A Semi-Supervised Approach for Trajectory Segmentation to Identify Different Moisture Processes in the Atmosphere

Benjamin Ertl[1,2][0000−0003−1431−2243], Matthias Schneider[2][0000−0001−8452−0035], Christopher Diekmann[2][0000−0002−8961−5241], Jörg Meyer[1][0000−0003−0861−8481], and Achim Streit[1][0000−0002−5065−469X]

[1] Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2] Institute for Meteorology and Climate Research (IMK-ASF), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{benjamin.ertl,joerg.meyer2,matthias.schneider,
christopher.diekmann,achim.streit}@kit.edu

Abstract. Different moisture processes in the atmosphere leave distinctive isotopologue fingerprints. Therefore, the paired analysis of water vapour and the ratio between different isotopologues, for example $\{H_2O, \delta D\}$ with δD as the standardized HDO/H_2O isotopologue ratio, can be used to investigate these processes. In this paper, we propose a novel semi-supervised approach for trajectory segmentation to extract information that enables us to identify atmospheric moisture processes. While our approach can be transferred to a variety of domains as well, we focus our evaluation on Lagrangian air parcel trajectories and modelled $\{H_2O, \delta D\}$ fields. Our final aim is to understand the free tropospheric $\{H_2O, \delta D\}$ pair distribution that is observable by satellite sensors of the latest generation. Our method adopts a recently developed density-based clustering algorithm with constrained expansion, *CoExDBSCAN*, which identifies clusters of temporal neighbourhoods that are only expanded with regards to a priori constraints in defined subspaces. By formulating a constraint for the correlation of $\{H_2O, \delta D\}$, we can segment trajectories into multiple phases and extract the regression coefficients for each phase. Grouping segments with similar coefficients and comparing them to theoretical values allows us to find interpretable structures that correspond to atmospheric moisture processes. The experimental evaluation demonstrates that our method facilitates an efficient, data-driven analysis of large-scale climate data and multivariate time series in general.

Keywords: Semi-Supervised Clustering · Multivariate Time-Series · Time-Series Segmentation · Climate Research.

1 Introduction

With advances in technology that translates into an increasing amount of data, researchers across many disciplines face new challenges analysing and gaining

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

knowledge from massive volumes of data. For example, the U.S. National Oceanic and Atmospheric Administration (NOAA) cites for their Big Data Program that tens of terabytes of data are generated from satellites, radars, ships, weather models, and other sources a day [13]. Unsupervised learning methods such as cluster analysis are particularly useful in analyzing large amounts of data since it allows domain experts to consider groups of objects rather than individual objects and to focus on a higher level representation of the data [19]. While many advances in cluster algorithms have been made for spatiotemporal data [11, 19], such as atmospheric model data, many proposed methods lack the exploitation of available a priori knowledge that might improve the output quality [5]. Especially semi-supervised learning clustering algorithms, which incorporate a prior knowledge into the clustering process, can improve the quality of the results [3]. In this paper, we propose a novel semi-supervised approach for subsequence time series clustering based on our recently developed density-based clustering algorithm with constrained expansion called *CoExDBSCAN* [4]. By applying *CoExDBSCAN* we can segment trajectories of $\{H_2O, \delta D\}$ pair distributions into multiple phases which can be associated with atmospheric moisture processes. Identifying such processes is an important scientific task to infer the dynamics of cloud-circulation systems. Investigating the atmosphere from a cloud-circulation system perspective is essential to address the significant uncertainty of climate predictions [1].

The two main contributions of our work presented here are:

- Adaptation of our *CoExDBSCAN* algorithm for trajectory segmentation by formulating a constraint on the $\{H_2O, \delta D\}$ pair distribution to differentiate multiple phases of a trajectory.
- Extracting information about the regression coefficients for each phase and comparing the distribution of coefficients to theoretical values to identify corresponding atmospheric moisture processes.

The remainder of this paper is organized as follows. Section 2 introduces related work and background knowledge about our data, the theory behind the relation of $\{H_2O, \delta D\}$ and atmospheric moisture processes. In Section 3 we present the experimental evaluation of our proposed approach and provide a discussion of the results in Section 4. We conclude the paper in Section 5 and provide an outlook on future research.

All dataset together with the code for this paper are publicly available in the supplementary GitHub repository[3].

