This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/

Spatio-temporal clustering of natural hazards

Zandovskis, Uldis

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT

Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. May. 2024

Spatio-temporal clustering of natural hazards

Uldis Zandovskis

Thesis submitted for the degree of Doctor of Philosophy (PhD)

Funded by an EPSRC Studentship

Department of Geography
King's College London
United Kingdom

2 July 2023

Table of Contents

List of Figures ..... **v**

List of Tables ..... **xii**

Abstract ..... **xiv**

Acknowledgements ..... **xv**

1. Introduction **1**

1.1. Motivation 2

1.2. Aim and objectives 8

1.3. Structure of the thesis 10

2. Critical review of spatio-temporal clustering methodologies **13**

2.1. Introduction 14

2.2. General review of spatio-temporal clustering methodologies 14

2.2.1. Critical review approach 15

2.2.2. Spatio-temporal clustering research timeline 19

2.2.3. Application of methodologies to various datasets and interpretation 20

2.3. Spatio-temporal clustering methodologies 22

2.3.1. Point events 23

2.3.2. Trajectories 28

2.4. Characterisation of clustering methodologies as global and local clustering measures 31

2.5. Combining global and local clustering measures in a spatio-temporal clustering analysis framework 32

2.6. Chapter conclusions 35

3. Natural hazard datasets: examples, lightning characterisation, and synthetic data **36**

3.1. Introduction 37

3.2. Examples of real-world spatio-temporal natural hazard data 38

3.2.1. Real-world point event data 38

3.2.2. Real-world trajectory data 40

3.3. Characterisation and modelling of lightning strikes as point events in time and space 41

3.3.1. Review of existing point event models 43

3.3.2. Data collection of lightning and storms 45

3.3.3. Methods to characterise lightning strike parameters in time and space 61

3.3.4. Results of variable characterisation 68
3.3.5. Modelling lightning strikes 78
3.3.6. Limitations of the characterisation and modelling approach 86
3.4. Simulation study of synthetic models and benchmark datasets 89
3.4.1. Point event benchmarks 89
3.4.2. Trajectory benchmarks 96
3.5. Chapter conclusions 99

4. Application of the spatio-temporal clustering methodologies to real-world and synthetic natural hazard datasets 101
4.1. Introduction 102
4.2. Results and analysis for point event simulation study 103
4.2.1. K-function analysis 103
4.2.2. Kernel density estimation 112
4.2.3. Space-time scan statistic 120
4.2.4. Initial evaluation of the spatio-temporal clustering framework 129
4.3. Results and analysis for synthetic benchmark trajectory datasets 130
4.4. Results and analysis for real-world point event data 133
4.4.1. K-function analysis 134
4.4.2. Kernel density estimation 136
4.4.3. Space-time scan statistic 137
4.4.4. Evaluation of limitations of the selected methodologies 140
4.5. Results and analysis for real-world trajectory datasets 141
4.6. Chapter conclusions 150

5. Improvements on the spatio-temporal clustering methodologies for application to point events produced by a moving source 151
5.1. Introduction 152
5.2. Improvements to the K-function analysis 152
5.3. Improvements to the space-time scan statistic 154
5.4. Results of improved clustering measures on synthetic datasets produced by the lightning strike models 157
5.5. Results of improved clustering measures on real-world datasets 169
5.6. Evaluation of improved clustering measures 173
5.7. Chapter conclusions 175

6. Summary, conclusions and further research directions 176
6.1. Introduction 176
6.2. Relationship of thesis to original research objectives 177
6.3. Further research directions 182
6.3.1. Thunderstorm movement tracking to characterise spatial spread 183
6.3.2. K-function spatial and temporal weights 185

6.3.3. Trajectory-OPTICS density-based clustering methodology 187
6.3.4. Further applications of thesis research to natural hazard processes 189
6.4. Concluding remarks 192

References 193

Appendix A. Critically reviewed publications 204

Appendix B. K-function derivation 218

Appendix C. Kernel density estimation derivation 219

Appendix D. Space-time scan statistic derivation 220

Appendix E. Trajectory-OPTICS implementation 222

Appendix F. Synthetic data generator algorithms 224

Appendix G. List of publications on modelling natural hazards as point events 226

Appendix H. Spatial extent and propagation of the lightning strike occurrences for both storm-systems 228

Appendix I. 2D spatial count of lightning strikes in easting and northing 230

Appendix J. Cutting time series of inter-event times 231

Appendix K. Least-squares plane-fit solution to lightning datasets 232

Appendix L. Linear least-squares fit to estimate storm movement speed in the spatio-temporal lag space 233

Appendix M. Bivariate Gaussian fit to spatial spread variable to sample easting and northing distances 239

Appendix N. Sampling, sorting and re-indexing procedure for inter-event time variable 239

Appendix O. Spatio-temporal clustering results for regular grid data 243

Appendix P. Spatial and temporal views of the three storm reference Model 4 and its variations 248

Appendix Q. Inclined elliptic space-time scan statistic spatio-temporal clustering results for Models 4-4.3 250

List of Figures

Chapter 1:

Figure 1.1. Examples of **(a)** synthetic and **(b)** real-world point event data 5
Figure 1.2. Examples of **(a)** synthetic and **(b)** real-world trajectory data..... 6
Figure 1.3. Various physical properties of the natural hazard process 7
Figure 1.4. Structure and workflow of the research done during this thesis..... 8

Chapter 2:

Figure 2.1. Critical review of spatial-temporal clustering 19
Figure 2.2. Distribution of the 'Applied datasets' review subcategory (**Table 2.1**) options .. 21
Figure 2.3. Distribution of the 'Applied datasets' and 'Interpretation' review subcategory .. 22

Chapter 3:

Figure 3.1. A single day of lightning strike data collected on 1 July 2015..... 39
Figure 3.2. The subset (20 year data from 1998-2017 over the North Atlantic basin) of the Atlantic hurricane database HURDAT2 40
Figure 3.3. The spatial extent of lightning strike occurrences 47
Figure 3.4. The lightning strike data subset available for assignment to Storm 1B 48
Figure 3.5. The availability of Doppler radar station location information..... 49
Figure 3.6. The assignment of lightning strikes to Storm 1B 50
Figure 3.7. Spatial extent and propagation of the lightning strike occurrences as recorded on 28 June 2012 52
Figure 3.8. Spatial extent for 28 June 2012 of the lightning strike occurrences..... 54
Figure 3.9. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for Storm 1A 56
Figure 3.10. Spatial extent and propagation of the lightning strike occurrences as recorded on 1 July 2015..... 58
Figure 3.11. Spatial extent for 1 July 2015 of the lightning strike occurrences 59

Figure 3.12. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for Storm 2A 60
Figure 3.13. Two subplots showing the removal of inter-event time $\Delta\tau$ values in real-time t (removal of the black line in upper subfigure) and the subsequent inter-event time-series in order of occurrence k (lower subfigure) 69
Figure 3.14. Violin plots of inter-event time $\Delta\tau$ distributions for all six selected storms 70
Figure 3.15. Violin plots of inter-event time $\Delta\tau$ distributions for each one-hour interval when lightning strikes were observed after cutting (**Figure 9**) of the time-series 72
Figure 3.16. Least-squares plane-fit solution to the lightning strike datasets 74
Figure 3.17. Least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. 75
Figure 3.18. 2D spatial count of lightning strikes per 1 km^2 of the easting and northing distances to the movement track in natural time k for lightning strikes in Storm 1A 77
Figure 3.19. Cumulative distribution function (CDF) for the inter-event time $\Delta\tau$ for all six identified thunderstorms and the combined supercell thunderstorm dataset 80
Figure 3.20. Sampling, sorting and re-indexing procedure based on the combined supercell inter-event time $\Delta\tau$ dataset and thunderstorm (**a#**) 1A, (**b#**) 2C. The first subfigure shows the original cut inter-event time time-series in natural time k 81
Figure 3.21. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for a bivariate Gaussian distribution based on Storm 1A with window size $\Delta w = 500$ events 82
Figure 3.22. A single run of the spatio-temporal model to produce a lightning strike dataset representative of Storm 1A and a synthetic three-storm-system with varying spatial spread parameters. 84
Figure 3.23. A synthetic three-storm-system that consists of individual synthetic storms that would be representative of Storms 1A, 1B and 1C 85
Figure 3.24. The selected initiations of the reference single thunderstorm model (**Model 3**), and its variations. 93
Figure 3.25. The selected initiations of the reference three thunderstorm model (**Model 4**) and its variations. 93
Figure 3.26. The two synthetic datasets used for spatio-temporal clustering analysis of trajectory data 98

Chapter 4:

Figure 4.1. Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to homogeneous point event data (**Model 1**) 103
Figure 4.2. Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to clustered point event data (**Model 2**). 106
Figure 4.3. The clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to the single thunderstorm reference model (**Model 3**) and its variations 107
Figure 4.4. The clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to the three thunderstorm reference model (**Model 4**) and its variations 109
Figure 4.5. The collapsed datasets of the reference model for three thunderstorms (**Model 4**) and its variation 110
Figure 4.6. Volume rendering of probability density values as returned by the kernel density estimation methodology for homogeneous (**Model 1**) and clustered (**Model 2**) point event datasets 113
Figure 4.7. Volume rendering of voxels with the highest 10% of the probability density values as returned by the kernel density estimation methodology for homogeneous (**Model 1**) and clustered (**Model 2**) point event datasets 114
Figure 4.8. Volume rendering of probability density values as returned by the kernel density estimation methodology for the single thunderstorm reference model (**Model 3**) and its variations 116
Figure 4.9. Volume rendering of voxels with highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the single thunderstorm reference model (**Model 3**) and its variations 117
Figure 4.10. Volume rendering of probability density values as returned by the kernel density estimation methodology for the three thunderstorm reference model (**Model 4**) and its variations 118
Figure 4.11. Volume rendering of voxels with highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the reference three thunderstorm model (**Model 4**) and its variations 119
Figure 4.12. Various angles of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **Model 2** 122

Figure 4.13. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **Model 3** and its variations. 124
Figure 4.14. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **Model 4** and its variations 127
Figure 4.15. The number of core trajectories n_{CT} for various neighbourhood distance threshold ε and minimum number of trajectories within given threshold n_{pts} combinations..... 130
Figure 4.16. Reachability distance D_R plots for homogeneous and clustered trajectories as numbered by their visiting order number N_{VO} 131
Figure 4.17. The visiting order and clustering results after the execution of the Trajectory-OPTICS methodology on clustered trajectories.. 133
Figure 4.18. The clustering measure $D = K(s,\tau)–K(s)K(\tau)$ values as applied to lightning strike location data. 135
Figure 4.19. Volume rendering of all and only highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the real-world lightning strike dataset 137
Figure 4.20. Various angles of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to the real-world lightning strike data 139
Figure 4.21. Real-world hurricane trajectory dataset after alignment of all trajectories at $t = 0$ and application of a duration threshold of $n_t = 30$, which corresponds to $T = 7.5$ d..... 143
Figure 4.22. Clustering results of the $N = 136$ real-world hurricane trajectories with alignment at $t = 0$ and duration cut-off of $T = 7.5$ d 144
Figure 4.23. Reachability distance D_R plots for various sets of input parameters ε and n_{pts}. 146
Figure 4.24. Reachability distance D_R plots and clustering results for various trajectory duration threshold as imposed by the number of selected time steps n_t 148

Chapter 5:

Figure 5.1. Various geometries of the space-time statistic for the lightning strike location (represented by arbitrary point events as black dots) dataset. 156
Figure 5.2. Three subplots of the standardised clustering measure as applied to the homogeneous (**Model 1**) and the clustered (**Model 2**) point event data..... 158

Figure 5.3. The standardised clustering measure $D' = K'(s,\tau)–K'(s)K'(\tau)$ as applied to the reference single thunderstorm model (**Model 3**) and its variations 160
Figure 5.4. The standardised clustering measure $D' = K'(s,\tau)–K'(s)K'(\tau)$ as applied to a single initiation of the homogeneous Poisson process (**Model 1**) with $N = 3{,}000$ point events 162
Figure 5.5. The results of a Monte Carlo (MC) test for significance of the standardised clustering measure *D'* results 163
Figure 5.6. The standardised clustering measure $D' = K'(s,\tau)–K'(s)K'(\tau)$ as applied to the reference three thunderstorm model (**Model 4**) and its variations 164
Figure 5.7. 3D views of the spatio-temporal domain showing the point events contained within the inclined elliptic space-time scan statistic as applied to **Model 4**. 167
Figure 5.8. The standardised clustering measure $D' = K'(s,\tau)–K(s)K(\tau)$ as applied to lightning strike location data 170
Figure 5.9. Various angles of the spatio-temporal domain showing the point events contained within the inclined elliptic space-time scan statistic as applied to the real-world lightning strike data. 172

Chapter 6:

Figure 6.1. Illustration of selecting a thunderstorm movement track. 184
Figure 6.2 Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$, as applied to homogeneous and clustered point event datasets with various temporal weights 186
Figure 6.3. The three components of the distance function used in measuring separation between two line segments 189

Appendices:

Figure E1. Reachibility distance calculation between core and other trajectories. 223
Figure H1. Additional angles of the 3D spatial extent and propagation of the lightning strike occurrences plots. 228
Figure H2. Additional angles of the 3D spatial extent and propagation of the lightning strike occurrences plots 229
Figure I1. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for thunderstorms 1A, 1B, 1C, 2A, 2B and 2C 230

Figure J1. Subplots showing the cutting of inter-event time Δτ time-series in real-time 231
Figure K1. Least-squares plane-fit solution to the lightning strike datasets for storms 232
Figure L1. Storm 1A (28 June 2012, n = 7,001 events, v = 56 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...233
Figure L2. Storm 1B (28 June 2012, n = 9,901 events, v = 48 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...234
Figure L3. Storm 1C (28 June 2012, n = 4,701 events, v = 60 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...235
Figure L4. Storm 2A (1 July 2015, n = 4,019 events, v = 101 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...236
Figure L5. Storm 2B (1 July 2015, n = 390 events, v = 66 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...237
Figure L6. Storm 2C (1 July 2015, n = 1,469 events, v = 107 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space...238
Figure M1. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1B.....239
Figure M2. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1C.....239
Figure M3. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 2A 240
Figure M4. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 2B.....240
Figure M5. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1C.....241
Figure N1. Sampling, sorting and re-indexing procedure based on the combined supercell inter-event time Δτ dataset and thunderstorms 242
Figure O1. The positions of N = 1,000 point events as a set of regular grid data in a {X = 20, Y = 20, T = 20} space-time volume 243
Figure O2. Three subplots of the clustering measure $D = K(s,\tau) - K(s)K(\tau)$ as applied to regular grid point event data.....244
Figure O3. Various spatial distances that can be measured from an arbitrary point event (coloured in red) for the regular grid dataset with the set spacing of 2 spatial units. 244

Figure O4. Various angles of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to regular grid data 246
Figure P1. The point events counts per 0.1 h of temporal separation for the reference three thunderstorm model (**Model 4**) and its variations. 248
Figure P2. The two-dimension spatial view of the point events as scatter point for the three thunderstorm reference model (**Model 4**) and its variations 249
Figure Q1. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **Model 4** and its variations 250
Figure Q2. 3D views looking perpendicular to the movement direction of the spatio-temporal domain showing the elliptic space-time volume with an inclination height..... 251

List of Tables

Chapter 1:

Table 1.1. Variables and their definitions for the three-dimensional feature space.....3

Chapter 2:

Table 2.1. Subcategories of the systematic literature review and their descriptions. 18

Chapter 3:

Table 3.1. Variables and their values used for the assignment procedure of individual lightning strikes to the six thunderstorms.....51
Table 3.2. Ranges for spatial and temporal coordinates for the six thunderstorms. 61
Table 3.3. Results of least-squares plane-fit and linear-fit solutions to lightning strike data. 76
Table 3.4. Summary of lightning strike characterisation results.....78
Table 3.5. Variables and their definitions for Model 1 and Model 2 dataset generation.....91
Table 3.6. Variables and their values for the generation of Model 3 and its variations.....92
Table 3.7. Variables and their values for the generation of Model 4 and its variations.....94
Table 3.8. Variables and their definitions for synthetic trajectory dataset generation.....97

Chapter 4:

Table 4.1. Identified spatial and temporal ranges $\{s_{MAX}, \tau_{MAX}\}$ for maximum clustering measure D values for **Models 3, 4** and their variations (**Models 3.1, 3.2, 3.3, 4.1, 4.2, 4.3)**. 112
Table 4.2. Input parameters and their values for the space-time scan statistic for **Model 2**. 121
Table 4.3. Identified input parameter values by the elliptic space-time scan statistic on clustered point event data (**Model 2**).....122
Table 4.4. Input parameters and their values for the space-time scan statistic for single storm reference Model **3** and its variations (**Models 3.1, 3.2, 3.3**).....123
Table 4.5. Identified input parameter values by the elliptic space-time scan statistic on single storm reference **Model 3** and its variations (**Models 3.1, 3.2, 3.3**). 125

