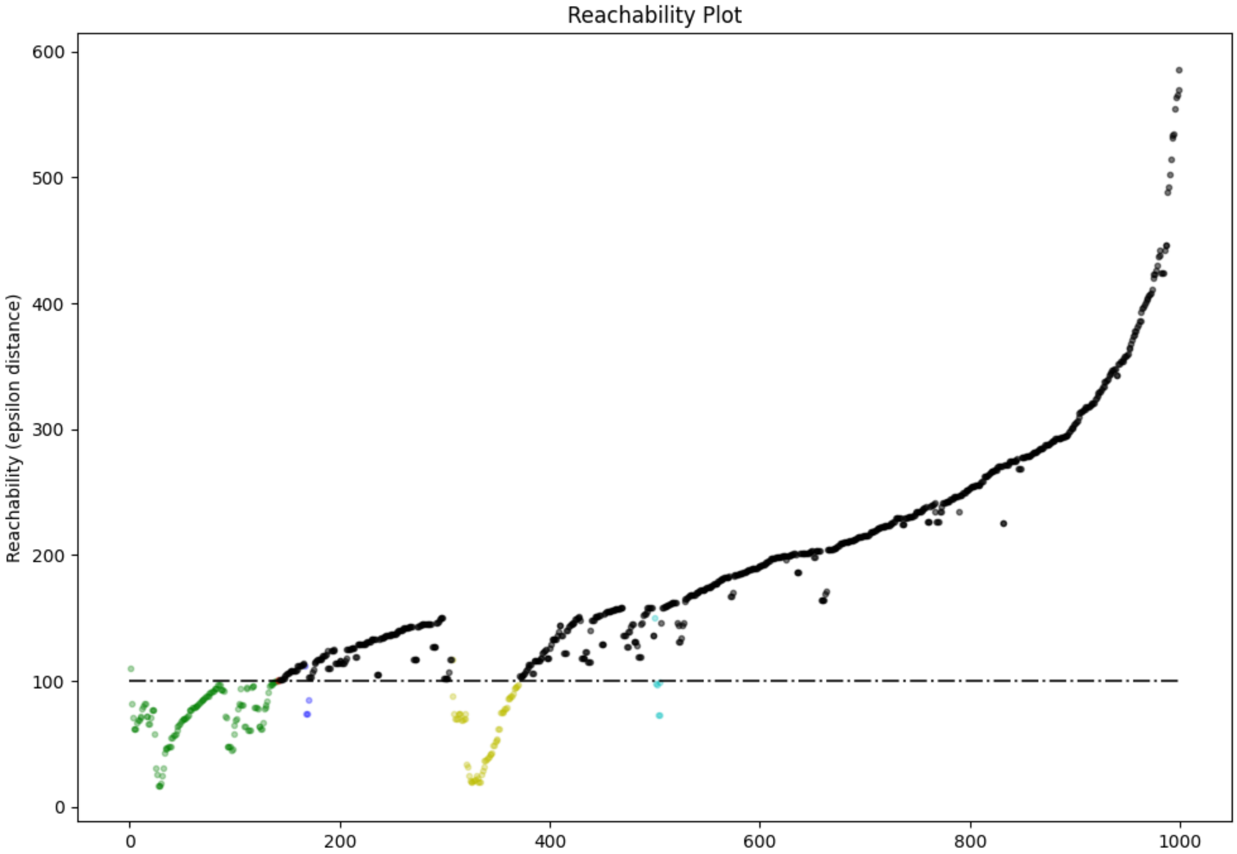
(c)

Fig. 9 Similar filtered images based on W_2 distance. Subfigures (a), (b) and (c) are similar thunderstorms at 16:20-03/30/2016, 17:30-03/24/2017 and 12:50-03/28/2018.