2 Background

2.1 Trajectory Clustering

Trajectory clustering and subsequence time series clustering are well established research fields. Trajectories can be described as sets of measurements which are

[3] https://github.com/bertl4398/iccs2021

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

measured as a function of an independent variable, typically time, where each individual trajectory measures a possible multidimensional response variable [7]. One of the first comprehensive methods for this type of data has been introduced by Gaffney et al. [7], who proposed a probabilistic mixture regression model applying the Expectation Maximization (EM) algorithm to cluster trajectories and demonstrated their approach analysing extratropical cyclones [8]. Since clustering whole trajectories can overlook common behaviour in partial segments of the trajectories, Lee at al. [10] proposed a partition-and-group framework and a trajectory clustering algorithm called *TRACLUS*, which they demonstrated among others in the field of climate research for hurricanes landfall forecasts. Following the given notion of trajectory data, there is no distinction to time series data in general, however the data records per individual trajectory can frequently be too short to be amenable to conventional time series modelling techniques, which requires specialized approaches [7]. Zolhavarieh et al. [21] compiled a well received survey about subsequence time series clustering algorithms and applications. More recently there have also been subsequence time series cluster algorithms proposed that are model-based [9], completely unsupervised [20] or semi-supervised [4]. We have recently developed the semi-supervised algorithm *CoExDBSCAN*, that utilizes the original *DBSCAN* algorithm proposed by Ester et al. [6], to find density-connected clusters in a defined subspace of features and restricts the expansion of clusters to a priori constraints. Because we can formulate an a priori constraint on the $\{H_2O, \delta D\}$ pair distribution of our data based on expert knowledge, *CoExDBSCAN* is a suitable choice to our problem. *CoExDBSCAN* has been demonstrated to be especially suited for spatiotemporal data, where one subspace of features defines the spatial extent of the data and another subspace the correlations between features. We can apply the algorithm to differentiate multiple phases in our trajectory data. However, by focusing on the temporal aspect of the trajectories, i.e. considering our data as multivariate time series, in distinction to the original algorithm we define the time space of the data as subspace for the distance based density computations. In this way we are able to find subsequences that follow our formulated constraint on the $\{H_2O, \delta D\}$ pair distribution. For a detailed explanation of the original *DBSCAN* and *CoExDBSCAN* algorithms as well as the pseudo code of the algorithms, we refer to the original papers by Ester et al. [6] and Ertl et al. [4] respectively.

2.2 Research Data

Our final research aim is to link the global $\{H_2O, \delta D\}$ pair distribution observed in the MUSICA IASI satellite-based remote sensing data set [17, 2] to different moisture processes that occurred prior to the observation. This dataset offers well-documented $\{H_2O, \delta D\}$ pair data from the year 2014 to 2020 with high quality and global coverage. The generation of this unique dataset has become only recently possible through advances in satellite sensor technology and retrieval theory. In this dataset and throughout this paper, H_2O indicates the water vapour concentration measured in parts per million by volume (ppmv); δD corresponds to the standardised ratio value between light and heavy water

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

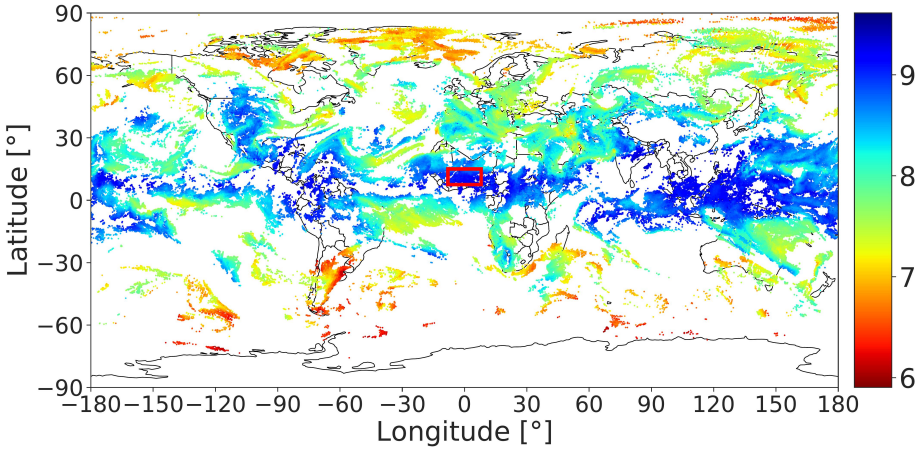


Fig. 1: Example global MUSICA IASI H_2O data for morning satellite overpasses at 2016-06-08. H_2O values are in parts per million by volume (ppmv) in logarithmic scale. The depicted data are limited to cloud free observations and have been filtered for best quality (retrievals with good sensitivity and low errors). The red square indicates our area of interest for the trajectory clustering.

i.e. H_2O and HDO [17]. Figure 1 and Figure 2 illustrate the characteristics of the MUSICA IASI dataset.

Figure 1 shows the global H_2O observations retrieved from the infrared atmospheric sounding interferometer (IASI) onboard the EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites) Metop-A and Metop-B (Meteorological Operational) platforms for morning overpasses at the 8^{th} June 2016 for about five kilometers altitude. For this single day 183,036 individual observations are available after filtering out cloudy and partly cloudy observations as well as observations with bad quality.

Figure 2 depicts the $\{H_2O, \delta D\}$ pair distribution starting from the same date at the 8^{th} June 2016 until the 30^{th} June 2016 for the area of interest (red rectangle in Figure 1). All MUSICA IASI $\{H_2O, \delta D\}$ data are shown as gray dots and the contours are at 2.5%, 10% and 50% levels, meaning the percentage of data lying outside the indicated area.