Table 4.6. Input parameters and their values for the space-time scan statistic for single storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**). 126
Table 4.7. Identified input parameter values by the elliptic space-time scan statistic on three storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**). 128
Table 4.8. Input parameters and their values for space-time scan statistic on the real-world dataset. 138
Table 4.9. Identified parameter values for elliptic space-time scan statistic on the real-world dataset. 139

Chapter 5:

Table 5.1. Input parameters and their values for the inclined elliptic space-time scan statistic for single storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**). 166
Table 5.2. Identified input parameter values by the inclined elliptic space-time scan statistic on three storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**). 168
Table 5.3. Input parameters and their values for the inclined elliptic space-time scan statistic as applied to the real-world dataset. 171
Table 5.4. Identified parameter values for the inclined elliptic space-time scan statistic as applied to the real-world dataset. 172

Appendices:

Table A1. Scientific journals with multiple critically reviewed publications. 204
Table A2. Critical literature review database of 143 publications on spatio-temporal clustering methodologies. 205
Table B1. Variables and their definitions for K-function derivation 218
Table C1. Variables and their definitions for kernel density estimation derivation. 219
Table D1. Variables and their definitions for space-time scan statistic derivation. 220
Table E1. Variables and their definitions for Trajectory-OPTICS derivation. 222
Table G1. Summary table of publications on using spatio-temporal point process models to model natural hazard processes. 226
Table O1. Input parameter values for space-time scan statistic on regular grid data.. 246

Abstract

Natural hazards are inherently spatio-temporal processes. Spatio-temporal clustering methodologies applied to natural hazard data can help distinguish clustering patterns that would not only identify point-event dense regions and time periods, but also characterise the hazardous process. In **Chapter 2**, spatio-temporal clustering methodologies applicable to point event and trajectory datasets representative of natural hazards are reviewed by critically examining 143 scientific publications from various fields of study. These methodologies include clustering measures that are either (i) global (providing a single quantitative measure of the degree of clustering in the dataset) or (ii) local (i.e. assigning individual point events to a cluster). A common application and analysis framework of combining global and local measures for application to point event data is proposed. For global measures, K-functions analysis and for local measures, a space-time scan statistic with kernel density estimation as an aiding methodology within the framework are selected. For trajectories, a density-based local clustering measure Trajectory-OPTICS is selected. In **Chapter 3**, to assess the performance of the methodology framework, real-world natural hazard data and synthetic datasets, either representative of natural hazards or used as performance benchmarks for application, are presented and characterised. A point event dataset of 12,521 lightning strikes recorded on 1 July 2015 over the UK is selected, where a severe three-storm system crossed the region with different convective modes. It is also used as a case study together with a dataset of 77,252 lightning strikes on 28 June 2012 over the UK to characterise and model lightning strikes as point events produced by a moving source. Each source has a set number of points events, initiation point in space and time, movement speed, direction, inter-event time distribution and spatial spread distribution. Movement speed, inter-event time and spatial spread distributions are characterised based on the two case studies. Inter-event time values range from below 0.01 s to over 100 s for individual storms from both case studies. A least-squares plane fit in the spatio-temporal domain estimates a range of representative movement speed values of 47–60 km h^{-1} for the first and 66–111 km h^{-1} for the second case study. Based on these values, single (Model 3) and three storm (Model 4) models are generated to form a simulation study of point event datasets representing various physical lightning characteristics, each with three variations in their movement speed and spatial spread input parameters. For trajectories, the Atlantic hurricane database (HURDAT2) is used to select a real-world dataset of 316 hurricanes. Homogeneous and clustered trajectory datasets are generated as benchmarks for Trajectory-OPTICS. In **Chapter 4**, the clustering methodology framework identified in **Chapter 2** is applied to all the real-world and synthetic datasets presented in **Chapter 3**. K-function analysis results are used to inform the range of bandwidth values for the kernel density estimation. A leave-one-out estimator is used to find the optimal values. A value threshold on the probability density values from the kernel density estimation is imposed to identify high probability density space-time volumes. These volumes are used as centroids for applying the scan statistic as a local clustering measure. The elliptic scan statistic is unable to identify individual lightning strike clusters within the same storm source for storm sources with small temporal separation (Model 4). **Chapter 5** extends the elliptic scan statistic by including an 'Inclination height' parameter as the temporal distance between the major axis points of the ellipse basis. With detailed selection of input parameter ranges, the inclined elliptic scan statistic is applied to Model 4 and its variations and is able to identify point event cluster produced by a moving source and the point events assigned to the cluster are from the same storm source.

Acknowledgements

I would like to thank Professor Paul Bates and Professor Mário Gonzalez Pereira who agreed to be the examiners of this thesis.

Throughout the years of working on this research and thesis, I would like to first of all thank my supervisors — Professor Bruce D Malamud and Dr Davide Pigoli. Without their continued support, encouragement and advice, the hardest periods of this thesis would have been much tougher to overcome. Working with them has not only made me appreciate the dedication, attention to detail and critical approach to learning with relation to scientific research, but also apply these qualities in any other endeavours. It was a privilege to be supervised by them.

I would like to thank all the people I encountered and shared experiences with in the Department of Geography and the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES). The latter of which I am thankful for funding my studentship and allowing me the opportunity to study in London, surrounded by inspiring people, events, locations and stories. Although interrupted by unforeseen world events, the time I shared with some many people within King's College London enriched my life in ways I probably still have not understood.

A personal thank you to every single person in Cohort IV of CANES. I appreciate the shared experiences during the MSc year and the PhD, the late working hours, the dedication, creativity and friendliness. Special thank you to my Bow group – Daniel, Davide and Giuseppe. I look back fondly on all the shared experiences, cycling trips, basketball games, Italian food recipes, everyday struggles and laughs. Also thank you to Ad Maiora for taking care of Og. Although still doing a lot of thinking and calculations, these shared evenings were exactly what I needed after long days working on this thesis.

Vislielākais paldies gan šī darba, gan visas dzīves garumā ir manai ģimenei. Bez viņu atbalsta gan attālināti, esot Londonā, gan klātienē, jau strādājot no mājām, nekas no šī nebūtu iespējams. Personīgs paldies pienākas ļoti daudziem cilvēkiem, kas atcerējās un pastūma mani uz priekšu tieši grūtākajos brīžos. Īpašs paldies Mārim par pārgājieniem Verdaņskā mājsēdes laikā un alus degustācijām pēc tās, Kasparam un Maksim par gatavību doties lekt aukstā ūdeni gada tumšākajā laikā un Reinim par staigājošām diskotēkām un visu Jūrmalas grants ceļu iepazīšanu.

1. Introduction

A natural hazard is defined as a natural process or phenomenon that may cause loss of life, injury or other negative health impacts, damage to property, loss of livelihoods, social, economic and environmental disruption [**UNISDR, 2009**]. Natural hazards can be divided into six distinct categories [**Gill and Malamud, 2014**]: geophysical, hydrological, shallow earth processes, atmospheric, biophysical and space hazards. Furthermore, each natural hazard can be described by its location, time of occurrence, intensity, frequency and probability. This thesis investigates methodologies for examining spatial-temporal clustering in the context of natural hazards. No preference for a particular category of natural hazard is made, rather, the relevancy of various hazards and what spatio-temporal clustering methodologies can be applied will be assessed based on their data type.

As the initiation and propagation of natural hazards pose several risk factors for both urban and natural environments, resulting in various economic, environmental and human costs, better understanding the spatio-temporal patterns of different types of natural hazards can aid us to quantify that risk better. The patterns of how point events (e.g. earthquakes) and trajectories of events (e.g. hurricane tracks) cluster in time and space inform the physical properties of these initiation and propagation processes. These processes can be characterised and modelled by examining these patterns over various spatial and temporal scales.

For example, in [**Camargo *et al.*, 2007**], tropical cyclone track data collected over several years can reveal groups of cyclones with similar movement patterns (i.e. similar genesis positions, angles of movement or duration as physical properties of the cyclones as a natural hazard). When a new tropic cyclone is identified as belonging to any of the previously found groups, the historical information on the physical properties of tropical cyclones belonging to this group can be used for forecasts. Furthermore, faster identification of the physical properties can improve the forecasts and aid in risk alleviation efforts on the ground as the more vulnerable areas and the temporal intervals of higher hazard intensity could be identified in advance.

Similar examples can be found for other types of natural hazards as point events (e.g. clustering of induced seismic events [**Leśniak and Isakow, 2009**] allows to improve mining safety procedures, clustering of lightning strike locations [**Strauss *et al.*, 2013**] can be used to track

convective events and clustering of wildfire events [**Tonini *et al.*, 2017**; **Parente *et al.*, 2016**] can identify triggering factors for ignitions and areas at risk and inform fire prevention policy decisions).

This chapter is organized as follows: **Section 1.1** provides a motivation for the limitations imposed on the types of hazards, and thus methodologies than can be applied to them. The discussion centres on the definition and description of the two selected data types – point events and trajectories – their relevancy to natural hazards as well as the definition and importance of clustering in the combined spatio-temporal domain. Afterwards, **Section 1.2** lists and expands on the details of the aims and objectives of this thesis. Each objective is presented with the relevant research directions undertaken for their completion and the relevant research results. Finally, **Section 1.3** provides an overview of how the rest of this thesis has been organized.

1.1. Motivation

There are various methods to perform clustering of events as recorded from a particular natural hazard process. Natural hazards intrinsically impose various properties on the data types and, thus, the clustering methodologies that can be applied to these data types. These properties can be understood by considering the types of natural hazard data that can be gathered, the features of the hazard process recorded during the data collection phase and the desired outcomes of the clustering process.

Depending on its complexity, geographical data can be broadly categorised into five distinct groups [**Kisilevich *et al.*, 2009**]:

- spatio-temporal point events,
- geo-referenced variables,
- geo-referenced time series,
- moving objects,
- trajectories.

The focus of this research is on the identification of clusters in space and time, so we will restrict our attention to the type of data where the position in space and time of occurrence of the natural hazard are the only variables of interest. Thus, within the scope of this thesis, *spatio-temporal point events* and *trajectories* were selected so as to have two distinct data types, each representative of several natural hazard categories, allowing the analysis framework to be

generalised to different natural hazards. *Geo-referenced variables* and *geo-referenced time series* were not included in this thesis as they observe a given process property at a static spatial location (e.g. meteorological phenomena at a grid point), thus extending the dimensionality of the data but without fully utilising the spatial domain due to the fixed locations of the variable data. *Moving objects* lack historical data on the movement track, reducing the completeness of the available spatio-temporal data, thus *trajectories* were chosen.

For *spatial-temporal point events*, which are used in this thesis, only data in the 2D space and 1D temporal dimension are considered. Additional properties (e.g. earthquake epicentre depth, wildfire cause of fire) could add additional dimensions for modelling the hazard process [**Serra *et al.*, 2014**; **Xu and Schoenberg, 2011**; **Møller and Díaz-Avalos, 2010**], but this can create a level of complexity not appropriate for the research framework of this thesis. Implications of the additional dimensionality are discussed throughout the thesis where appropriate.

To visually illustrate the differences between spatio-temporal point events and trajectories, both data types are constrained to a three-dimensional feature space – with one temporal and two spatial dimensions. The variables and their definitions necessary for this section are shown in **Table 1.1**.

Table 1.1: Variables and their definitions for the three-dimensional feature space (one temporal and two spatial dimensions).

Variable	Definition
X	Length of the first spatial dimension
Y	Length of the second spatial dimension
T	Length of the temporal dimension
x	Spatial coordinate along $[0, X]$
y	Spatial coordinate along $[0, Y]$
t	Temporal coordinate along $[0, T]$
s	Source producing an event

While both data types incorporate the location and time of occurrence information of the individual events as recorded for the particular natural hazard process, the sources producing the individual events have varying properties. The definitions of the two data types and their examples as applied in this thesis are as follows:

- *Point events*: a spatially fixed source $s(x, y)$ produces an individual point event at time t (e.g. lightning strike location (**Figure 1.1b**), earthquakes, tornado touchdown locations, wildfires).
- *Trajectories*: a moving source $s(x(t),y(t))$ produces a point at time t. A track connecting such points spatially for the duration of the moving source is defined as an individual trajectory event (e.g. hurricanes (**Figure 1.2b**), tornadoes).

Examples of both synthetic and real-world data are shown in **Figures 1.1** and **1.2** for the point event and trajectory data types, respectively. The synthetic data are generated in a three-dimensional (two spatial and a temporal dimension) $\{X, Y, T\}$ space-time volume (further details on various approaches to synthetic point event dataset generation are in **Section 3.4**), while for real-world data, datasets of single-day (1 July 2015) lightning strike locations over the United Kingdom [**Gaffard *et al.*, 2008**] and single-year (2005) North Atlantic hurricane tracks [**Landsea and Franklin, 2013**] are displayed (further details on these real-world datasets are in **Section 3.2**).

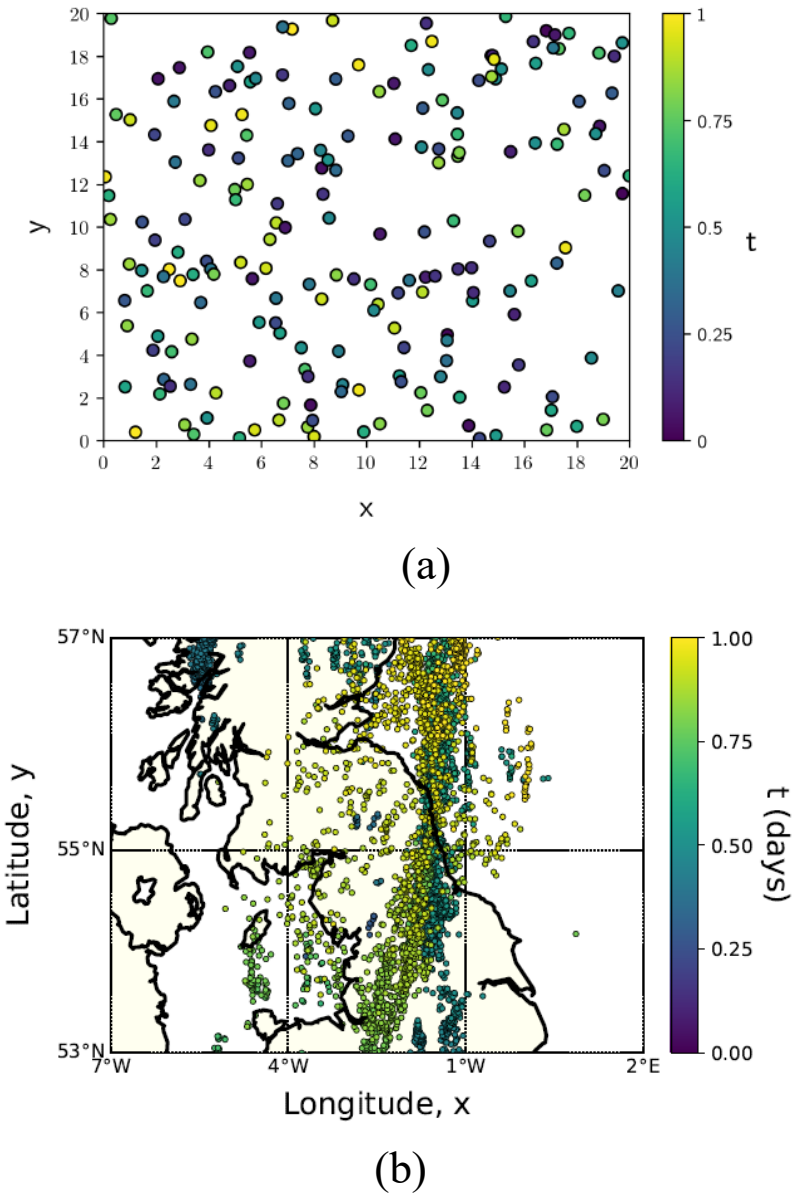

Figure 1.1. Examples of **(a)** synthetic and **(b)** real-world point event data with points from both data types defined by their coordinates (x, y) on a two-dimensional space. The colour of the data points indicates their time of occurrence t. For synthetic point event data, the points are defined by their location on a $\{X = 20, Y = 20, T = 1\}$ space-time domain. The real-world data example is represented by lightning strike locations for a single day of increased lightning activity (1 July 2015 with time coordinate $t = 0$ corresponding to 00:00 UTC) over the United Kingdom.

For the point event data type in **Figure 1.1**, each event is static in both spatial and temporal domains and thus can be fully described by its coordinates $\{x,y\}$ and its time of occurrence t. In the real-world example, the point events seem to produce an elongated pattern in the SW–NE direction. While such systems are moving sources, the tracking information to identify individual trajectories connecting the point events is unavailable. The assumption of the thunderstorm being a spatially fixed source at the moment of a lightning discharge is thus made.