D. Clustering Results for Thunderstorm Sequences

In our study, we extended our analysis beyond examining pairwise similarities by employing the OPTICS algorithm [11] to cluster thunderstorm snapshots within our dataset. Specifically, we selected 1,000 thunderstorm samples with number of thunderstorm points being between 100 and 500, at random from the entire dataset for our clustering analysis. This subset of 1,000 samples provides a manageable yet sufficiently diverse representation of the overall dataset, allowing us to effectively apply the OPTICS algorithm.

For the OPTICS clustering, we set the epsilon distance to 100. The results of this clustering process are presented in Figure 10. The analysis revealed two clusters within these samples, highlighted in green and yellow in the reachability plot. These clusters indicate groupings of thunderstorms with similar characteristics.To further illustrate these findings, Figure 11 showcases two snapshots that correspond to the lowest points, or "bottoms", of these valleys in the reachability plot. These snapshots represent the most central or typical examples of the thunderstorms within each identified cluster.



(a) 20:45

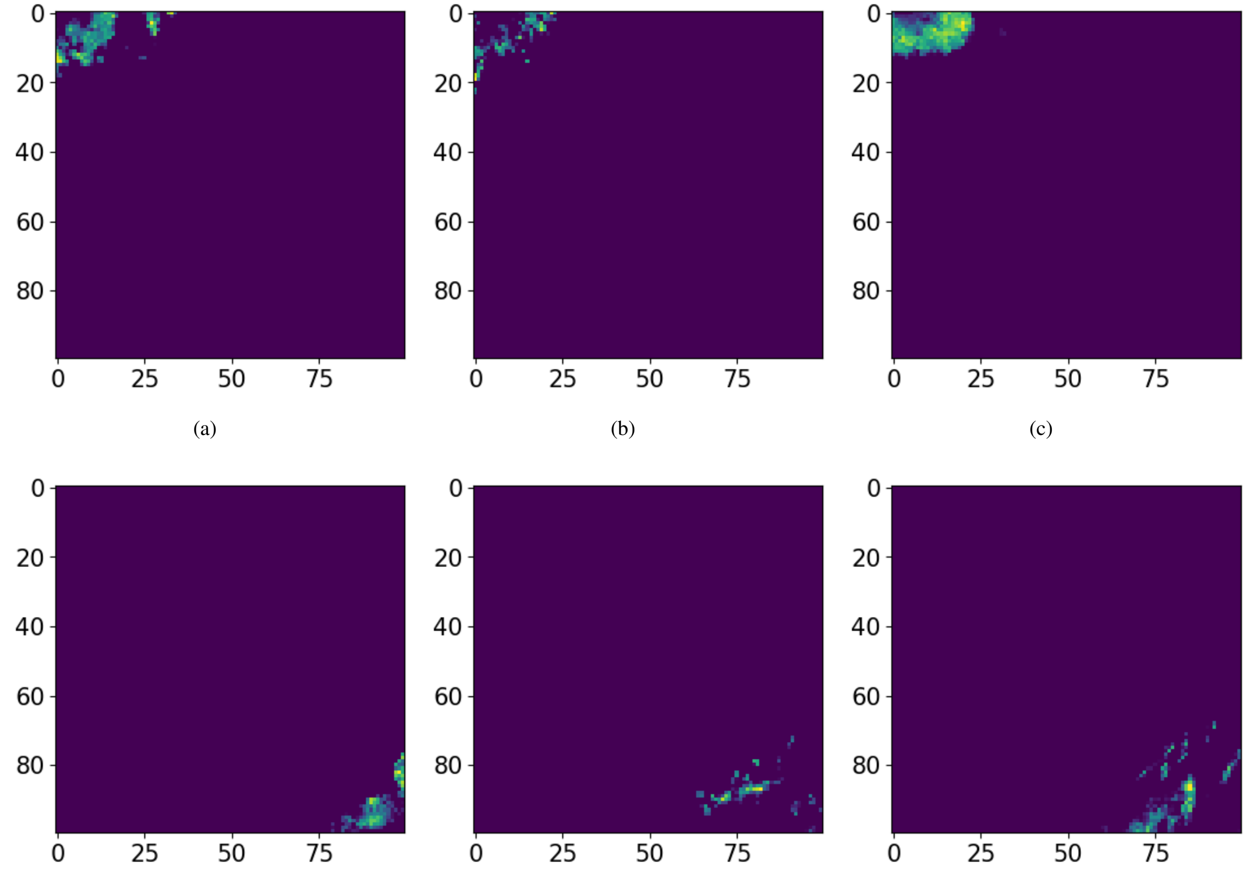
Fig. 10 Reachability plot from the OPTICS algorithm showing clustering of thunderstorm snapshots.