Different water cycle processes affect the isotopic composition of atmospheric water differently, for example lighter isotopes evaporate preferentially while heavier isotopes condense preferentially. The red lines in Figure 2 illustrate the theoretical dependencies of δD as a function of H_2O. Noone et al. differentiate between five processes that leave a distinct trace in the $\{H_2O, \delta D\}$ value space [14]:

1. **Rayleigh pseudoadiabatic** process in which the liquid water that condenses is assumed to be removed as soon as it is formed, by idealized instantaneous precipitation (red dotted line in Figure 2)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

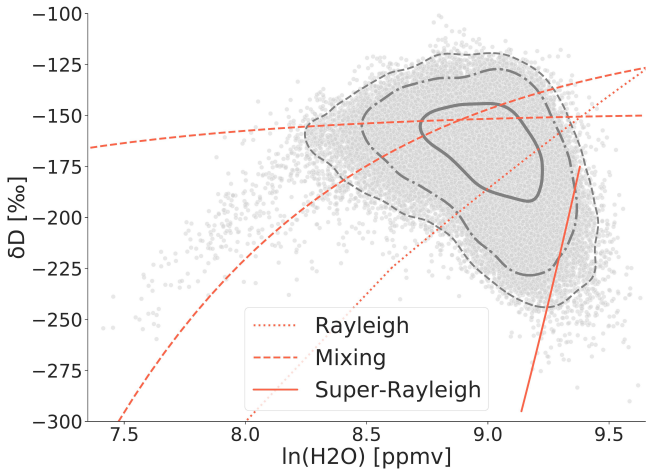


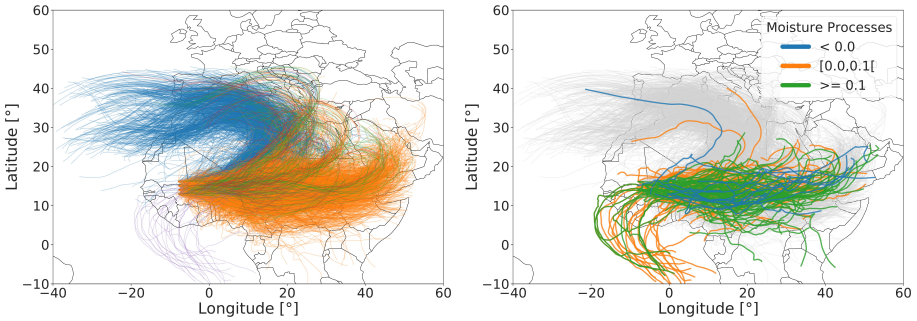
Fig. 2: Example MUSICA IASI $\{H_2O, \delta D\}$ pair distribution. The 61,283 observations are for morning and evening satellite overpasses from 2016-06-08 to 2016-07-30 with H_2O in logarithmic scale in the area of interest (see red rectangle in Figure 1); red lines indicate theoretical lines.

2. **Super-Rayleigh** remoistening associated with isotopic exchange as raindrops evaporate into a subsaturated layer (red solid line in Figure 2)
3. **Reversible moist adiabatic** process with a transition to a Rayleigh process when condensation is to ice and irreversible (not shown, would be a line with a weaker slope as the dotted line in Figure 2)
4. **Mixing** of two different mixing members having a specific $\{H_2O, \delta D\}$ characteristic (red dashed lines in Figure 2 show two examples)
5. **Terrestrial transpiration** mixing with land source (not shown)

Noone's work establishes a theoretical basis for using isotope ratio observations paired with the water vapor mixing ratio to identify different water sources, condensation processes, and transport pathways in the troposphere. Moreover, Noone et al. were able to derive slope and intercept of the linear relationship between H_2O and δD from measurements of the isotope ratio of water vapor at the Mauna Loa Observatory [15].

In this paper, we apply our segmentation algorithm to the $\{H_2O, \delta D\}$ pairs modelled along Lagrangian air parcel trajectories. As model data we use the high-resolution data from the regional isotope-enabled atmospheric model COSMO-iso [12] and the trajectories are determined with the tool LAGRANTO [18]. The trajectories' calculation setup is oriented towards the overpass times and altitudes representative for the MUSICA IASI data. Analysing the model data enables us to reveal the kind of moisture processes that can be observed in the MUSICA IASI $\{H_2O, \delta D\}$ pair data. We utilize the theoretical and observational findings by Noone in our experimental evaluation to identify atmospheric moisture processes that correspond to different $\{H_2O, \delta D\}$ pair distribution for different segments of our trajectory data. Our focus will be on rain events, where

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23



(a) Example DBSCAN Clustering Result (b) Example Moisture Processes

Fig. 3: Illustration of 3,194 trajectories out of 11,853 that have been colored according to their geographical closeness, bearing similarity and height difference along each individual trajectory (a). (b) illustrates the association of individual trajectories to different moisture processes as a result of the *CoExDBSCAN* segmentation.