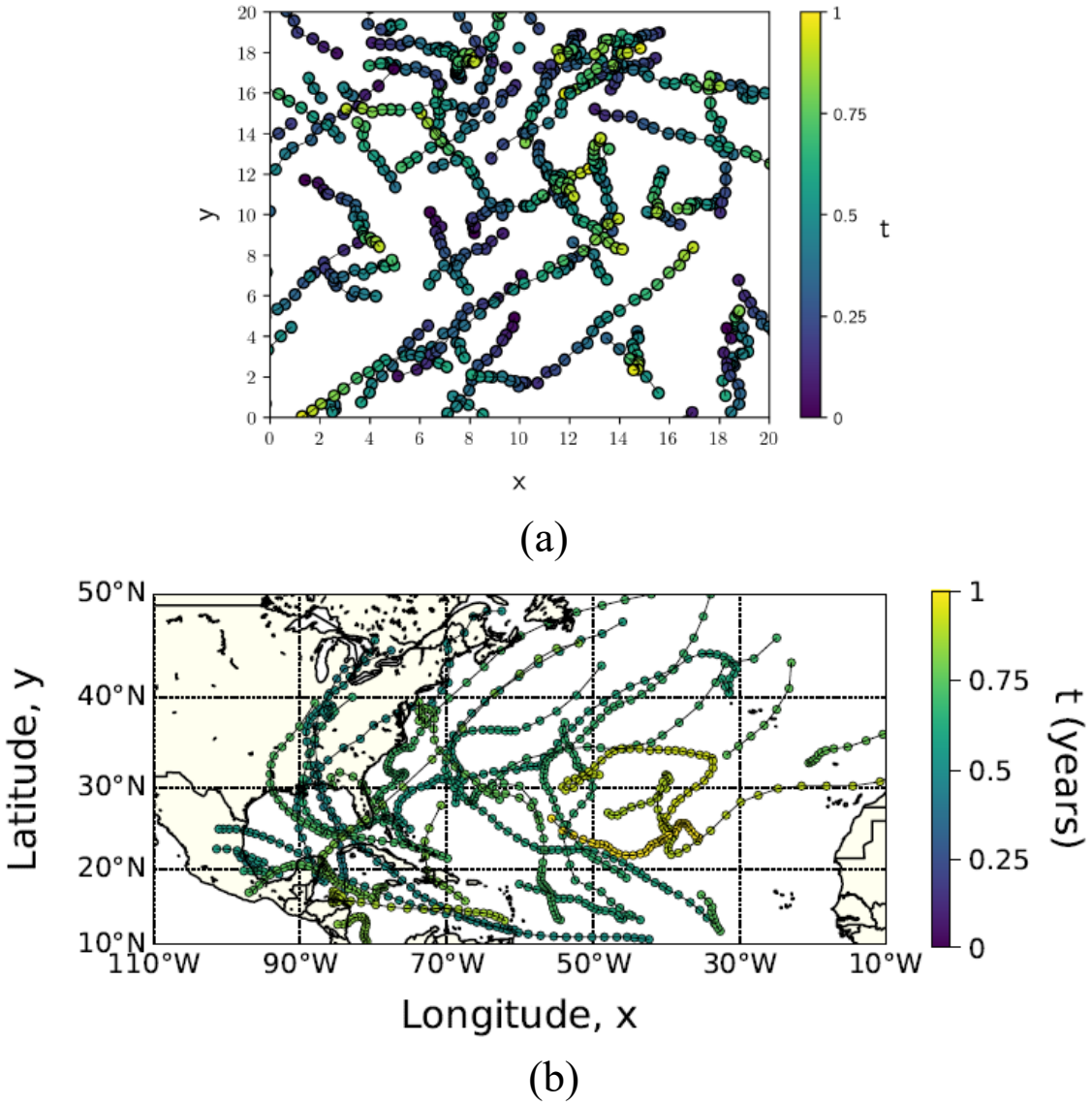

Figure 1.2. Examples of **(a)** synthetic and **(b)** real-world trajectory data with trajectory tracks for both data types defined by their time-dependent coordinates $x(t)$, $y(t)$ on a two-dimensional space. The colour of individual points along a track indicates their time of occurrence t. For synthetic trajectory data, the tracks are generated in a $\{X = 20, Y = 20, T = 1\}$ space-time domain. The real-world data are represented by North Atlantic hurricane tracks for year 2005 [**Landsea and Franklin**, **2013**].

For the trajectory data type in **Figure 1.2**, each event can be considered a collection (i.e. a track in the context of natural hazards) of point events for a particular moving source producing them. Hence, additional information can be gathered from the spatial and temporal separation between consecutive points along the tracks and how this separation evolves with time. Additionally, any potential differences in the initial positions, general or instantaneous direction of movement and total duration (i.e. the total number of individual point events within each track or the total temporal length) of the trajectory tracks add further complexity. These

intricacies can be seen in both the synthetic and real-world examples as the tracks are initiated at various points in both the spatial and the temporal dimensions, evolve in different directions with respect to their initial positions with varying spatial separations and have different temporal length (total duration) values.

The complexity of the two data types illustrates the importance of considering the spatial and temporal domains together, instead of applying clustering methodologies in either domain separately. If considered separately in the context of natural hazards, each domain can give insight into the different properties of the hazardous process producing the events. Nonetheless, as natural hazards are inherently spatio-temporal processes, combining the two domains allows one to gain a more sophisticated insight and characterisation of the extent and evolution of the hazardous process. A visual representation of the value of combining the two domains with the various properties that can be extracted is illustrated in **Figure 1.3**.

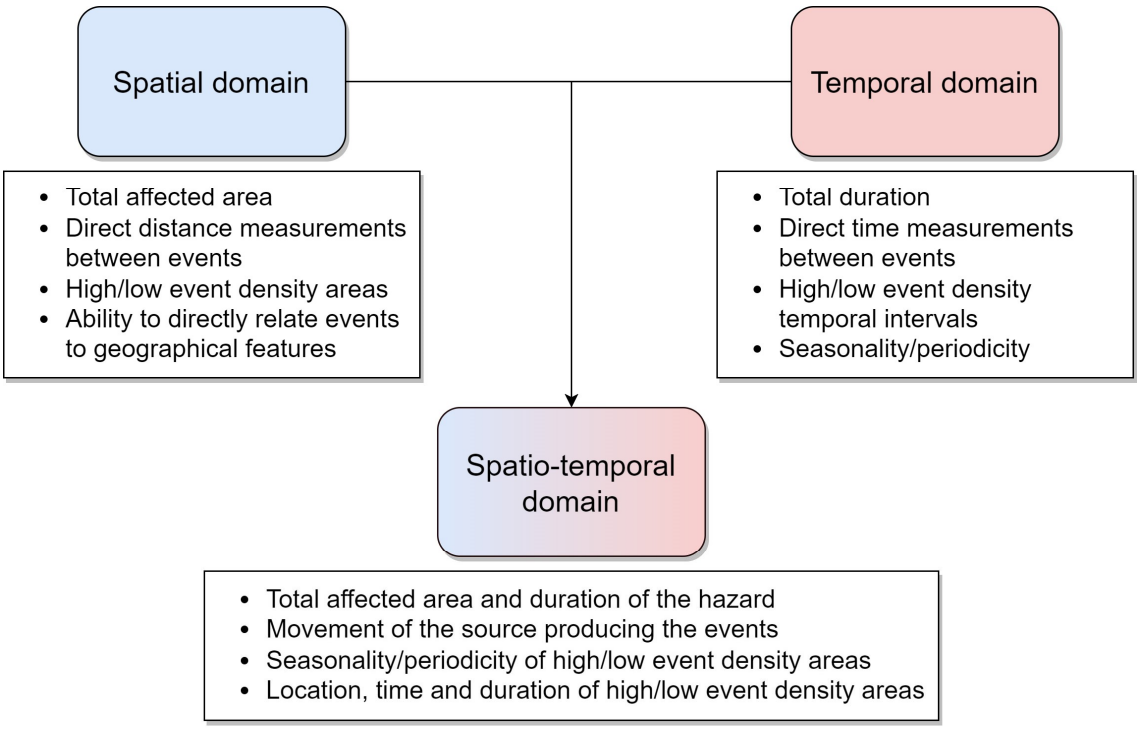

Figure 1.3. Various physical properties of the natural hazard process that can be extracted when analysing the natural hazard data in either spatial or temporal domains separately or in the combined spatio-temporal domain.

1.2. Aim and objectives

The research aims of this thesis are

- *to synthesise into a common framework and apply spatio-temporal clustering analysis methodologies*;
- *to develop spatio-temporal models to reproduce observed natural hazard clustering patterns*.

To achieve these research aims, the research workflow is divided into four main objectives and various research directions necessary to accomplish each objective. After the completion of each individual objective, the research results of the objective can be gathered as the final outcome or as the necessary research direction for the completion of the next objectives. A visual summary of the research work flow for this thesis is shown in **Figure 1.4**.

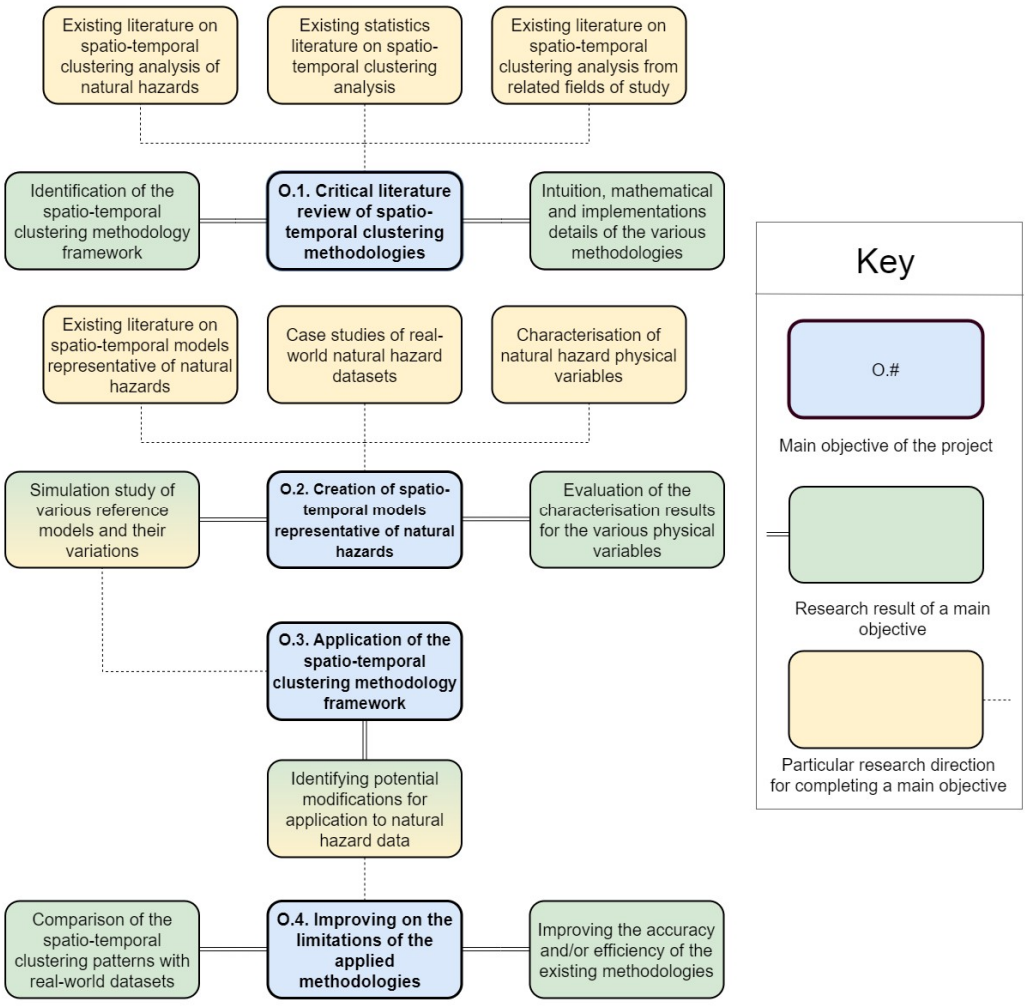

Figure 1.4. Structure and workflow of the research done during this thesis. The main individual objectives, the research directions taken to accomplish them and the research results after completion of each objective are shown with their interconnections coloured and marked as per the figure key.

To provide a more detailed overview with additional context and particular research activities for each objective **(O.1)** – **(O.4)**, an expanded list of objectives is given below:

(O.1) Critical literature review of spatio-temporal clustering methodologies.

- The critical review process includes identifying spatio-temporal clustering methodologies for natural hazard application, but the focus is placed on identifying spatio-temporal clustering methodologies which are general (i.e. designed for the particular data type) and applicable to any field of study.
- This allows an overview of spatio-temporal clustering, acknowledging other fields of study and focusing on the given data types instead of a particular natural hazard.

(O.2) Creation of spatio-temporal models representative of natural hazards.

- To create a spatio-temporal model representative of natural hazards, it is necessary to overview the available models implemented for the given data type for natural hazard modelling.
- For this objective, lightning strikes as point events produced by a moving source are chosen as the natural hazard process.
- Two case studies are selected to use real-world lightning strike data to characterise three physical variables describing the natural hazard process.
- These variables are used as input parameters for a spatio-temporal point event model.

(O.3) Application of the spatio-temporal clustering methodology framework.

- A set of synthetic datasets representative of lightning strikes from objective (**O.2**) is selected to form a simulation study together with purely mathematical models to represent homogeneous and clustered point datasets.
- For both the point event and trajectory data types, these sets of synthetic datasets allow one to assess and evaluate the performance of the individual spatio-temporal clustering methodologies.

(O.4) Improving on the limitations of the applied spatio-temporal methodologies from objective (**O.3**).

- The improvements to the existing spatio-temporal clustering methodologies are presented and applied to the same simulation study and real-world datasets.

1.3. Structure of the thesis

To address the four individual objectives of this research, this thesis is structured into the following chapters and their subsections:

- **Chapter 2. Critical review of spatio-temporal clustering methodologies**:- Presents a critical review of spatio-temporal clustering methodologies. Justification for the approach to the general critical literature review and its results (**Section 2.2**) are given in the context of completing objective (**O.1**). The identified main spatio-temporal methodologies for the two selected types of data are presented in mathematical and implementation detail (**Section 2.3**). Categorisation of the identified clustering methodologies into two broader groups is discussed and presented (**Section 2.4**). Such categorisation leads to a proposed and implemented spatio-temporal clustering framework (**Section 2.5**). **Research results:** The spatio-temporal clustering methodologies relevant to the point event and trajectory data types are identified. The methodology research timeline, individual methodology components, implementation intuition and application nuances to the given data type are analysed and presented. Potential categorisation of the methodologies is given based on their properties and ability to assess the presence of spatio-temporal clustering. The ability of the methodologies to complement each other and be used as part of a general spatio-temporal clustering framework is assessed.

- **Chapter 3. Natural hazard datasets: examples, lightning characterisation, and synthetic data**: This chapter addresses objective (**O.2**). The chapter first provides examples of real-world point event and trajectory datasets (**Section 3.2**). Then two case studies of increased lightning activity are detailed and used to characterisation various lightning strike physical variables and use them as input parameters for a synthetic spatio-temporal point event model representative of lightning strikes (**Section 3.3**). The characterisation results are contextualised with value ranges found in other studies and implementation details (i.e. definitions of all input parameters, their selected values and pseudocode where necessary) of the model are provided. Using purely mathematical models and initiations of the lightning strike point event model with various input parameter combinations, a simulation study is presented (**Section 3.4**). For trajectories, two approaches are detailed to generate homogeneous and clustered trajectory datasets.

Research results: The characterisation of the lightning strike physical variables gives an insight into the physical properties (i.e. range of values, empirical distribution, variations between different storm types) of individual thunderstorms producing the lightning strikes. The resultant spatio-temporal point event model provides a simple tool for generating synthetic datasets that are representative of point produced by a moving source. By using the physical variable ranges found during the characterisation process, this model then generates a set of synthetic datasets representative of lightning strikes.

- **Chapter 4: Application of the spatio-temporal clustering methodologies to real-world and synthetic natural hazard datasets.** This chapter addresses objective (**O.3**), presenting the results and analyses after applying the selected methodologies from **Chapter 2** to the simulation study and real-world datasets as described in **Chapter 3**. The results and analyses are separate for the point event simulation study (**Section 4.2**), synthetic trajectory benchmarks (**Section 4.3**), real-world point event data (**Section 4.4**) and the real-world trajectory data (**Section 4.5**). The advantages and limitations of the selected methodologies and the methodology framework are discussed and evaluated in the context of the presented results in each section. **Research results:** The application of the spatio-temporal clustering methodology framework to the simulation study and real-world datasets identifies limitations of the framework in terms of clustering accuracy and efficiency. These limitations are examined and their sensitivity to the physical variables (i.e. input parameter of the model) is evaluated.

- **Chapter 5. Improvements on the spatio-temporal clustering methodologies for application to point events produced by a moving source.** This chapter addresses objective (**O.4**), presenting the improvements to the clustering measures that can be made to improve their performance on the synthetic and real-world point event data from **Chapter 3**. The improvements are presented separately for the global (**Section 5.2**) and local (**Section 5.3**) clustering measures. These measures are again applied to the synthetic (**Section 5.4**) and real-world (**Section 5.5**) datasets to evaluate the changes in the clustering results and their implications for the spatio-temporal clustering framework (**Section 5.6**). **Research results**: Any improvements in the accuracy and performance of the individual methodologies and the methodology framework are

analysed. The implications of these changes are discussed, any further limitations of methodologies are noted, and potential further research directions are proposed.