Figure 11 presents two distinct clusters of thunderstorm images, identified using the OPTICS algorithm. These clusters are visually differentiated by color for clearer distinction. As is shown, the green and yellow are two main clusters and the rest of them only have few points, therefore, we only focus on the green and yellow clusters. Subfigures (a), (b), and (c) represent the first cluster, which is highlighted in green color, indicating one group of thunderstorm patterns. Subfigures (d), (e), and (f) depict the second cluster, marked in yellow color. This color differentiation aids in visually distinguishing between the two sets of thunderstorm patterns, each representing a unique grouping as determined by the OPTICS clustering algorithm.

VI. Conclusions

In this paper, we introduce a novel method for identifying similar thunderstorm snapshots and sequences, leveraging the principles of optimal transport theory. The core of our approach is leveraging the Wasserstein distance to measure the similarity between thunderstorm probability distributions, represented by the thunderstorm intensity data points. We present two innovative techniques—incorporation of temporal information into coordinates and the application of filter functions—to tailor the probability distributions according to specific requirements in airline decision making.

Furthermore, we discuss the computational complexity of our proposed method and delineate its distinctions from



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 11 Figure presents two distinct clusters of thunderstorm images, identified using the OPTICS algorithm. These clusters are visually differentiated by color for clearer distinction. Subfigures (a), (b), and (c) represent the first cluster, which is highlighted in green color, indicating one group of thunderstorm patterns. Subfigures (d), (e), and (f) depict the second cluster, marked in yellow color. This color differentiation aids in visually distinguishing between the two sets of thunderstorm patterns, each representing a unique grouping as determined by the OPTICS clustering algorithm.

traditional Fractions Skill Score (FSS) methods. Additionally, we explore the application of clustering algorithms within the probability distribution space of thunderstorms. This approach aids in identifying common thunderstorm categories within a particular airspace.

The effectiveness of our method is demonstrated through practical application to real-world weather data in the Dallas Fort Worth airspace region. The results showcase our method's capability in providing insightful analysis and understanding of thunderstorm patterns, thereby contributing valuable tools for meteorological research and practical applications in airline operations.

Acknowledgments

We express our sincerest gratitude to the National Science Foundation under Award No. 2047390, for partially financially supporting this research. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government. We thank the staff and facilities at the NCAR Research Applications Laboratory for the use of their laboratory equipment. Prof. Matthias Steiner was critical in shaping both the direction and execution of this study, and for this, we are deeply thankful. We really appreciate Timothy Niznik, Yuxi Xiao, Billy Wang, and our colleagues from American Airlines provided insightful feedback for

which we are appreciative.

References

[1] Dixon, M., and Wiener, G., “TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A Radar-based Methodology,” *Journal of Atmospheric and Oceanic Technology*, Vol. 10, No. 6, 1993, pp. 785 – 797.

[2] Matthews, M., and Delaura, R., “Assessment and Interpretation of En Route Weather Avoidance Fields from the Convective Weather Avoidance Model,” *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, 2010. https://doi.org/10.2514/6.2010-9160.

[3] Yang, B., Gao, X., Han, Y., Zhang, Y., and Gao, T., “A Thunderstorm Identification Method Combining the Area of Graupel Distribution Region and Weather Radar Reflectivity,” *Earth and Space Science*, Vol. 7, 2020. https://doi.org/10.1029/2019EA000733.