Rayleigh pseudoadiabatic, Super-Rayleigh, and reversible moist adiabatic processes affect the $\{H_2O, \delta D\}$ pair distribution, in contrast to non-rain events, where air mass mixing processes are dominating the $\{H_2O, \delta D\}$ pair distribution. Our semi-supervised approach with the adaptation of the *CoExDBSCAN* algorithm with the appropriate constraint formulation for our research objective is detailed in the next section.

3 Experimental Evaluation

For our experimental evaluation, we organize our Lagrangian air parcel trajectory dataset as multivariate time series, e.g. backward trajectories for individual air parcels, and focus on an area of interest with the arrival of all trajectories above West Africa at pressure levels 575 and 625 hectopascal (hPa). The trajectories are calculated daily for local morning (9 am) and evening (9 pm) times during the period from June 8, 2016 to July 30, 2016, resulting in 12,720 individual trajectories (11,853 after filtering) with 169 time steps each with a time delta of one hour; each trajectory comprises a time frame of 7 days. Figure 3a illustrates our trajectory dataset, depicting 3,194 individual trajectories out of 11,853 that have been colored according to their similarity, using *DBSCAN* on a precomputed distance matrix. For the precomputed distance matrix, each trajectory has been converted to a $4 \cdot 169 = 676$ dimensional vector; the latitudinal (1) and longitudinal (2) difference for each time point to the arrival coordinates, the bearing (3) for each consecutive point and the scaled height difference (4) for each consecutive point, 169 points each. This initial clustering is only done to associate and compare individual segments as a result of our *CoExDBSCAN* segmentation, see Figure 3b, and is completely independent from our trajectory clustering.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

For the trajectory segmentation we adopt the *CoExDBSCAN* algorithm by defining the temporal order of the data points, the time dimension, as the spatial subspace. Following the definition of the *CoExDBSCAN* ϵ-neighbourhood, see [4] and Definition 1 for reference, with the time dimension as the spatial subspace the ϵ-neighbourhood describes a neighbourhood of lagged points, similar to a time window, where the maximum lag in time for the initial data points is defined by the ϵ parameter and the minimal amount of data points that are required to form a cluster is defined by the *minPts* parameter.

Definition 1. *Let DB be a database of points. The CoExDBSCAN ϵ-neighbourhood of a point p, denoted by $N_\epsilon(p)$, is defined by*

$$N_\epsilon(p) = \{q \in DB | dist(p_S, q_S) \leq \epsilon \wedge constraints(p_R, q_R)\} \tag{1}$$

where p_S, q_S are the subspace representations of point p and q of the user-defined spatial subspace S, p_R, q_R are the subspace representations of point p and q of the user-defined constraint subspace R and the `constraints` *function evaluates* `true` *for each constraint C_i in a user-defined set of constraints $C = \{C_1, C_2, ..., C_n\}$.*

In general, we consider a trajectory as a time series T of size m as an ordered sequence of real-value data, $T = (t_1, t_2, ..., t_m)$, and a subsequence of length n as $T_{i,n} = (t_i, t_{i+1}, ..., t_{i+n-1})$, where $1 \leq i \leq m - n + 1$; a subsequence is considered an arranged sequence of data that omits some elements without changing the order of the remaining element [21]. The algorithm starts with an initial time point and considers all temporal neighbours to form a cluster following Definition 1. Our constraint on the $\{H_2O, \delta D\}$ pair distribution to differentiate multiple phases of individual trajectories has been conceptualized together with domain experts and the constraint parameter δ empirically determined, see Definition 2. We constrain the expansion of clusters e.g. subsequences by including a neighbouring point in the ϵ-neighbourhood only if the residuals of an ordinary least squares linear regression for the current cluster points without the neighbouring point deviates only by a certain factor δ from the ordinary least squares linear regression including the neighbouring point, formulated in Definition 2.

Definition 2. *A point t_j belongs to a subsequence of points $T_{i,n} = (t_i, t_{i+1}, ..., t_{i+n-1})$, where $1 \leq i \leq m - n + 1$, iff*

$$(Y_{t_j} - \hat{Y}_{t_j})^2 < \delta \cdot \frac{1}{n} \sum_{k=i}^{(i+n-1)} (Y_{t_k} - \hat{Y}_{t_k})^2 \tag{2}$$

where Y and $\hat{Y}$ are the dependent variable and fitted value of the linear regression respectively.