- **Chapter 6. Summary, conclusions and future research directions:** This chapter focuses on synthesising the research results of the thesis chapters by contextualising them within the overall aims and the individual research objectives of the thesis **O.1** – **O.4** (**Section 6.2**). Identifies potential further research directions (**Section 6.3**).

For the four substantive chapters of this thesis, **Chapters 2** to **5**, a summary of each chapter is given at the beginning of each, along with a set of questions that the chapter aims to address through the research results. This thesis also has 17 appendices (**Appendices A**–**Q**), which are referred to at the appropriate locations throughout the thesis. These appendices aim to either provide additional context to sections of the thesis or include additional results that did not require further analysis in the main results.

2. Critical review of spatio-temporal clustering methodologies

Chapter summary

This chapter presents a critical literature review for clustering methodologies, which are applicable to two selected natural hazard data types. The steps taken to complete the critical review and the identified relevant subcategories (e.g. pre-processing steps, cluster shape) are noted and detailed. Focus is not on methodologies specifically applied to natural hazards, but the data types themselves. Examining the frequency of different methodologies and their development timeline aids an audience of natural hazard researchers to better understand the wider range of methodologies available for the two selected spatio-temporal data types. A total of 143 publications are critically reviewed and selected for inclusion in the literature database. Various review subcategories are created to synthesise the review results. The review results are visualised (e.g. timeline of the research intensity, frequency of various fields of study). The major methodologies are detailed in terms of their mathematical intuition and implementation, with full details given in the appendices. For point events, categorising the methodologies into global and local clustering measures is justified. The exact details for the terminology of global and local clustering measures are discussed and a common methodology application framework of K-functions, kernel density estimation and space-time scan statistic is presented. The existing categorization of trajectory clustering methodologies is noted for trajectories, and the density-based clustering methodology category is selected for analysis and evaluation.

This chapter aims to answer the following questions:

Q2.1. What spatio-temporal clustering methodologies are available for the particular data types – point events and trajectories?

Q2.2. What is the usage frequency of various spatio-temporal clustering methodologies?

Q2.3. Which methods are well-established, and which are still being developed (i.e. publication timeline)?

Q2.4. Which fields of study apply clustering methodologies in the spatio-temporal domain?

Q2.5. Are there particular methodologies that have been extensively used in only particular fields of study?

Q2.6. Can the reviewed methodologies be categorized in broader groups depending on their application purpose and used in a common framework?

2.1. Introduction

This chapter presents a critical literature review of spatio-temporal clustering methodologies applicable to the point event and trajectory data types. **Section 2.2** presents the justification for the approach to the general critical literature review of spatio-temporal clustering methodologies and its results. The results of the literature review are given in terms of the reviewed fields of study in which the relevant methodologies are found, their prevalence through various fields of application, distribution in time (i.e. research timeline) and overall frequency of use in literature. Further analysis is presented on applying methodologies to various synthetic and/or real-world datasets and following the interpretation of the results. For the spatio-temporal clustering methodologies selected for further application, the results of the critical review are presented separately for the two selected data types — point events (**Section 2.3.1**) and trajectories (**Section 2.3.2**) — with focus on the most common methodologies.

After reviewing the most common and relevant methodologies, they are compared and categorised as two larger groups of spatio-temporal clustering methodologies for point events (**Section 2.4**). The global clustering measure group comprises methodologies that measure spatio-temporal interaction between point events over given spatial and temporal scales. These measures do not identify exact locations in the space-time domain of event clusters, which is the objective of applying local clustering measures. A combined spatio-temporal clustering analysis framework combining the two groups of methodologies to create an efficient and accurate way of examining spatio-temporal point event datasets is presented and discussed in **Section 2.5**.

2.2. General review of spatio-temporal clustering methodologies

The critical literature review approach comprised multiple steps to ensure temporal coverage of the research timeline, various fields of study and identifying different review subcategories to identify commonalities and differences within the methodologies. Within this section, the literature review general approach, keywords, review subcategories and sources are presented in **Section 2.2.1**. The research timeline of the methodologies over various fields of study and application is presented and discussed in **Section 2.2.2**. Analysis of the application of the review clustering methodologies to various synthetic and/or real-world datasets is presented in **Section 2.2.3**.

2.2.1. Critical review approach

To complete the general review and identify relevant spatio-temporal clustering methodologies, a detailed approach was created to select the sources and relevant keywords for the scientific literature search and any further steps to ensure completeness. The approach consists of the following review steps:

1. **Identifying sources**: Through the review process, relevant literature was gathered through two main sources: Google Scholar™ and Web of Science™. To identify the most common methodologies and their application framework, these two sources provide an overview of the relevant methodologies. Only peer-reviewed academic journals and books, conference papers, theses and dissertations and preprints were considered from both sources.
2. **Selecting keywords:** The relevant keywords were identified that would encompass the relevant literature. Keywords were summarised into three distinct parts to reflect the motivation of this thesis (**Section 1.1**):

    - **focusing on point events and trajectories**: 'point events', 'point patterns', 'point process', 'trajectories', 'moving objects'.
    - **combining spatial and temporal domains**: 'spatio-temporal', 'spatiotemporal', 'space- time', 'spatial-temporal'.
    - **clustering interpretations**: 'clustering', 'hotspots', 'patterns'.

    These various keyword permutations were chosen to cover as wide of a range of keywords as possible. This necessity is due to either different research background (i.e. particular wording might be prevalent in a given field of study) or language differences. While the permutation of the keywords encompassed the motivation of this thesis, their singular vs plural forms were not considered, introducing an initial limitation to the final literature database.
3. **Identifying highly cited and review publications:** Publications were selected that have a 'high' citation count (every publication with at least 100 citations was identified), indicating their importance for the particular methodology and/or field of study. Identifying such publications allowed understanding the research timeline and developments of particular methodologies. A weakness of this approach is that

publications in more recent years would have fewer citations. Review publications, when available in a particular field or for a subset of methodologies/data types, further allowed to synthesise the research timelines of various methodologies and consider various categorisation approaches. While identifying the most common methodologies, such an approach introduces bias towards more active recent fields of study, where the citation numbers would be higher.

4. **Exploring the citation network:** Once the highly cited papers were identified, the publications that led to the development of the clustering methodology and any related methodologies and improvements that have been done since the publication were explored. This approach inherently adds bias towards the particular methodology as additional publications implementing it are selected for the literature database. This is acceptable for this thesis as, once the common methodologies are identified, potential improvements and variations in their implementation are used to help inform the research for **Chapters 4** and **5** (i.e., application of the spatio-temporal clustering methodologies to real-world data, and improvement of the methodologies).

Following the above four critical literature review steps, approximately 300 publications (with estimated 80% from Google Scholar™ and 20% from Web of Science™) were identified and skimmed to identify their relevance. Throughout the review process, publications were discarded due to any of the following limitations:

- **Year of publication**: in the citation networks of the more common methodologies, their development could be traced to publications dated decades before them. As the objective of this thesis is to find currently used methodologies, only the most cited of the older publications were left to provide general context for their development.
- **Separation of spatio-temporal domain**: publications that did not consider the combined spatio-temporal domain but implemented methodologies to spatial and temporal domains separately were discarded.
- **Limited applicability**: methodologies limited to a particular field of application due to their assumptions and implementation were discarded.

A total of 143 relevant scientific publications were identified through this approach and comprised the final spatio-temporal clustering methodologies literature database. These selected publications were further systematically reviewed within a common review framework. The review framework consisted of four main categories:

1. **General information**: in addition to the title, authors, journal and year of publication, the field of study and the field of application were identified. The field of study was chosen based on the journal the publication was in, while the field of application, if the methodology in question was applied, was assessed only based on the datasets analysed in the paper (i.e. a methodology presented in a 'Statistics' field of study could be applied to a dataset of earthquake spatial locations and times of occurrence – 'Earthquakes' field of application).
2. **Pre-processing steps**: before any methodology is applied to data, there are several pre-processing steps in the spatial and/or temporal domains. Identifying and noting such steps aids in understanding the intuition and limitations of the particular methodologies.
3. **Methodology**: individual components of the methodologies are identified (e.g. dataset embedding, clustering method, cluster shape) in order to compare and contrast them. Additional components are identified for particular methodologies and details on these components are given in **Section 2.3.1** and Section **2.3.2**.
4. **Metadata**: it is noted whether the particular methodology has been applied to synthetic and/or real-world datasets. For the real-world dataset applications, any set magnitude thresholds, particular study areas and time period limitations are identified and noted.

Within each of these four categories, a full list of all the subcategories within each and their respective descriptions in the context of this thesis is given in **Table 2.1**. The synthesised literature database on some of the identified review subcategories ('Title', 'Authors', 'Journal', 'Year', 'Field of study', 'Field of application', 'Clustering method', 'Applied datasets', 'Interpretation') is presented in **Appendix Table A2**.

Table 2.1. Subcategories of the systematic literature review and their descriptions.

Review subcategory	Description
Title	Title of the publication
Authors	Publication authors
Journal	Publication journal
Year	Year of publication
Field of study	Research field of the journal
Field of application	Research field of the selected application datasets
Pre-processing	Specifies pre-processing steps necessary for the application of the methodology
Dataset embedding	Changes made to the dimensionality of the data for the application of the methodology
Clustering method	Spatio-temporal clustering methodology
Cluster shape	Shape of the cluster identified by the methodology
Cluster separation	Specifies whether the methodology can identify individual clusters ('+' for 'Yes', '–' for 'No')
Magnitude threshold	Thresholds imposed on the applied dataset
Applied datasets	Specifies whether the methodology was applied to synthetic and/or real-world datasets
Region	Spatial extent of the applied dataset
Period	Time period of the applied dataset
Interpretation	Specifies whether authors provide interpretation of the clustering results ('+' for 'Yes', '–' for 'No')

To assess the coverage of various fields of study within the critical literature review, the list of scientific journals with two or greater publications in the literature database is shown in **Appendix Table A1**. Only three journals have three or more reviewed publications that have been added to the database and each is assigned a different field of study: '*Journal of Geophysical Research*' is categorised as 'Geophysics', '*PLOS One*' as 'Medical Science' and '*Environmental and Ecological Statistics*' as 'Statistics'. A total of 24 journals as listed in **Table A1** have two or more publications, indicating that the reviewed literature database has a widespread coverage of various fields of study.

2.2.2. Spatio-temporal clustering research timeline

As the coverage of various fields of study is established, to understand better the development of the methodologies, particular fields of study that lead the development and have been using the particular methodologies more frequently are identified. Initial evaluation of the temporal coverage of the critically reviewed literature database, temporal distribution of the publication years for the reviewed fields of study and application are given in **Figure 2.1**.

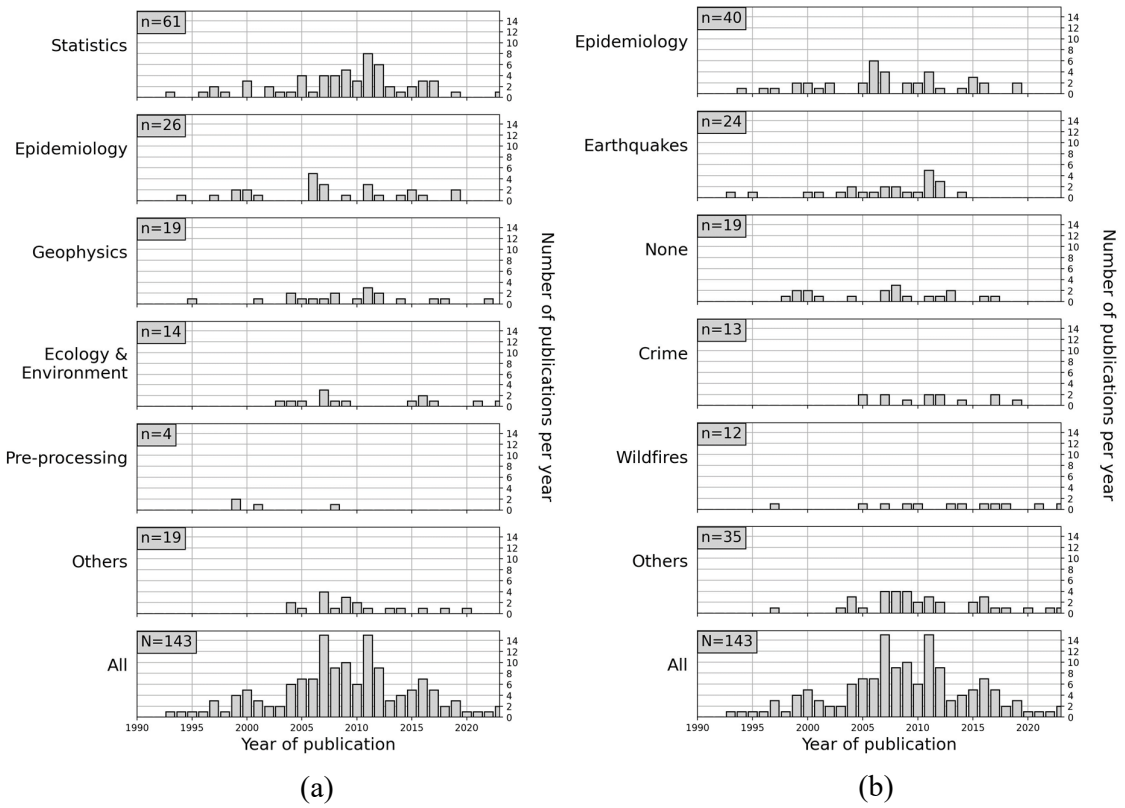

Figure 2.1. Critical review of spatial-temporal clustering for publications from 1992–2022 (N = 143 publications): **(a)** fields of study (general field of the journal the publication was published in) and **(b)** fields of application (field of the datasets used for application). The five fields for (a) and (b) are arranged by the total number of publications in each field, with only the five highest displayed. Any other fields of study/application are marked as 'Others', while the cumulative numbers of publications per year are displayed as 'All' in the lowest subfigures. The total number of publications within each field *n* is displayed at the upper left corner of each subfigure.

From the temporal distribution of all publications (lower left and right subfigure with 'All' in **Figure 2.1**), it can be seen that the publication period is from 1992–2022 with 143 publications. Only 24 publications cover the 12–year period from 1992–2003 (2 publications/year), indicating few publications that mention particular aspects of the methodologies during this period. The majority of the publications (114 out of 143) are over the 16–year period 2004–2019 (7 publications/year).

For both the fields of study and application, the five most common fields are singled out for each and the temporal distribution of publications within each is presented in separate subfigures within **Figure 2.1**. For fields of study, 'Statistics' was the most common (n = 61 publications) and 'Medical Science' second (n = 27). For the fields of application, the field of 'Epidemiology' has the highest number of publications (n = 40). There are n = 14 publications without any application to datasets. This would imply that a large part of the publications in statistical journals ('Statistics' field of study) have applied the methodologies to epidemiology or have presented just the methodology without an application.

The review did not focus on only natural hazards, but encompassed a wide range of fields of study and application to ensure that the methodologies are not only applicable for specific cases. General spatio-temporal clustering methodologies applicable to the two selected data types (i.e. point events and trajectories) are preferable. As the field of study 'Statistics' was the most common field of study, the generality of methodologies is evident.

The most common general spatio-temporal clustering methodologies identified during the critical literature review as applicable to one of the two data types are presented in **Section 2.3**. Their intuitive objective is summarised, brief description of the implementation steps (with expanded details in the **Appendices C–E**) is provided with a mention of the necessary input parameters and their purpose in their implementation. **Section 2.3.1** focuses on spatio-temporal clustering methodologies applicable to the point event and **Section 2.3.2** to the trajectory data type.

2.2.3. Application of methodologies to various datasets and interpretation

This thesis aims to establish real-world and synthetic point event and trajectory datasets for application of spatio-temporal clustering methodologies. Such a framework allows the analysis and interpretation of clustering results based on the clustering patterns observed and contrasted between the various datasets. The use of real-world and/or synthetic datasets in the critically reviewed publication was a key focus of the review (i.e. the 'Metadata' section of the review framework as listed in **Section 2.2.1**). To examine and present the frequency of each of the various options for the 'Applied datasets' review subcategory (**Table 2.1**), a pie chart detailing the division between real-world data ('Real'), synthetic data ('Synthetic'), both ('Real'), or no application ('None') is shown in **Figure 2.2**.