[4] Huang, Y., Fan, Y., Cai, L., Cheng, S., and Wang, J., “A new thunderstorm identification algorithm based on total lightning activity,” *Earth and Space Science*, Vol. 9, No. 4, 2022. https://doi.org/https://doi.org/10.1029/2021EA002079.

[5] Roberts, N., and Lean, H., “Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events,” *Monthly Weather Review*, Vol. 136, 2008.

[6] Ran He, Z. S., Xiang Wu, and Tan, T., “Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. abs/1708.02412, 2017.

[7] Rubner, Y., Tomasi, C., and Guibas, L. J., “The earth mover’s distance as a metric for image retrieval,” *International journal of computer vision*, Vol. 40, No. 2, 2000, pp. 99–121. https://doi.org/https://doi.org/10.1023/A:1026543900054.

[8] Xue, B., Wu, L., Wang, K., Zhang, X., Cheng, J., Chen, X., and Chen, X., “Multiuser gesture recognition using sEMG signals via canonical correlation analysis and optimal transport,” *Computers in Biology and Medicine*, 2021.

[9] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1996, p. 226–231.

[10] Zhang, T., Ramakrishnan, R., and Livny, M., “BIRCH: an efficient data clustering method for very large databases,” *ACM sigmod record*, Vol. 25, No. 2, 1996, pp. 103–114. https://doi.org/https://doi.org/10.1145/235968.233324.

[11] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J., “OPTICS: Ordering points to identify the clustering structure,” *ACM Sigmod record*, Vol. 28, No. 2, 1999, pp. 49–60. https://doi.org/https://doi.org/10.1145/304181.304187.

[12] Adler, R. F., and Fenn, D. D., “Thunderstorm intensity as determined from satellite data,” *Journal of Applied Meteorology and Climatology*, Vol. 18, No. 4, 1979, pp. 502–517. https://doi.org/https://doi.org/10.1175/1520-0450(1979)018<0502:TIADFS>2.0.CO;2.

[13] Mittermaier, M. P., “A “Meta” Analysis of the Fractions Skill Score: The Limiting Case and Implications for Aggregation,” *Monthly Weather Review*, Vol. 149, No. 10, 2021, pp. 3491 – 3504. https://doi.org/https://doi.org/10.1175/MWR-D-18-0106.1.

[14] Roberts, N. M., and Lean, H. W., “Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events,” *Monthly Weather Review*, Vol. 136, No. 1, 2008, pp. 78 – 97. https://doi.org/https://doi.org/10.1175/2007MWR2123.1.

[15] Peyré, G., and Cuturi, M., “Computational Optimal Transport,” *arXiv*, 2020.

[16] Villani, C., “Optimal Transport: Old and New,” Springer, 2008.

[17] Amburn, S. A., and Wolf, P. L., “VIL Density as a Hail Indicator,” *Weather and Forecasting*, Vol. 12, No. 3, 1997, pp. 473 – 478. https://doi.org/https://doi.org/10.1175/1520-0434(1997)012<0473:VDAAHI>2.0.CO;2.

[18] Cuturi, M., “Sinkhorn distances: Lightspeed computation of optimal transport,” *Advances in neural information processing systems*, Vol. 26, 2013.

[19] Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G., “Faster Wasserstein distance estimation with the Sinkhorn divergence,” *Advances in Neural Information Processing Systems*, Vol. 33, 2020, pp. 2257–2269. https://doi.org/https://doi.org/10.48550/arXiv.2006.08172.

[20] Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., Blake, B. T., Brown, J. M., Olson, J. B., Hu, M., Smirnova, T. G., Ladwig, T., Kenyon, J. S., Ahmadov, R., Turner, D. D., Duda, J. D., and Alcott, T. I., "The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description," *Weather and Forecasting*, Vol. 37, No. 8, 2022, pp. 1371 – 1395. https://doi.org/10.1175/WAF-D-21-0151.1.