Algorithm 1 gives a pseudo code representation of our semi-supervised trajectory segmentation approach. The algorithm takes a set of trajectories as input, as well as the parameters and subspace selections for *CoExDBSCAN*. To exemplify our approach and demonstrate the flexibility to segment likewise only

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

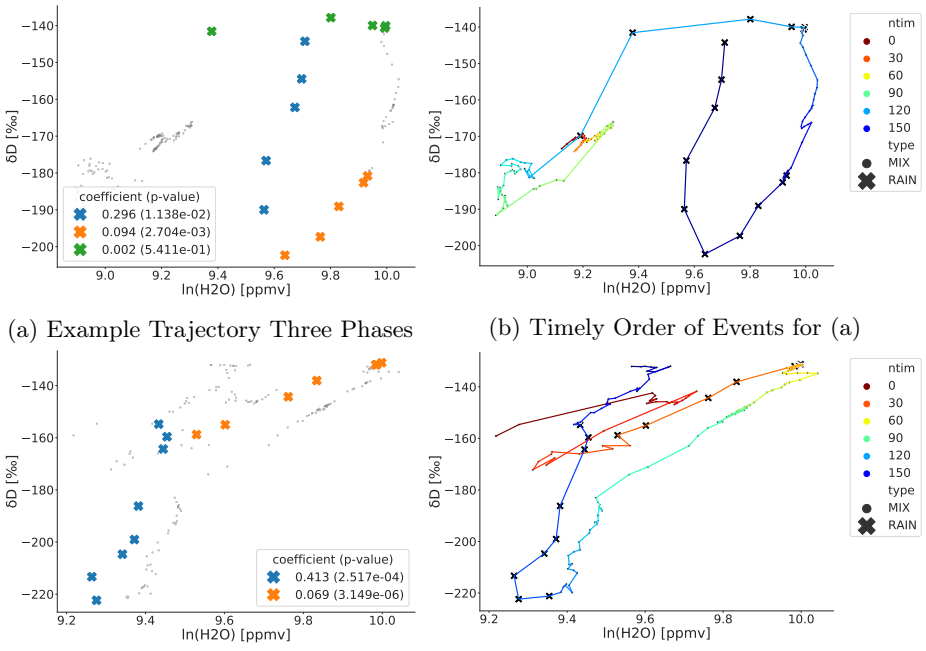
a subset of individual trajectories, we focus on rain events within trajectories without any loss of generality. These events have been identified by retrieving the time points where the moving average over three consecutive time points for the relative humidity are above a certain threshold and extending the event to at least three time points if necessary.

Algorithm 1: Semi-Supervised Trajectory Segmentation

```
   input  : trajectories T
   input  : time radius ε
   input  : density threshold minPts
   input  : residual threshold δ
   output: point labels per trajectory label_t initially undefined
 1 foreach trajectory t in trajectories T do
 2     timePoints = sortByTime(t, ascending);
 3     phasesAscending = CoExDBSCAN(timePoints.time,
         timePoints.{ln(H2O),ln(deltaD/1000 + 1)}, ε, minPts, δ);
 4     timePoints = sortByTime(t, descending);
 5     phasesDescending = CoExDBSCAN(timePoints.time,
         timePoints.{ln(H2O),ln(deltaD/1000 + 1)}, ε, minPts, δ)
 6     foreach phaseAscending in phasesAscending do
 7         foreach phaseDescending in phasesDescending do
 8             if sum(OrdinaryLeastSquares(phaseAscending).residuals^2) < sum(
                 OrdinaryLeastSquares(phaseAscending).residuals^2) then
 9                 label_t ← phaseAscending;
10             else
11                 label_t ← phaseAscending;
```

The identified trajectories with a subset of rain events are sorted by time in ascending and descending order, see Line 2 and 4 in Algorithm 1. For each time ordering we compute the labels using *CoExDBSCAN* with the time dimension, `timePoints.time`, as the spatial subspace and the natural logarithm of the water vapour values together with the natural logarithm of the isotopologue ratio value divided by 1,000 plus one, `timePoints.{ln(H2O),ln(deltaD/1000 + 1)}`, as the constraint subspace, see Line 3 and 5. The remaining parameters have been empirically determined and proposed by domain experts and set to $\epsilon = 2$, $minPts = 3$ and $\delta = 2$ in this example. Computing the segmentation in both temporal orders is necessary, because the outcome of the linear regression in our constraint depends on the deviation of the residuals from the current cluster points, which can be different following the trajectory points in ascending or descending temporal order. The final segmentation of the trajectory, or the subset of time points within a trajectory, is the selection of phases from the ascending and descending *CoExDBSCAN* run where the outcome with the squared residual sum is lowest, Line 6 to 11 in the algorithm.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23



(a) Example Trajectory Three Phases

(b) Timely Order of Events for (a)

(c) Example Trajectory Two Phases

(d) Timely Order of Events for (c)

Fig. 4: Example trajectories with two rain sequences, which can be identified by the crosses. (a) with additional segmentation of a continuous sequences based on the defined constraint for the regression coefficients and (c) without additional segmentation. (b) and (d) illustrate the timely order of events according to the number of hours before arrival from 168 to 0 (dark blue to dark red).