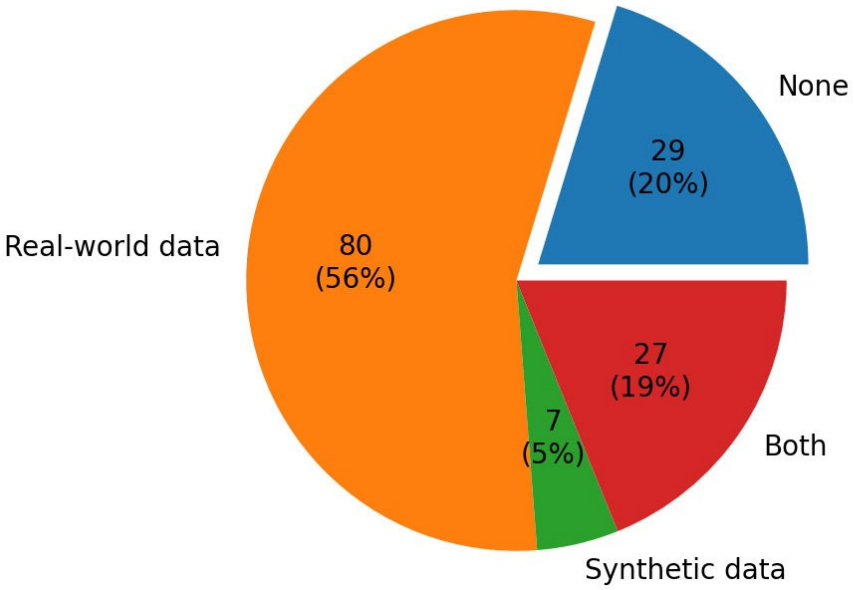

Figure 2.2. Distribution of the 'Applied datasets' review subcategory (**Table 2.1**) after applying the review framework to $N = 143$ publications. Proportion for the four subcategories is given in percentage over each division with the total number of publications for each option given in brackets.

In **Figure 2.2**, the distribution of the 'Applied datasets' review subcategory shows 80% of all publications used data to apply the clustering methodologies. The $n = 29$ publications without application can be attributed to the 'Statistics' field of study, where advancements and changes to methodologies were presented without application. The majority of 56% of the publications used only real-world data, while 5% used only synthetic data. A total of $n = 27$ publications applied spatio-temporal clustering methodologies to both real-world and synthetic data.

This thesis uses synthetic data to establish performance benchmarks through a simulation study (**Section 3.4**). The real-world data (**Section 3.1**) establishes clustering patterns and evaluates their correspondence to the synthetic benchmarks. The interpretation of these results allows to evaluate the methodology performance on 'non-idealised' real-world data compared to the simulation study. To examine the presence of interpretation of the spatio-temporal clustering results, the previous 'Applied datasets' subcategory is further divided by the 'Interpretation' subcategory (**Table 2.1**). The 'Interpretation' subcategory indicates whether the authors further interpreted the clustering results with respect to their implications for the region, temporal scales, seasonality, and other aspects. The results are presented in a joint pie chart in **Figure 2.3**.

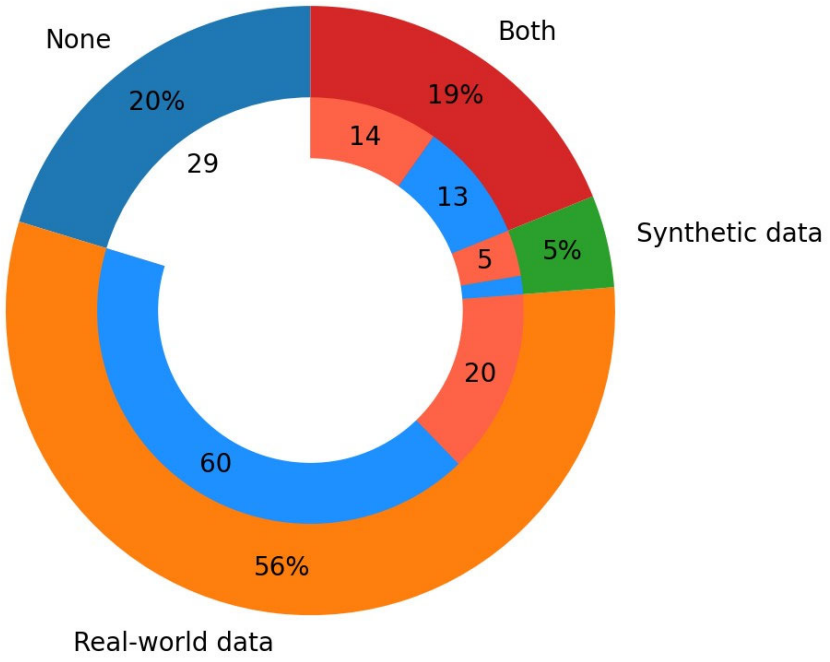

Figure 2.3. Distribution of the 'Applied datasets' and 'Interpretation' review subcategory (**Table 2.1**) after applying the review framework to $N = 143$ publications. Proportion for the four subcategories for 'Applied datasets' is given in the outer pie chart in percentage over each division. The inwards pie chart represents the further division of the 'Interpretation subcategory with blue colour showing that the clustering results have been interpreted ('Yes' or '+') and red showing that they have not ('No' or '–') .

In **Figure 2.3**, the additional 'Interpretation' review subcategory indicates the use of real-world data to interpret the clustering results as 75% (60 out of 80) of the publications provided further interpretation of the clustering results. With only synthetic data, there were only two publications [**Nanni and Pedreschi, 2006**; **Wiegand and Moloney, 2004**] that provided further interpretation of synthetic data generated to incorporate properties of natural hazards. When combining the real-world and synthetic data, only 13 out of 27 publications further interpreted the results, which is one of the objectives of this thesis.

2.3. Spatio-temporal clustering methodologies

This section details the mathematical intuition and implementation details of the most common and the selected spatio-temporal methodologies for point events (**Section 2.3.1**) and trajectories (**Section 2.3.2**). Additional methodologies that were not implemented will be mentioned with their variations from the implemented methodologies discussed. A common methodology framework for spatio-temporal clustering is presented later in **Section 2.5**.

2.3.1. Point events

From the identified N = 143 publications, 114 (80%) presented spatio-temporal clustering methodologies for the point event data type. As review publications were used to understand already established categorisation of methodologies (see the review framework in **Section 2.2.1**), such an uneven split occurred due to the lack of existing review papers for point-event spatio-temporal clustering. Point processes have also been studied extensively and have a wider range of applications due to their generality [**González *et al.*, 2016**]. Each application of point process then focuses on the intricacies for the given field of application. The review papers that were identified for the point events focused primarily on the spatial domain and for a particular field of application (e.g. ecology [**Perry *et al.*, 2006**; **Dale *et al.*, 2002**] and crime [**Adepeju *et al.*, 2016**]). Spatio-temporal clustering methodologies were mentioned as potential extensions to the analysis of only the spatial domain without noting their development timeline. Publications specific to particular methodologies had to be further identified after identifying highly cited and review publications (see **Section 2.2.1**).

2.3.1.1. K-function analysis

As a methodology, K-function analysis measures space-time interaction between point events. Initial measures for estimating space-time interaction were developed as the Knox statistic in [**Knox and Bartlett, 1964**]. The K-function analysis framework is based on the contingency table approach in Knox test. For every given point event, spatial and temporal thresholds are used to assess whether point events fall within these set thresholds of the given event and a 2 × 2 'Yes'/'No' contingency table allows for generation of simplistic evaluation metrics. The concept of a K-function was first developed in [**Ripley, 1976**], but only applicable to spatial point patterns analysis (i.e. estimating spatial homogeneity). An extension of the K-function to the three-dimensional space–time domain as a tool for quantifying space-time interaction is provided in [**Diggle *et al.*, 1995**] and its main objective is to estimate the spatial and temporal ranges around an arbitrary point event over which space-time interaction with other point events is evident and significant.

For application specifically to natural hazard data, the K-function analysis framework has been used for various hazard types – landslides [**Tonini and Cama, 2019**], wildfires [**Tonini *et al.*, 2019**; **Costafreda-Aumedes *et al.*, 2016**; **Hering *et al.*, 2009**] and floods [**Wang *et al.*, 2021**]. The exact implementation details of the K-function analysis framework for analysing space-

time interaction between point events is given in **Appendix B**, but it can be concisely summarised with the following steps:

1. **Selecting the spatial and temporal ranges**: based on the extent of the dataset, sets of values for the spatial range s and the temporal range τ need to be selected. No automated methodologies for their selection were identified during the review framework. Their value ranges depend on the size of the study space-time volume and visual assessment of the dataset, which will be the approach taken in this thesis. These ranges will be used to estimate how spatio-temporal clustering varies with different spatial and temporal range values combinations.
2. **Calculating the spatio-temporal K-function**: for a given combination of spatial and temporal distances, the expected number of point events that are within both of the given spatial and temporal distance values are calculated and return the spatio-temporal K-function $K_{ST}(s,\tau)$.
3. **Calculating the product of spatial and temporal K-functions**: individual K-functions are calculated for given spatial and temporal range values separately ($K_S(s)$ for the spatial and $K_T(\tau)$ for the temporal domain) and their product is returned as the expected spatio-temporal K-function under the assumption of no clustering in the spatio-temporal domain.
4. **Calculating the difference between spatio-temporal K-function and product of spatial and temporal K-functions**: if there is no space-time interaction present between the point events in the dataset, the spatio-temporal K-function $K_{ST}(s,\tau)$ will have a similar value to the product of separate K-functions $K_S(s)K_T(\tau)$, while a large deviation indicates clustering at the given spatial and temporal ranges. The clustering measure $D(s,\tau)$ can therefore be expressed as

$$D(s,\tau) = K_{ST}(s,\tau) - K_S(s)K_T(\tau) \tag{2.1}$$

2.3.1.2. Kernel density estimation

The kernel density estimation is a widely used spatio-temporal clustering methodology which is prevalent in various fields of study and application (e.g. urban studies [**Gao, 2015**], meteorology [**Strauss *et al.*, 2013**], epidemiology [**Zhang *et al.*, 2011**]) and is used to identify space-time volumes with higher probability density values relative to surroundings. In the

spatio-temporal domain, the main implementation difference for this methodology is in the representation of the two-dimensional space as either a bivariate [**Brunsdon *et al.*, 2007**] or two univariate [**Li and Racine, 2007**] kernels. From the total of 12 identified publications to use kernel density estimation as the only or one of many clustering methodologies, 7 of them applied it to crime datasets with both representations used (e.g. univariate [**Hu *et al.*, 2018**], bivariate [**Nakaya and Yano, 2010**]). To allow for more generality in applying the methodology, two univariate kernels were selected for this thesis. A detailed summary for the implementation is provided in [**Hu *et al.*, 2018**] with the main objective to provide an estimate for the probability density of point events over the space-time domain and extract space-time volume of higher density values.

For application specifically to natural hazard data, the kernel density estimation in the spatio-temporal domain has been used for various hazard types: landslides [**Tonini and Cama, 2019**], wildfires [**Tonini *et al.*, 2017**] and earthquakes [**Adelfio and Ogata, 2010**]. The exact implementation details of the space-time kernel density estimation with univariate kernels is given in **Appendix C**, but can be concisely summarised with the following steps:

1. **Selecting a kernel**: for any coordinate in the space-time volume, the probability density value is calculated from all the point event data points, depending on how close they are to the coordinate. The kernel is a weighting function as it takes the difference between the coordinate value and the data point and assigns a probability density value to it. Epanechnikov and Gaussian kernels were considered, but the Gaussian kernel was selected for implementation in this thesis as other publications [**Hu *et al.*, 2018**; **Nakaya and Yano, 2010**] did not identify a significant impact on the density value estimates from the choice of the kernel function. Both kernel types were used when performing initial clustering analysis research for **Chapter 4**, and the lack of impact on the density value estimates was confirmed, but the results are not presented.
2. **Selecting kernel bandwidths**: the difference between the coordinate and the data point is not directly used as input for the kernel, but is divided by a smoothing parameter – kernel bandwidth. A small bandwidth value will be assigned with almost no weight to the further data points, while a larger bandwidth value will 'bring' the events closer. A data-driven approach to maximise the expected leave-one-out density estimator is selected due to its exhaustive computational search for optimal values [**Hu *et al.*, 2018**].

3. **Calculating the probability density**: after selecting the bandwidth values, a probability density is calculated for a grid of points over the space-time volume, generating *voxels* – space-time volumes with appropriate units for which the probability density value is calculated [**Hu *et al.*, 2018**; **Nakaya and Yano, 2010**]. Within the space-time volume, smaller volumes of high probability density values are identified by introducing probability density thresholds as a percentile of the values. As these high probability density value estimates indicate space-time volumes with relatively higher densities of point events, the value of this threshold is dependent on the specific dataset and number of voxels. As part of the spatio-temporal clustering methodology framework (**Section 2.5**), this threshold is adjusted visually to identify a voxel subset for local clustering measures to use and decrease the computational complexity.

2.3.1.3. Space-time scan statistic

The space-time scan statistic methodology was the most widely used spatio-temporal clustering methodology during the critical literature review phase with a total of $n = 27$ publications. From these, 15 were in the field of study 'Medical Science' and 16 were applied to the field of application 'Epidemiology'. While applied to other fields of application, the origin and development of this particular methodology has been mostly within the field of epidemiology to detect emerging disease clusters. After initially introduced for a one-dimensional data setting [**Naus, 1965**], the scan statistic was first extended to multi-dimensional data setting in [**Kulldorff, 1997**] where it was applied to the two-dimensional spatial domain. Shortly afterwards, it was extended to the spatio-temporal domain in [**Kulldorff *et al.*, 1998**] and applied to an epidemiology dataset.

As the scan statistic was mainly developed for application to epidemiology datasets, it had made the assumption of having 'population-at-risk data' producing the individual point events. In [**Kulldorff *et al.*, 2005**], the methodology was developed further to only necessitate the point events as input without auxiliary data. Variations in the shape of the statistic, the importance of which will be expanded on later in this subsection, have included an elliptical cylinder [**Kulldorff *et al.*, 2006**] and a square pyramid [**Iyengar, 2004**] as examples. In general, the main objective of any spatio-temporal scan statistic is to identify a volume within the space-time domain of highest point event density relative to what would be expected as based on a selected background point process.

For application specifically to natural hazard data, the space-time scan statistic has been used for various hazard types: landslides [**Tonini *et al.*, 2022**; **Tonini and Cama, 2019**], wildfires [**Parente *et al.*, 2016**; **Pereira *et al.*, 2015**; **Orozco *et al.*, 2012**; **Tuia *et al.*, 2008**], and floods [**Wang *et al.*, 2021**]. The exact implementation of the space-time scan statistic with the mathematical justification and derivations is shown in **Appendix D**. The implementation can be concisely summarised with the following steps:

1. **Selecting a background point process**: to establish a benchmark for how many point events would be expected in a given space-time volume, background point process (a subcategory used during the critical literature review phase (**Table 2.1**) was selected to identify the various approaches used in the literatures) has to be selected, denoting a random distribution of point events. For the implementation of the methodology in this thesis, a homogeneous Poisson process is selected.
2. **Defining the volume parameters**: to find the space-time volume with the highest point event density relative to the background point process, locations in the space-time domain and various geometric parameters must be iterated computationally. The set of geometric parameters depend on the selected scanning window shape. For publications using space-time scan statistics, only 6 out of 27 (22%) used elliptic cylinder scanning volume. For this thesis, considering that the elliptic cylinder parameter set includes a circular base for the cylinder, an elliptic cylinder is selected to allow for determination of directionality in the datasets. Thus, the final set of volume iteration parameters comes from permutations of centroid locations, length of the semi-major axis, shape of the elliptic base (ratio between the length of the semi-major and the semi-minor axes), angle between the semi-major axis and first spatial dimension and height. The shape of the scan volume used for the scan statistic was identified as a subcategory of the critical literature review (**Table 2.1**).
3. **Calculating the likelihood ratio**: given the number of point events within the particular iteration of the elliptic cylinder, a likelihood ratio is calculated using the assumed background homogeneous Poisson process. Other possibilities identified during the literature review were the Bernoulli [**McNally *et al.*, 2009**; **McNally and Colver, 2008**] and Hypergeometric [**Tuia *et al.*, 2008**] models. For publications using space-time scan statistics, only the three given here (out of 27, 11%) used models other than the homogeneous Poisson process. The homogeneous Poisson process is selected

for the clustering analysis of point events without any auxiliary geo-referenced variables (as discussed in **Section 1.1**). As the selected volume is assumed to contain a higher density of point events, a point process creating point events is then selected to be an inhomogeneous Poisson process. Therefore, the cluster of point events selected by the space-time scan statistic will be events contained within the elliptic cylinder with the highest likelihood ratio (LR) value given as

$$\mathrm{LR} = \frac{Likelihood\ for\ an\ inhomogeneous\ Poisson\ process}{Likelihood\ for\ a\ homogeneous\ Poisson\ process} \tag{2.2}$$

2.3.2. Trajectories

In contrast to point events, the critical literature review on spatio-temporal clustering of the trajectory data type contained several methodology and field of study review publications that were not focused on a particular field of application [**Atluri *et al.*, 2018**; **Yuan *et al.*, 2017**; **Zheng, 2015**; **Long and Nelson, 2013**; **Kisilevich *et al.*, 2009**; **Palma *et al.*, 2008**]. Review publications allowed for better understanding of the established categorisation of the spatio-temporal clustering methodologies and their development timeline. The identified review publications cover an extensive time period and provide a common categorisation for the various spatio-temporal clustering methodologies available for trajectories:

- **Model-based methodologies**: derive a global mathematical mixture model that would describe the full dataset and identify individual clusters.
- **Distance-based methodologies**: embeds the dataset into a feature space, where each dimension is considered separately, and a distance measure is used within the framework of a generic clustering algorithm.
- **Density-based methodologies**: density thresholds in terms of high number of trajectories within some neighbourhood regions are established around trajectories to separate inter-class trajectories from noise.
- **Visual-aided methodologies**: user input is used to guide and adjust various steps and input parameters of clustering methodologies based on visual information output at various stages of the methodologies.
- **Micro-clustering**: clustering is performed on piece-wise segments of the trajectories, attempting to establish smaller segments of the trajectories that would exhibit a higher level of spatio-temporal clustering.