Figure 4 shows the segmentation result of two example trajectories. Figure 4a and Figure 4c depict the results of our semi-supervised segmentation approach with dots indicating non-rain events (air mass mixing) and crosses indicating rain events, which have been segmented into different coloured phases with different regression coefficients. Figure 4b and Figure 4d visualize the temporal order of the corresponding trajectories' events coloured according to the number of hours before arrival from 168 to 0 (dark blue to dark red). The first example trajectory in Figure 4a has two rain events with three distinct phases, where the first event from hour 168 to hour 159 before arrival (see Figure 4b) has been segmented into two different phases with different regression coefficients with statistical significance (p-value < 0.05, blue crosses and orange crosses); the second event starts at hour 124 and ends at hour 119 before arrival with a steady regression coefficient, however not statistically significant (green crosses). The second example trajectory in Figure 4c shows again two rain events, the first from hour 140 to hour 133 before arrival and the second from hour 38 to hour 32 before arrival. The regression coefficients are unvarying with statistical significance (blue crosses and orange crosses). We clearly observe different slopes of the linear regression lines fitted to the two different rain events.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

Comparing the rain segments from both example trajectories with the theoretical evolution of δD as a function of H_2O, see Section 2.2, we can interpret the segmented event in Figure 4a (blue and orange crosses) and the two timely separated phases in Figure 4c (blue and orange crosses) as different kind of rain processes. For example, the first phase in Figure 4c (blue crosses) is in line with super-Rayleigh processes, i.e. there is some interchange between condensed moisture and vapour. The second phase (orange crosses) is close to a Rayleigh process, i.e. can be explained by condensation and direct rainout, without significant interchange between condensed moisture and vapour.

These two example trajectories demonstrate that our approach can **(1) identify timely separated events** as well as **(2) split timely connected events** with varying regression coefficients. The latter gives novel opportunities for identifying details of the related rain processes. Being able to distinguish these fine-grained structures in large volumes of data emphasizes the benefit of applying our method, which we discuss in the following section.

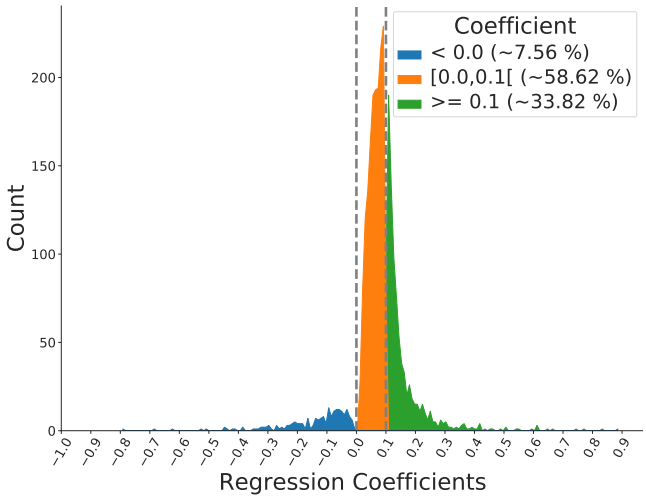
The histogram in Figure 5a outlines the distribution of regression coefficients with statistical significance (p-value < 0.05) for all rain segments as a result of our semi-supervised trajectory segmentation, analysing all 11,853 trajectories, clipped to the interval $[-1.0, 1.0]$ (omitting four segments < -1.0 and three segments > 1.0). Of all rain events 7.6% have a negative slope, 58.6% have a slope between 0.0 and 0.1, and 33.8% have a slope greater than 0.1.

Figure 5b depicts the distribution of the $\{H_2O, \delta D\}$ pairs modelled in the area of interest at the altitudes that are representative for the MUSICA IASI data. The grey dots show all model data, i.e. represent the same as the grey dots in Figure 2, but modelled instead of measured data. The lines are 50% contour lines. The grey contour line is for the whole data set, see Figure 2. The blue, orange, and green contour lines are for data points representing air masses that experienced rain events during the last five days prior to arrival. The colours are according to the coefficients that characterise the rain events (i.e. they are in line with Figure 5a). If an air mass experienced rain events with two different characteristics, the respective data point belongs to both groups. We can clearly identify air masses that experienced a Super-Rayleigh process (green contour line is almost completely below the theoretical Rayleigh line).

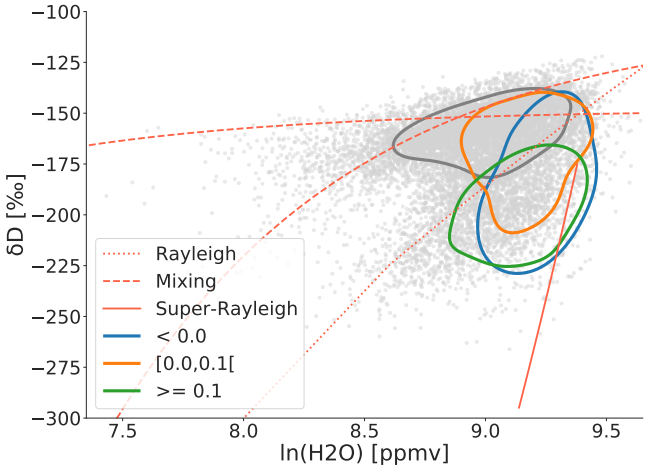
4 Discussion

As our experimental evaluation demonstrates, our approach can successfully segment time series, e.g. trajectories or subsequences of trajectories, into sequences that contain temporal close points which follow a priori constraints. With our constraint formulated in Definition 2 each segment is differentiated by the deviation from the ordinary least squares regression residuals, which in effect splits sequences if their individual linear regression is a better fit than their combined linear regression. This constraint is particularly useful for our research objective to analyse $\{H_2O, \delta D\}$ pair distributions that follow theoretical linear relations. However, in general this constraint restricts the cluster expansion to the cor-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23