- **Flocks**: instead of determining which trajectories are similar based on some measure of similarity, attempts to discover groups of trajectories that move together in time.

Each of the clustering methodology categories has their disadvantages for the generality of the applicable spatio-temporal methodologies. Model-based methodologies will have varying components of the model based on the particular application, losing their generality. Distance-based methods lose trajectory track information by transforming the dataset and treating each dimension separately. Visual-aided methodologies require user input for adjusting parameters during some methodological steps and assessing intermediate performance, which depends on the dataset and field of application. Micro-clustering excludes parts of the dataset and flocks do not perform clustering as defined in the context of this thesis (**Section 1.1**).

For implementation during this thesis, the density-based methodology category is selected. Dependent on the spatio-temporal distance function, these methodologies do not transform the dataset and have the same set of input parameters independent of the field of application (although their values will differ based on the applied datasets). By identifying clusters of increased inter-cluster density of trajectories, these methodologies are robust to noise. The development of density-based clustering methodologies in the context of the trajectory data type is presented in **Section 2.3.2.1**.

2.3.2.1. Density-based clustering

In the context of trajectory data clustering, the development of the density-based methodologies was traced with 10 publications and the Trajectory-OPTICS algorithm [**Nanni and Pedreschi, 2006**] was selected. The density-based methodologies are relatively recent and developed from the general DBSCAN algorithm [**Ester *et al.*, 1996**] without any initial modifications for the trajectory data type. The OPTICS algorithm [**Ankerst *et al.*, 1999**] was introduced to extend the DBSCAN algorithm to account for clusters of varying densities and provides a more intuitive way of identifying clusters. While initially not accounting for the trajectory data type and applying to point events, it was extended to the Trajectory-OPTICS [**Nanni and Pedreschi, 2006**] algorithm and its main objective is to identify individual trajectories with relatively high number of neighbouring trajectories and extract sequences of trajectories with low distance measure values relative to them.

The exact implementation of the Trajectory-OPTICS algorithm is given in **Appendix E**, but the main steps can be summarised as below:

1. **Identifying 'core' trajectories**: to determine which trajectories have a relatively high number of neighbouring trajectories within a given spatio-temporal distance, two input parameters are required: (i) minimum number of neighbouring trajectories and (ii) neighbourhood distance within which such a number of trajectories would be present. Such trajectories are identified as being 'core' trajectories of the dataset. In the context of trajectories, the distance between a pair of trajectories can be calculated using various similarity measures, but the spatio-temporal distance as in [**Nanni and Pedreschi, 2006**] is implemented in this thesis. The similarity measure (i.e. various spatio-temporal distances) used for the various density-based trajectory clustering methodologies was identified as a subcategory of the critical literature review. Additional similarity measures and various approaches are discussed as a further research direction of this thesis (**Section 6.3.3**).
2. **Calculating distances between trajectories**: from each of the identified 'core' trajectories, the aforementioned spatio-temporal distance as a similarity measure is calculated to all other trajectories of the dataset.
3. **Ordering trajectories**: all trajectories are visited consecutively, with the order being determined by the smallest spatio-temporal distance with respect to any already visited 'core' trajectory. This distance value is then stored for that particular trajectory and the trajectory is not visited again.
4. **Identifying sequences of low reachability distance values**: individual clusters can be identified if the spatio-temporal distances calculated in the previous step are plotted against the visiting order of the trajectories. As individual trajectories are not visited twice during the algorithm, once all trajectories within a dense region are visited, a local maximum in the spatio-temporal distance values will follow as the algorithm 'jumps' to the next cluster of trajectories. Therefore, identifying low-value sequences will return all individual clusters.

2.4. Characterisation of clustering methodologies as global and local clustering measures

After defining the two selected data types and the various properties of the spatio-temporal domain (**Section 1.1**), the definition of spatio-temporal clustering is expanded in the context of this thesis and relevant publications, proposing a similar division between the methodologies. Clustering can be understood as a global measure of the dataset (i.e. spatial and temporal scales over which space-time interaction is present) or division of objects (i.e. point events or trajectories) into individual clusters based on a given measure of similarity. The former will be referred to as 'global' clustering measure, while the latter as a 'local' clustering measure with the following definitions:

- **Global clustering measure**: provides a single quantitative measure of clustering present in the dataset given spatial and temporal scales. The global clustering measure assesses the extent of the interaction between the spatial and temporal domains [**Diggle *et al.*, 1995**]. The space-time interaction can be understood as correspondence between the two domains (i.e. if objects are close in space, in the case of global clustering, they are also close in time). As a global clustering measure, the given methodology does not return object clusters, but rather is applicable as an initial assessment for the dataset.
- **Local clustering measure**: identifies individual objects that belong to individual clusters as they are close to each other in the spatio-temporal domain based on some similarity measure. An example of the local clustering measure definition is given in [**Estivill-Castro, 2002**] as 'the objects (i.e. point events or trajectories) are clustered or grouped based on the principle of maximising the inter-class similarity and minimising the intra-class similarity'. To accomplish this, there are two general approaches:
    - **Top-down approach**: a heterogeneous population is used as input to separate it into a number of relatively homogeneous subgroups.
    - **Bottom-up approach**: events within a population are grouped based on a measure of similarity, which depends on the domain and particular application.

  The focus for any local clustering measure in the context of this thesis is on how to measure the similarity in the spatio-temporal domain. While in the spatial domain, the Euclidean distance metric can be used and in the temporal domain the events can be

directly clustered based on time differences, in the combined spatio-temporal domain, the optimal way of combining the distance metrics into a single measure of similarity is not obvious and is highly dependent on the particular application and data type.

For the trajectory data type, there was no clear categorisation of the clustering methodologies by this framework. The various trajectory clustering categories in **Section 2.3.2** all consist of methodologies identifying individual trajectory clusters, fitting into the local cluster measure category. For the point event data type, this particular categorisation of methodologies and their varying purpose is mentioned when applying combinations of these methodologies: in epidemiology [**McNally *et al.*, 2008**; **Diggle *et al.*, 1995**], crime [**Nakaya and Yano, 2010**] and in natural hazard research [**Wang *et al.*, 2021**; **Tonini and Cama, 2019**] and [**Orozco *et al.*, 2012**]. In [**Wang *et al.*, 2021**], it is noted that the "global cluster indicators only consider one dimension, disregarding the interaction between space and time". Such a property is only valid when global clustering measures are applied separately to the spatial and temporal domains.

When applied to only the spatial domain, the individual K-functions can be extended into L-functions [**Tonini *et al.*, 2017**; **Hering *et al.*, 2009**] and [**Besag, 1977**] or O-ring statistic [**Wiegand and Moloney, 2004**], measuring global clustering for a ring of certain spatial width around an arbitrary point event. In this thesis, the motivation and aim of the research indicate that the space-time interaction is the exact property of the data that a global clustering measure would quantify. The K-function analysis framework in the spatio-temporal domain is then a valid methodology for the definition of global clustering measure presented in this section. In the literature database, other identified space-time interaction tests that would be valid for application in the spatio-temporal domain include Jacquez k-Nearest Neighbor test [**Jacquez, 1996**], Mantel test [**Mantel, 1967**] and working in fractal dimensions [**Lovejoy *et al.*, 1986**].

2.5. Combining global and local clustering measures in a spatio-temporal clustering analysis framework

As the global and local clustering measures are used for different purposes and measure different spatio-temporal clustering aspects, they can be combined into a common spatio-temporal clustering analysis framework. If considering only the three-point event clustering methodologies from **Section 2.3.1**, various combinations of these methodologies have been

applied to natural hazard datasets: landslides [**Tonini *et al.*, 2022**; **Tonini and Cama, 2019**], wildfires [**Tonini *et al.*, 2017**], and floods [**Wang *et al.*, 2021**]. By combining the various methodologies into a common application framework, there are two objectives:

(i) to provide insight into the global clustering scales and identify individual clusters and

(ii) to use the results of one methodology to aid the computational complexity and clustering performance of other methodologies from the application framework.

In the reviewed publications, there is a clear accomplishment of the first objective as K-function and space-time clustering as a pair of methodologies are used. Only [**Tonini and Cama, 2019**] accomplished the second objective with a framework consisting of all three of the clustering methodologies reviewed in this section.

In [**Tonini and Cama, 2019**], the K-function analysis framework results are used as input for the bandwidth selection procedure for the kernel density estimation methodology. It is argued that the best values of the bandwidth parameter (for a bivariate spatial kernel) are the distance values displaying the highest cluster behaviour detected over the cluster measure *D*. No other link is established between the three methodologies, with only temporal aggregation of point events suggested as a measure to reduce computational complexity. In this thesis, it is argued that having the spatial and temporal ranges values for maximum clustering measure *D* values will cause the methodologies to be unable to identify smaller structures within the given ranges.

From the kernel density estimation implementation details (**Section 2.3.1.2**), the leave-one-estimator can identify the optimal bandwidth given an input set of potential values. These potential bandwidth values can be selected as sets of spatial $\{s_1,\ldots, s_{MAX}\}$ and temporal range $\{\tau_1,\ldots, \tau_{MAX}\}$ values, where $\{s_{MAX}, \tau_{MAX}\}$ are then the range combination that gives the maximum clustering measure *D* value. The exact distribution of the values within each set then depends on the available computational resources.

Once the kernel density estimation methodology returns space-time volumes with relatively higher probability density values, these volumes can be considered to likely contain any individual clusters that a local clustering methodology will identify. From the space-time scan statistic implementation details (**Section 2.3.1.3**), the importance of defining the space-time

scanning volume (i.e. cylinder) parameters is noted. As the ranges of these parameters are extended, the number of permutations that the methodology has to iterate through increases. The space-time volumes identified by the kernel density estimation can then be used to decrease the number of scanning volume centroid locations. The K-function analysis framework also established the spatial and temporal ranges over which there is clustering present in the data. These ranges can be used as limiting the height (i.e. the temporal extent) or the radius or length of the semi-major axis (i.e. for cylinder or elliptic cylinder as scanning volume) parameters of the scanning volume.

Based on this, the spatio-temporal clustering methodology framework proposed in this thesis can be summarised by application of the three reviewed methodologies (**Section 2.3.1**):

1. **K-function analysis**: as a global clustering measure, identifies spatial and temporal ranges $\{s, \tau\}$ over which space-time interaction is present in the dataset and $\{s_{MAX}, \tau_{MAX}\}$ for which the clustering measure D is maximum.
2. **Kernel density estimation**: with K-function analysis indicating the spatial $\{s_1,\ldots, s_{MAX}\}$ and temporal $\{\tau_1,\ldots, \tau_{MAX}\}$ ranges over which space-time interaction is present and significant, these values can be used to create a set of potential bandwidth values. The kernel density estimation then returns particular space-time volumes with higher probability density values due to higher density of point events contained within. From visual inspection, these space-time volumes indicate the potential number of clusters within the dataset if they are well separated.
3. **Space-time scan statistic**: as a local clustering measure, the space-time scan statistic identifies individual clusters of point events. It uses space-time scanning volumes, defined by a set of parameters depending on the volume. Kernel density estimation limits the set of centroids for the volume to visit, while the spatial and temporal ranges from K-function analysis limits the height and the spatial extent of the volumes.

Each methodology from the framework is implemented numerically without use of existing software and libraries. This allows the ability to fully control the implementation details, evaluate any proposed changes to the methodologies and output various additional results.

2.6. Chapter conclusions

This chapter's critical literature review generated 143 publications for a literature database on point event and trajectory spatio-temporal clustering methodologies. Multiple review publications on trajectories were identified [**Atluri *et al.*, 2018**; **Yuan *et al.*, 2017**; **Zheng, 2015**; **Long and Nelson, 2013**; **Kisilevich *et al.*, 2009**; **Palma *et al.*, 2008**], while for point events there was a lack of a common review framework due to the complexity and extent of the field. This chapter further defined and extended the research on global and local clustering measure categories in [**Tonini *et al.*, 2022**; **Wang *et al.*, 2021**; **Tonini and Cama, 2019**; **Tonini *et al.*, 2017**], focusing on the combined use of these measures in a common spatio-temporal clustering methodology application framework.

3. Natural hazard datasets: examples, lightning characterisation, and synthetic data

Chapter summary

This research chapter presents the real-world and synthetic datasets used in applying spatio-temporal clustering methodologies to point event and trajectory data types. A lightning strike dataset of 12,521 individual strikes over the United Kingdom over a time period of one day (1 July 2015) of increased lightning activity is selected for point events. For trajectories, the Atlantic hurricane database (HURDAT2) is used to select a 20-year window (1998-2017) of 316 hurricanes that belong to the North Atlantic basin. The chapter focuses on further characterising the already selected day of increased lightning activity and additional day (28 June 2012) as two case studies. Physical variables such as movement speed, inter-event time distribution and the distance to the movement track (i.e. spatial spread) are characterised from these case studies to establish their use in a spatio-temporal model framework. Overview the available models for producing synthetic point event models is presented. Based on this, a synthetic model for point event datasets produced by a moving source as a representative model for lightning strikes as part of a thunderstorm-system is detailed and implemented. This model incorporates the three physical parameters returned from case studies as input parameters. Finally, a set of spatio-temporal models are selected and detailed for a simulation study. These models include homogeneous Poisson and Thomas (clustered) point processes as the benchmarks for random and clustered point event datasets. The aforementioned spatio-temporal model representative of lightning strikes are split into single storm and three storm models. Each of these models has variations in its input parameters based on the characterisation results for the three physical variables and additional parameters.

This chapter aims to answer the following questions:

Q3.1. What are the established synthetic models for point-event natural hazards?

Q3.2. How are the physical properties of the natural hazards incorporated in these models?

Q3.3. For lightning strikes as point events produced by a moving source (i.e. thunderstorm-system) – what physical properties of the hazard can be used to create a synthetic model?

Q3.4. What input parameter ranges can be classified as ‘representative’ of lightning?

Q3.5. What synthetic datasets representative of lightning strikes can be established in order to use them for application of the spatio-temporal clustering methodologies?

3.1. Introduction

This chapter overviews the real-world natural hazard datasets selected for use in this thesis and presents the synthetic models used to model the particular natural hazard. Already established models are overviewed for the point event and trajectory data types, and a novel synthetic model is detailed for point events produced by a moving source. Particular emphasis is placed on explaining the significance of the model input parameters in relation to the physical properties of the particular natural hazard. A link between the real-world natural hazard properties and models that can represent them as accurately as possible is established.

As discussed in **Section 1.1**, this thesis focuses on the two larger groups of natural hazard data: point events and trajectories. For both groups, the most prevalent methodologies were presented in **Section 2.5**. To examine the capabilities and limitations of each of the discussed methodologies, this section will present the datasets to which these local and global clustering methodologies are applied. This section can be divided into three distinct subsections:

1. *Examples of real-world spatio-temporal natural hazard data* (**Section 3.2**). Datasets with real-world limitations, including varying levels of coverage, measurement precision and other factors that will be able to demonstrate and compare the performance of the various methodologies.
2. *Characterisation and modelling of lightning strikes as point events in time and space* (**Section 3.3**). Characterisation and modelling results of lightning strikes as point events produced by a moving source (thunderstorm system). Modelled datasets that are representative of lightning as a natural hazard have their input parameters related to the physical parameters of the natural hazard process.
3. *Synthetic models and benchmark datasets* (**Section 3.4**). A simulation study of various spatio-temporal datasets intended as a representation of idealised conditions or conditions representative of the natural hazard (lightning strikes produced by a moving source). These synthetic datasets can be used for establishing benchmarks for how the methodologies are supposed to perform. For point event data type, these conditions are inferred based on the characterisation results of lightning strikes from **Section 3.3**.

Section 3.2 presents both the trajectory and point event real-world datasets. As both categories have varying purposes for their inclusion, justification is given for the particular dataset

choices. For point events, **Section 3.3** analyses two case studies of increased lightning activity over the United Kingdom. It briefly overviews the existing natural hazard spatial-temporal models (for point events). It also introduces a more detailed model for point event natural hazard produced by a moving source that would incorporate the physical parameters of the natural hazards examined in this thesis (lightning strikes as part of a thunderstorm-system).