(a) Histogram of regression coefficients



(b) Distribution of regression coefficients

Fig. 5: Histogram (a) of regression coefficients (slope of linear regression line) for all rain segments as a result of our semi-supervised trajectory segmentation with statistical significance (p-value < 0.05); and $\{H_2O, \delta D\}$ distributions (b) of the model data in the area of interest for all data points (grey dots and contour line) and for data points representing air masses that experienced rain events (blue, orange and green colours are as in (a) and represent rain events having different regression coefficients); contour levels are at 50%.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

relation of time point values and can be transferred to different domains and datasets. For example motion data in 2D or 3D space can be segmented with our approach and the same constraint formulation as well, identifying recurring, similar motions that follow similar linear correlations.

Segmenting our Lagrangian air parcel trajectories enables us to better understand the complex dynamics that cause the $\{H_2O, \delta D\}$ pair distribution observable in the MUSICA IASI dataset. Since our approach is able to distinguish fine-grained structures in large volumes of data, it is an effective data-driven analysis method for this purpose. For instance, with the grouping of trajectories we can draw conclusions on atmospheric moisture transport patterns: Figure 3b suggests that if we observe a $\{H_2O, \delta D\}$ data pair below the Rayleigh line, the air mass has very likely been transported from East to West Africa.

Ongoing research in the area of observational atmospheric data orientated towards the combination of different sensors might soon allow the global detection of the vertical distribution of $\{H_2O, \delta D\}$ pairs (a method for such synergetic combination has recently been demonstrated using CH_4 as an example, Schneider et al. [16]). By using state-of-the-art reanalysis datasets, for calculating the trajectories our approach could then be used to directly identify different moisture processes in the measured $\{H_2O, \delta D\}$ fields.

5 Conclusion

In this paper we propose a novel semi-supervised approach for trajectory segmentation and demonstrate our approach to identify different processes in the atmosphere. We adopt our recently developed algorithm *CoExDBSCAN*, density-based clustering with constrained expansion, for trajectory segmentation by formulating a constraint on the deviation from the ordinary least squares regression residuals of combined or split segments. By further extracting information about the regression coefficients for each segment and comparing the distribution of coefficients to theoretical values we are able to identify corresponding atmospheric processes. This approach is demonstrated in our experimental evaluation for Lagrangian air parcel trajectories together with data from the regional isotope-enabled atmospheric model COSMO-iso [12]. We are able to successfully extract segments from subsequences of temporal continuous events and compare these segments with the theoretical evolution of our dependent variable δD as a function of H_2O. By orienting the trajectory calculations towards the overpass times and altitudes represented by the satellite data, we can use this model evaluation to better understand the satellite data.

For future work, we plan to further improve the extraction of specific $\{H_2O, \delta D\}$ sequences in the model domain and use the model data of rain, cloud etc. to clearly relate specific $\{H_2O, \delta D\}$ distributions to distinct moist processes. For instance we want to demonstrate that our method can be used to distinguish between air masses that experienced shallow convection or deep convection. The long-term perspective is to generate multisensor $\{H_2O, \delta D\}$ observation data, i.e. to create observation data that offer $\{H_2O, \delta D\}$ pair information at different

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

altitudes. Then use trajectories calculated from state-of-the-art reanalysis fields and apply the method directly to the observational data. Data-driven analysis of observation data could significantly improve the insight into the dynamics of such cloud-circulation systems and would help to improve the significant uncertainty of climate predictions.