The characterisation results and modelling approach from **Section 3.3** are used to establish the simulation study approach detailed in **Section 3.4**. The details for the relevant input parameters and generation of the respective datasets are given in **Section 3.3,** with the particular parameter choices and final datasets detailed in **Section 3.4**. For the point event data type, the input parameters for the model are related to the physical variables of lightning as the selected natural hazard. For trajectories, the selected synthetic benchmarks are based on mathematical models without any input from characterising case studies of representative natural hazards.

3.2. Examples of real-world spatio-temporal natural hazard data

If synthetic datasets are used to establish dataset benchmarks on how the various spatio-temporal clustering methodologies perform on homogeneous or clustered datasets, real-world datasets allow assessment of the methodologies' performance in more complex situations. The application of the methodologies is no longer on idealised cases with given input parameters. This allows evaluation of the robustness of the methodologies and identifies potential research directions for improvements on the accuracy and/or precision of the individual methodologies.

3.2.1. Real-world point event data

The selected real-world point event dataset is a subset of the ATDnet [**Gaffard *et al.*, 2008**] lightning strike location dataset. The ATDnet network uses surface-based very low-frequency radio receivers located across Europe that have been active throughout the United Kingdom since 1987. Lightning density analysis studies have been done using data collected by this network, but the number of point events collected during the summer months is on the scale of millions per month [**Anderson and Klugmann, 2014**]. Such large datasets are not feasible for applying clustering methodologies with given computational resources and a subset must be selected based on a particular case study of increased lightning activity.

In [**Lewis and Silkstone, 2017**], a three-storm system, which crosses the United Kingdom on 1 July 2015 is analysed. This event was notable due to the different convective modes producing each of the storms in close succession with various propagation speeds. A clustering behaviour in the lightning strike locations over the spatio-temporal domain would then be expected in this situation, with each storm potentially producing a distinct cluster of point events. As the study area for the storm system is clearly defined in the study, further restriction on the dataset is applied spatially, allowing a decrease of the computational costs of implementing the clustering methodologies. Only events recorded within the ATDnet domain (49°N to 61°N, 14°W to 6°E) are selected. As a result, N_L = 12,521 point events are recorded on a single day (1 July 2015) over the ATDnet spatial domain as shown in **Figure 3.1**.

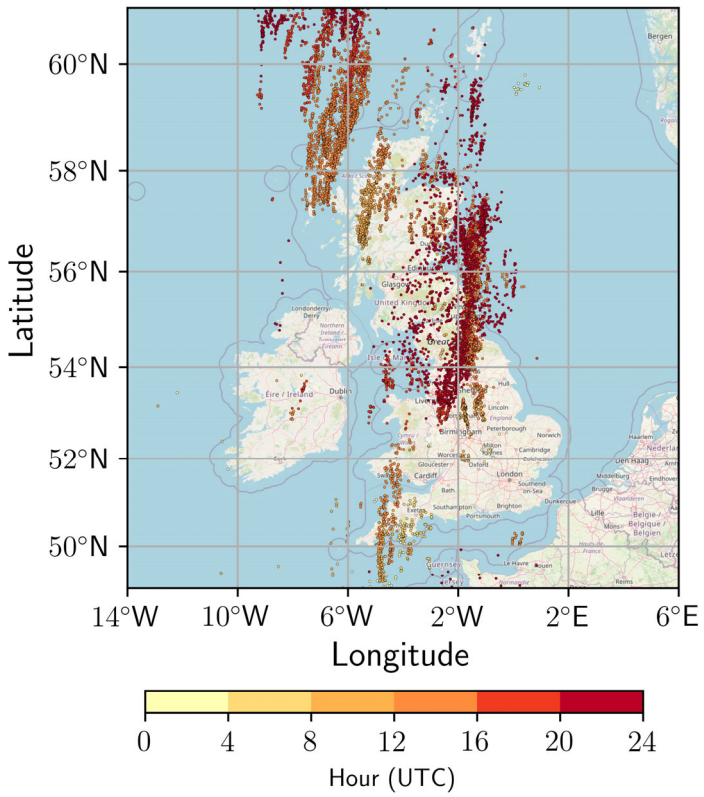

Figure 3.1. A single day of lightning strike data collected on 1 July 2015 over the ATDnet [**Gaffard *et al.*, 2008**] lightning strike location dataset selected for spatio-temporal clustering analyses. Shown is the 2D spatial view of the dataset (total of N_L = 12,521 point events) with the timestamp of each point event given by the colour, according to the colour scale.

3.2.2. Real-world trajectory data

The selected real-world trajectory dataset is a subset of the Atlantic hurricane database (HURDAT2) as collected by the National Hurricane Center (NHC) [**Landsea and Franklin, 2013**]. The NHC conducts a post-storm analysis of each hurricane in order to re-create its track. In addition, once new methodologies are introduced, the database is updated to ensure homogeneity of the tracking data quality throughout history. All hurricane tracks are recorded in six-hour intervals, noting their latitude, longitude, time, central pressure, and maximum wind speed. Data has been gathered since 1851; a total of N = 1,832 individual North Atlantic basin hurricanes have been identified. Despite the re-analysis, authors of the database note that data gathering has been drastically improved since satellite data were used since 1972.

To ensure that the selected subset is of similar data quality and to lower the computational requirements of implementing the clustering algorithms, a 20-year period from 1998 to 2017 over the North Atlantic basin was selected separately. No spatial restrictions are placed on the dataset and trajectories are selected in their entirety both spatially and temporally. Over the 20-year period, a total of N_T = 316 trajectory tracks have been identified as shown in **Figure 3.2**.

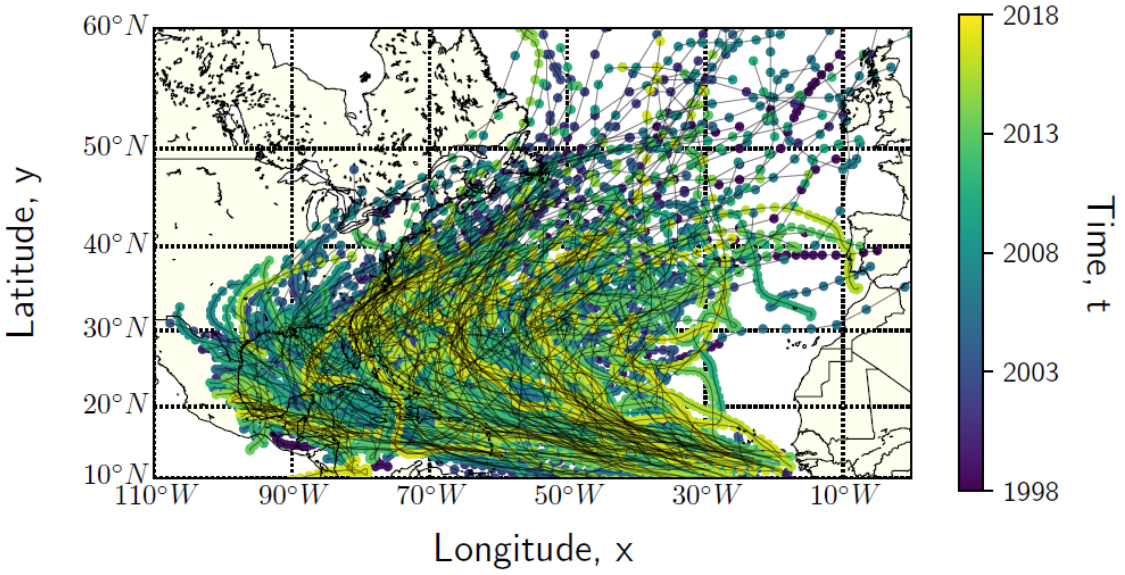

Figure 3.2. The subset (20 year data from 1998-2017 over the North Atlantic basin) of the Atlantic hurricane database HURDAT2 [**Landsea and Franklin, 2013**]. Shown is the 2D spatial view of the N_T = 316 hurricane tracks with the timestamp of each event along the tracks, according to the colour scale.

3.3. Characterisation and modelling of lightning strikes as point events in time and space

This chapter focuses on the characterisation and modelling of lightning strike spatial location and times of occurrence data. A lightning strike dataset recorded by the ATDnet system [**Nash *et al.*, 2006**] over an area between 13°W and 5°E and 49°N and 61°N and the time period 2008 to 2016 is used to separate two case studies of individual days of significant lightning activity. The data used in these case studies can be categorised for characterisation and modelling methodologies as spatio-temporal point events produced by a moving point source [**Kisilevich *et al.*, 2009**] (i.e. a thunderstorm producing a cluster of lightning strike point events).

For spatio-temporal natural hazard data, both spatial and temporal domains, if considered separately, can give insight into different variables of the hazardous process producing the events. The extent and density of lightning flashes or strikes can be gridded over global [**Mezuman *et al.*, 2014**] or continental (e.g. [**Anderson and Klugmann, 2014**; **Enno *et al.*, 2020**]) spatial scales and varying time periods. The combined spatio-temporal domain allows the observation of dependencies among measurements in the spatial and temporal dimensions (i.e. autocorrelation and heterogeneity) [**Atluri *et al.*, 2018**]. Therefore, natural hazard point events are structurally related to each other in space and time and show varying properties in different spatial regions, time periods and their combined spatio-temporal windows.

As natural hazards are inherently spatio-temporal processes, characterising the natural hazard over the combined spatio-temporal domain allows us to gain more sophisticated insight into the extent and evolution of the hazardous process. For wildfires, spatio-temporal interaction can be modelled directly as one of the parameters for a given point process intensity function (e.g. [**Serra *et al.*, 2014**; **Møller and Díaz-Avalos, 2010**]). For earthquakes, the spatio-temporal domain can be used with earthquake magnitude to inform the separation of foreshock, mainshock and aftershock clusters [**Zaliapin *et al.*, 2008**; **Zaliapin and Ben-Zion, 2013**]. The spatio-temporal domain can not only directly inform the modelling of a given natural hazard process, but also provide additional insight into the natural hazard and its properties.

For lightning strikes, a spatio-temporal point event model was developed in [**Finke, 1999**], while physics-based models are more common (e.g. [**Rakov and Uman, 2003**; **Yair *et al.*, 2009**]). The main objective of this study is to extend the work in [**Finke, 1999**] by

characterising various physical lightning variables based on real-world case study data and incorporating them into a general spatio-temporal framework to statistically model spatio-temporal point events produced by a moving point source. This framework can be used to generate multiple synthetic point event datasets that would be representative of lightning strikes data. The section is organised as follows:

- **Section 3.3.1** expands on the existing literature on characterising and modelling lightning, identifying individual approaches to represent and analyse spatio-temporal point event data.
- **Section 3.3.2** details two case studies from the ATDnet dataset that will be considered in this work. Each case study refers to a separate day of increased lightning activity with available synoptic analysis. These analyses identify individual thunderstorms that were producing lightning strikes and from location, time of occurrence, Doppler radar and additional data, lightning strikes can be assigned to particular identified thunderstorms.
- **Section 3.3.3** details the methodologies used to characterise physical variables for all the identified thunderstorms and model these physical variables as input parameters for a spatio-temporal point event model.
- **Section 3.3.4** then discusses the results of these methodologies and the resulting characterisation of lightning strikes events.
- **Section 3.3.5** details the approach taken for modelling the lightning strikes as point events produced by a moving point source by including the characterised physical variables as parameters of a statistical model. It further gives examples of using the modelling approach to create point event datasets representative of lightning strikes. It also demonstrates the ability of flexible input parameters to create various point events datasets for other applications (e.g., validation of spatio-temporal clustering methodologies).
- **Section 3.3.6** critically assesses the approach taken in this study, its limitations, potential improvements, and future directions.

3.3.1. Review of existing point event models

3.3.1.1. Lightning

Lightning from a physics perspective can be defined as an electric spark (typically 5–10 km in length), caused when the resistance of the intervening air between areas of positive and negative charge is overcome [**Dwyer and Uman, 2014**]. Thunderstorms that produce lightning have a basic dipolar charge structure consisting of a negative charge at mid-levels of the storm and a positive charge in the upper level [**Shao and Krehbiel, 1996**]. This creates intracloud (IC) lightning strikes within the storm, intercloud lightning between different storms, and cloud-to-air and cloud-to-ground (CG) lightning.

Characterisation of lightning as a physical phenomenon has been done on data recorded directly from a current transducer [**Chowdhuri *et al.*, 2005**], from digital high-speed cameras [**Ballarotti *et al.*, 2012**], from electromagnetic radiation in the very high frequency (VHF) and ultra-high frequency (UHF) bands [**Shao and Krehbiel, 1996**; **Richard *et al.*, 1986**] or as point events by on-ground detection networks [**Galanaki *et al.*, 2015**; **Anderson and Klugmann, 2014**; **Steiger *et al.*, 2002**] or satellite detectors [**Beirle *et al.*, 2014**]. This study does not focus on characterising the physics of lightning initiation but rather on characterising physical variables of lightning events as point events. The physical variables describing the point event set are characterised based on the available data and are used to statistically model lightning strikes as spatio-temporal point events produced by a moving point source.

The recording methodology establishes the lightning variables that are characterised. For example, by measuring current directly, variables such as peak current, waveshape and return stroke velocity can be characterised [**Chowdhuri *et al.*, 2005**]. This study focuses on characterising lightning strikes as spatio-temporal point events produced by a moving point source and as recorded by an on-ground detection network over a short temporal scale. Previous studies have used lightning data of various spatial (i.e. global, continental, regional or national), but consistently long temporal scales (from 5 [**Anderson and Klugmann, 2014**] to 11 [**Steiger *et al.*, 2002**] years). Characterised variables include flash density (strikes per km^2 per year), strike distribution by month, strikes per hour over land or sea and the percent of positive flashes (availability dependent on the capabilities of the detection network).

3.3.1.2. Modelling spatial-temporal characteristics of natural hazard point events

Multiple approaches are possible to model spatio-temporal point events, depending on the desired purpose. In this study, two properties are required of the model: (1) the inclusion of characterised physical variables and (2) the ability to produce synthetic data representing the hazard process. A brief literature review was conducted to overview the various approaches to model spatio-temporal point event data of any natural hazard. For each of the 20 publications, this review (**Appendix Table G1**) identified:

- **Data attributes**: the size and spatial and temporal scales of the data.
- **Variables that inform the model**: physical variables incorporated into the model framework.
- **Main mathematical model**: the statistical model that has been implemented.

As the main mathematical model, various statistical point process models [**González *et al.*, 2016**] were used to model 6 different natural hazards:

- avalanches [**Eckert *et al.*, 2010**],
- volcano eruptions [**Jaquet *et al.*, 2012**],
- tornadoes [**González *et al.*, 2020**],
- earthquakes [**Van Lieshout and Stein, 2012**],
- hurricanes [**Economou *et al.*, 2014**],
- wildfires [**Serra *et al.*, 2014**; **Xu and Schoenberg, 2011**; **Møller and Díaz-Avalos, 2010**].

Different point process models were extensively used for wildfire modelling in particular. Existing studies used the Cox point process to incorporate different variables and covariates (e.g. cause of fire, spatial covariates and meteorological variables).

A purely statistical model approach was not found to be common for lightning strikes. There is extensive research for modelling the possible modes of charge transfer to ground during lightning discharges (i.e. dart-leader–return-stroke sequence, continuing current and M-

component [**Rakov and Uman, 2003**]). Lightning physics has also been modelled as a coupled network of thunderstorm cells [**Yair *et al.*, 2009**]. These thunderstorm clouds consist of electrically active cells that can be interacting or non-interacting. The electric field in a specific region of the thundercloud is modelled and, when passing its maximum, generates a lightning flash.

While able to describe and model lightning physics, these above models do not apply to model lightning as point events. For the spatial and temporal domains, point processes have been used for counting the number of total lightning strikes [**Artigas, 2012**] and modelling lightning only in the temporal domain [**Telesca *et al.*, 2008**]. In the combined spatio-temporal domain, a Gaussian cloud approach to modelling propagation of thunderstorms and generating lightning point events has been proposed in [**Finke, 1999**]. The distribution function of lightning point locations is approximated by a cloud of constant circular shape, moving with constant velocity. It is assumed that the lightning point event occurrence times are distributed uniformly.

In this thesis, a similar approach is proposed to the one in [**Finke, 1999**], but with a data-driven strategy for choosing the parameters in the model. For example, a uniform distribution of lightning strike occurrence times is not assumed but the aim is to empirically estimate that from the data. The occurrence times are incorporated by empirically sampling inter-event time distributions (**Section 3.3.3.2** and **3.3.5.1**). The movement speed (**Section 3.3.3.3**) and spatial spread (**Section 3.3.3.4**) are lightning variables that are also incorporated into the model, but by different methodologies (**Sections 3.3.4.2, 3.3.4.3**) than [**Finke, 1999**]. All three model variables are informed by or sampled from real-world lightning strike data from two case studies of increased lightning activity (**Section 3.3.2**). The model implementation details and synthetic data examples are given in **Section 3.3.5**.