References

1. Bony, S., Stevens, B., Frierson, D.M.W., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T.G., Sherwood, S.C., Siebesma, A.P., Sobel, A.H., Watanabe, M., Webb, M.J.: Clouds, circulation and climate sensitivity. Nature Geoscience **8**(4), 261–268 (Apr 2015). https://doi.org/10.1038/ngeo2398
2. Borger, C., Schneider, M., Ertl, B., Hase, F., García, O.E., Sommer, M., Höpfner, M., Tjemkes, S.A., Calbet, X.: Evaluation of MUSICA IASI tropospheric water vapour profiles using theoretical error assessments and comparisons to GRUAN vaisala rs92 measurements. Atmospheric Measurement Techniques **11**(9), 4981–5006 (2018). https://doi.org/10.5194/amt-11-4981-2018
3. Dinler, D., Tural, M.K.: A Survey of Constrained Clustering, pp. 207–235. Springer International Publishing, Cham (2016)
4. Ertl., B., Meyer., J., Schneider., M., Streit., A.: CoExDBSCAN: Density-based Clustering with Constrained Expansion. In: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,. pp. 104–115. INSTICC, SciTePress (2020). https://doi.org/10.5220/0010131201040115
5. Ertl., B., Meyer., J., Streit., A., Schneider., M.: Application of Mixtures of Gaussians for Tracking Clusters in Spatio-Temporal Data. In: Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,. pp. 45–54. INSTICC, SciTePress (2019). https://doi.org/10.5220/0007949700450054
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
7. Gaffney, S., Smyth, P.: Trajectory Clustering with Mixtures of Regression Models. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 63–72. KDD '99, Association for Computing Machinery, New York, NY, USA (1999). https://doi.org/10.1145/312129.312198
8. Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J., Ghil, M.: Probabilistic clustering of extratropical cyclones using regression mixture models. Climate Dynamics **29**(4), 423–440 (Sep 2007). https://doi.org/10.1007/s00382-007-0235-z
9. Hallac, D., Vare, S., Boyd, S., Leskovec, J.: Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 215–223. KDD '17, Association for Computing Machinery, New York, NY, USA (2017)
10. Lee, J.G., Han, J., Whang, K.Y.: Trajectory Clustering: A Partition-and-Group Framework. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. p. 593–604. SIGMOD '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1247480.1247546

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23

11. Maciąg, P.S.: A Survey on Data Mining Methods for Clustering Complex Spatiotemporal Data. In: Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. (eds.) Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation. pp. 115–126. Springer International Publishing, Cham (2017)
12. Miltenberger, A.K., Pfahl, S., Wernli, H.: An online trajectory module (version 1.0) for the nonhydrostatic numerical weather prediction model COSMO. Geoscientific Model Development **6**(6), 1989–2004 (2013)
13. NOAA: National Oceanic and Atmospheric Administration Big Data Program. https://www.noaa.gov/organization/information-technology/big-data-program, Accessed: 2020-11-30
14. Noone, D.: Pairing Measurements of the Water Vapor Isotope Ratio with Humidity to Deduce Atmospheric Moistening and Dehydration in the Tropical Midtroposphere. Journal of Climate **25**(13), 4476 – 4494 (01 Jul 2012). https://doi.org/10.1175/JCLI-D-11-00582.1
15. Noone, D., Galewsky, J., Sharp, Z.D., Worden, J., Barnes, J., Baer, D., Bailey, A., Brown, D.P., Christensen, L., Crosson, E., Dong, F., Hurley, J.V., Johnson, L.R., Strong, M., Toohey, D., Van Pelt, A., Wright, J.S.: Properties of air mass mixing and humidity in the subtropics from measurements of the D/H isotope ratio of water vapor at the Mauna Loa Observatory. Journal of Geophysical Research: Atmospheres **116**(D22) (2011). https://doi.org/10.1029/2011JD015773
16. Schneider, M., Ertl, B., Diekmann, C., Khosrawi, F., Röhling, A.N., Hase, F., Dubravica, D., García, O.E., Sepúlveda, E., Borsdorff, T., Landgraf, J., Lorente, A., Chen, H., Kivi, R., Laemmel, T., Ramonet, M., Crevoisier, C., Pernin, J., Steinbacher, M., Meinhardt, F., Deutscher, N.M., Griffith, D.W.T., Velazco, V.A., Pollard, D.F.: Synergetic use of IASI and TROPOMI space borne sensors for generating a tropospheric methane profile product. submitted to Atmospheric Measurement Techniques (2021)
17. Schneider, M., Wiegele, A., Barthlott, S., González, Y., Christner, E., Dyroff, C., García, O.E., Hase, F., Blumenstock, T., Sepúlveda, E., Mengistu Tsidu, G., Takele Kenea, S., Rodríguez, S., Andrey, J.: Accomplishments of the MUSICA project to provide accurate, long-term, global and high-resolution observations of tropospheric $\{H_2O,\delta D\}$ pairs – a review. Atmospheric Measurement Techniques **9**(7), 2845–2875 (2016). https://doi.org/10.5194/amt-9-2845-2016
18. Sprenger, M., Wernli, H.: The LAGRANTO Lagrangian analysis tool – version 2.0. Geoscientific Model Development **8**(8), 2569–2586 (2015). https://doi.org/10.5194/gmd-8-2569-2015
19. Wang, S., Cai, T., Eick, C.F.: New Spatiotemporal Clustering Algorithms and their Applications to Ozone Pollution. In: 2013 IEEE 13th International Conference on Data Mining Workshops. pp. 1061–1068 (2013). https://doi.org/10.1109/ICDMW.2013.14
20. Zhang, Q., Wu, J., Zhang, P., Long, G., Zhang, C.: Salient Subsequence Learning for Time Series Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9), 2193–2207 (2019). https://doi.org/10.1109/TPAMI.2018.2847699
21. Zolhavarieh, S., Aghabozorgi, S., Teh, Y.W.: A Review of Subsequence Time Series Clustering. The Scientific World Journal **2014**, 312521 (Jul 2014). https://doi.org/10.1155/2014/312521

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_23