3.3.2. Data collection of lightning and storms

This section first presents the ADT lightning data collection system and its parameters (**Section 3.3.2.1**). To identify and assign lightning strikes to individual thunderstorms, two case studies of increased lightning activity over two individual days are detailed (**Sections 3.3.2.2** and **3.3.2.3**). Each case study includes additional synoptic analysis information from available literature [**Lewis and Silkstone, 2017**; **Clark and Webb, 2013**] and presents the lightning strikes assigned to three individual thunderstorms in each case study.

3.3.2.1. ATDnet lightning detection system and available data

The lightning data used in this study is recorded by a very low frequency (VLF) long-range lightning location system (LLS). Over the United Kingdom, the Met Office operates the ATDnet LLS [**Nash *et al.*, 2006**], which locates lightning discharges using the arrival time difference (ATD) method [**Lee, 1986**]. ATDnet predominantly detects sferics (electromagnetic waves in the VLF range propagating in the earth–ionosphere waveguide [**Rakov and Uman 2003**] created by CG lightning strikes. As the energy and polarisation of sferics created by CG return strikes mean that they can travel more efficiently in the Earth-ionosphere waveguide, they are more likely to be detected at longer ranges [**Enno *et al.*, 2016**].

The system takes advantage of the long propagation paths of sferics to cover large areas with only a limited number of sensors. The effective range of ATDnet encompasses Europe, northern Africa, and northern parts of the Atlantic, but in our study, the range is focused on an area between 13°W and 5°E and 49°N and 61°N and over the time period 2008 to 2016 as provided by the Met Office [**2017**].

From this large dataset, two days are selected, one for each case study. The first day of increased lightning activity is 28 June 2012, and the corresponding synoptical analysis is provided by [**Clark and Webb, 2013**]. The second day occurred on 1 July 2015 and was analysed by [**Lewis and Silkstone, 2017**]. Both synoptic analyses give initiation and propagation details of three individual thunderstorms with auxiliary data from Doppler radars in the region and observations of other natural hazards. For both days, the spatial extent and propagation (in 4 hour intervals) of the available lightning strike data subsets are shown in **Figure 3.3**. This figure will be discussed in more detail in **Section 3.3.2.2** and **Section 3.3.2.3**.

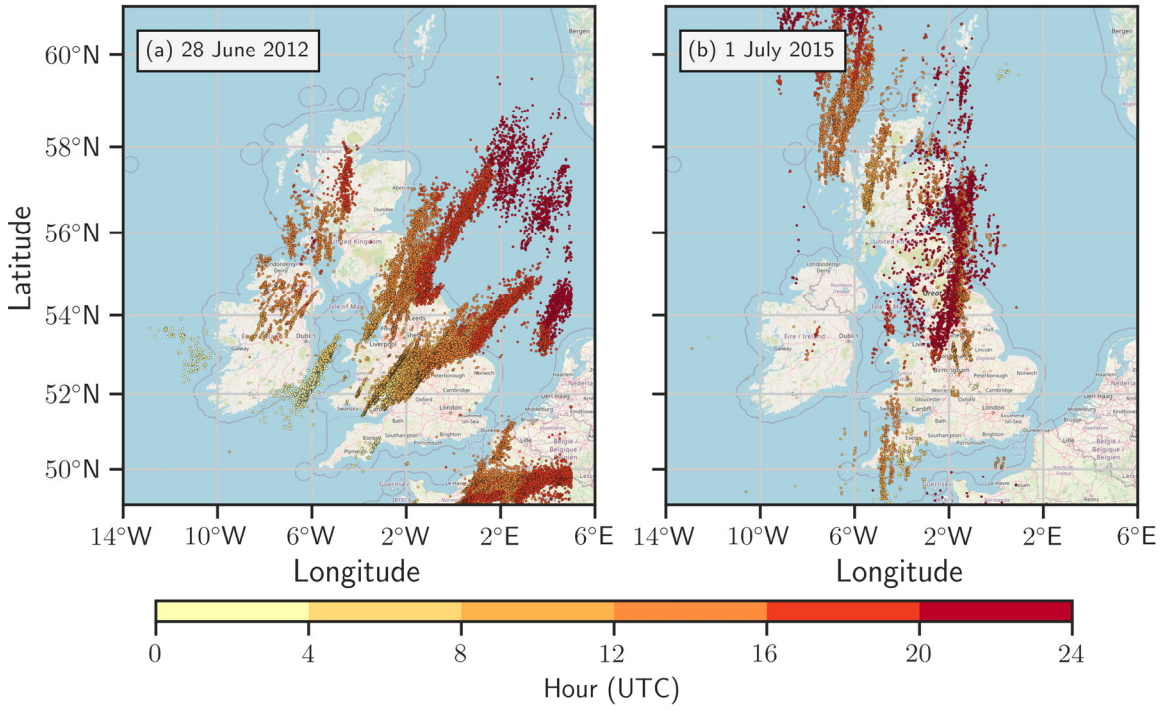

Figure 3.3. The spatial extent of lightning strike occurrences as recorded by the ATDnet system [**Gaffard *et al.*, 2008**] from 00:00 to 23:59 UTC on **(a)** 28 June 2012 and **(b)** 1 July 2015. Timestamps of all lightning strikes are categorised into 4-hour intervals with times indicated by the colour scale. Lightning strike data from Met Office from 2008–2016 (2017) and covering an area between 13.0°W and 5.0°E and 49.0°N and 61.0°N. Background map information from OSM (2022, ©OpenStreetMap Contributors).

3.3.2.2. Case study 1: Three supercell thunderstorms across English Midlands on 28 June 2012

The first case study in **Figure 3.3a** is based on a synoptic analysis of three supercell thunderstorms that crossed the English Midlands on 28 June 2012 during the 24-hour time period 00:00 to 23:59 UTC [**Clark and Webb, 2013**]. The three thunderstorms are identified by their approximate time and location of initiation, propagation path in descriptive terms and snapshots of the location around Doppler radar stations. The three thunderstorms are denoted as Storm 1A, 1B and 1C, corresponding respectively to supercells 'A', 'B' and 'C' from the synoptic analysis. In the Storm 1B case study, the location and temporal information of the observed hail and tornado are used to assign lightning strikes to individual thunderstorms.

To use the available identifying information from the synoptic analysis, the lightning strike location data, a total of N_L = 77,252 strikes, is projected onto the OSGB 1936 (EPSG:27700) northing/easting coordinates. In this three-dimensional spatio-temporal space, planes are drawn

to separate the lightning strike dataset. Points (i.e. position and time of observation) on these planes are selected based on the synoptic analysis and auxiliary data from Doppler radars in Clee Hill and Ingham and observations of a tornado in Sleaford.

Using Storm 1B as an example for this assignment procedure, the initial limitation on the available lightning strike point events is placed both spatially and temporally. From the synoptic analysis [**Clark and Webb, 2013**], Storm 1B formed shortly after 10:00 UTC, started merging with smaller cells around 13:00 UTC and was tracked until 14:37 UTC. Spatially, the study area is focused on 350–600 km in easting and 200–450 km in northing coordinates. This spatial area includes the Clee Hill station, from which Storm 1B was always observed eastwards and places Sleaford close to the centroid of the study area. The lightning strike data subset available for assigning lightning strikes to Storm 1B is shown in **Figure 3.4**.

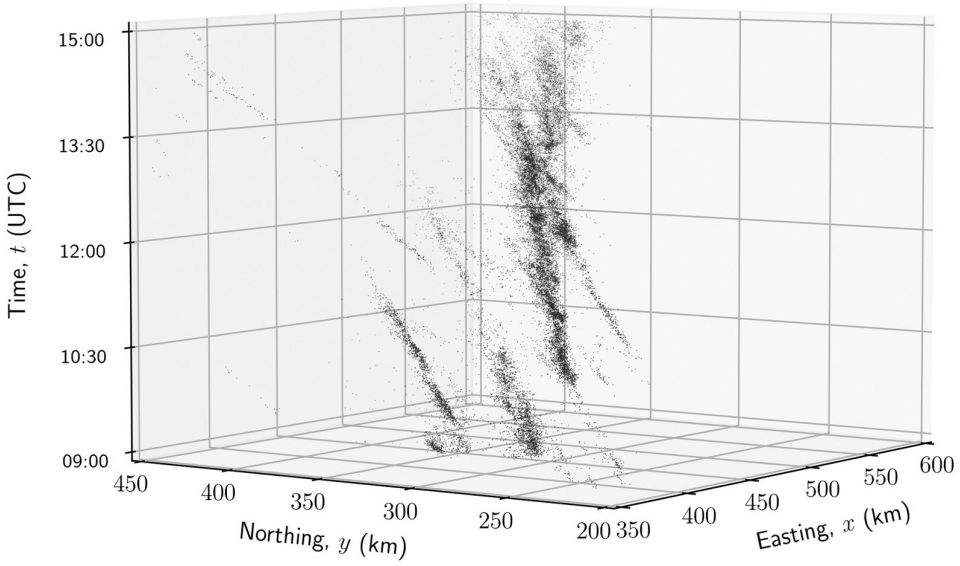

Figure 3.4. The lightning strike data subset available for assignment to Storm 1B of the first case study. The case study is based on a synoptic analysis of three supercell thunderstorms that crossed the English Midlands on 28 June 2012. Spatial extent was restricted to x=[350,600] km and y=[200,450] in the OSGB 1936 (EPSG:27700) easting/northing coordinates respectively. Temporal extent was considered from 09:00–15:00 UTC. All available lightning strikes are indicated as black scatter points.

Spatial and temporal scales can be further restricted from the auxiliary Doppler radar data available from the Clee Hill and Ingham stations. Snapshots of the Storm 1B location relative to these stations are available at 10:00, 12:00 and 14:00 UTC. Unprocessed reflectivity data are also available in 10-minute intervals from 11:02–11:52 UTC in Clee Hill and in 10- or 15-minute intervals from 12:42–13:37 UTC in Ingham. The snapshot frames move with the storm, so are not useful in determining the exact location of Storm 1B relative to the stations, but aid

in understanding the evolution of the storm. The relative positions of the Clee Hill and Ingham radar stations to the lightning strike subset from **Figure 3.4** are shown in **Figure 3.5** at 12:00 and 14:00 UTC.

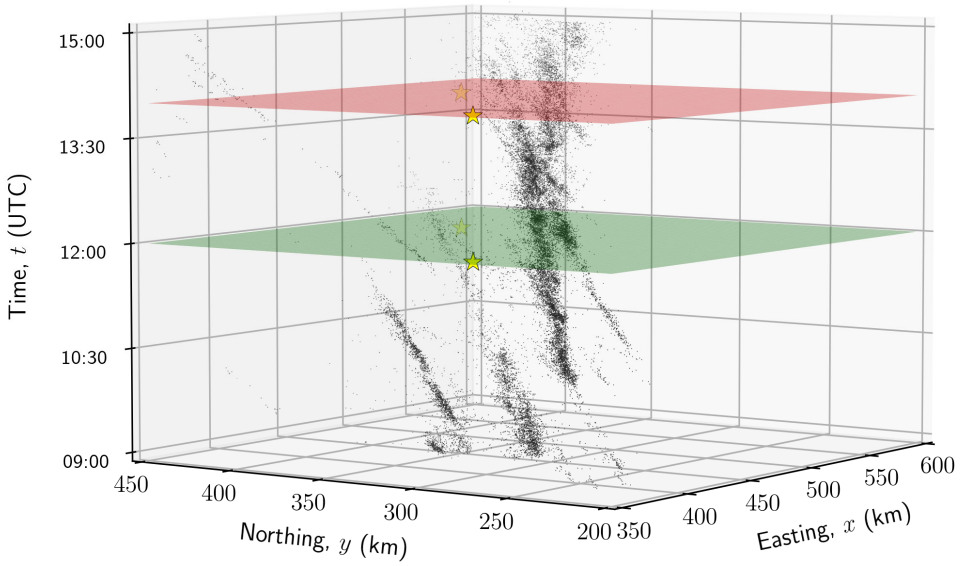

Figure 3.5. The availability of Doppler radar station location information for assignment of lightning strikes to Storm 1B of the first case study. All available lightning strikes are indicated as black scatter points. The locations of the Clee Hill and Ingham radar stations are marked as yellow start scatter points at two time points – 12:00 (green plane) and 14:00 (red plane) UTC.

Observation and tracking information available on the Sleaford tornado and hail further confirm the development and propagation of Storm 1B. The storm is tracked from 11:22–13:52 UTC through hail observation and radar data. The Sleaford tornado developed between 13:12 –13:37 UTC. Both time intervals include the merge with smaller storm cells to the south of Storm 1B, complicating the exact assignment of lightning strikes. The lightning strikes assigned to Storm 1B based on this auxiliary data and their relative position to the Sleaford tornado observation are shown in **Figure 3.6**.

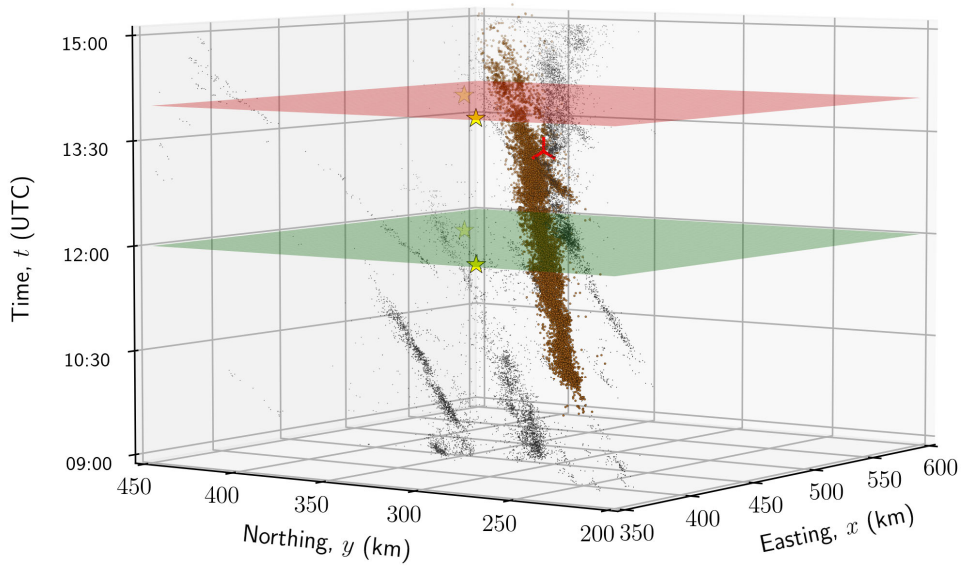

Figure 3.6. The assignment of lightning strikes to Storm 1B (dark brown scatter points) of the first case study. The locations of the Clee Hill and Ingham radar stations are marked as yellow start scatter points at two time points – 12:00 (green plane) and 14:00 (red plane) UTC. The location and first approximated time of occurrence (13:12 UTC) for the Sleaford tornado is marked as a red triangular line marker. Remaining unassigned lightning strikes are indicated as black scatter points.

Figure 3.6 shows that for the majority of the lightning strikes, assignment to Storm 1B was not complicated due to the clear separation of supercell thunderstorms as noted in the synoptic analysis [**Clark and Webb, 2013**]. The exact formation time was not noted in the synoptic analysis, but was chosen to be any time after 10:00 UTC due to Storm 1B not being identified at the 10:00 UTC Doppler radar snapshot. The noted merging of Storm 1B and local cells to the south at approximately 13:00 UTC is the main source of assignment uncertainty for individual lightning strikes. The exact termination time was also not noted in the synoptic analysis and was extended until 15:00 UTC to include any lightning strikes that still followed the general propagation direction of Storm 1B.

The summary of the initial spatial and temporal limitations on the lightning strike dataset, the time instances used from Doppler radar data and any additional use of auxiliary data is shown in **Table 3.1**.

Table 3.1. Variables and their values used for the assignment procedure of individual lightning strikes to the six thunderstorms on 28 June 2012 (Storms 1A, 1B, 1C) and 1 July 2015 (Storms 2A, 2B, 2C). UTC = Coordinated Universal Time.

Date	Storm	Spatial extent x_{min}, x_{max} y_{min}, y_{max} (km)	Temporal extent t_{min}, t_{max} UTC	Time instances of available Doppler radar data UTC	Doppler radar stations	Auxiliary data
28 June 2012	1A	200, 600 150, 500	06:00, 12:00	10:00, 12:00	Clee Hill, Ingham	-
	1B	350, 600 200, 450	09:00, 15:00	12:00, 14:00		Hail, Tornado
	1C	400, 600 150, 400	12:00, 15:00	12:00, 14:00		-
1 July 2015	2A	400, 475 420, 1000	12:00, 17:00	14:08	Hameldon Hill, High Moorsley, Ingham	-
	2B	350, 435 400, 700	14:00, 24:00	19:02, 19:52		-
	2C	200, 1000 200, 1000	19:00, 24:00	21:51		-

A three-dimensional plot of the assigned lightning strikes and all unassigned strikes recorded between 06:00 and 15:00 UTC is displayed in **Figure 3.7**. Additional perspectives of the same plot in the three-dimensional space can be found in **Appendix Figure H1**.

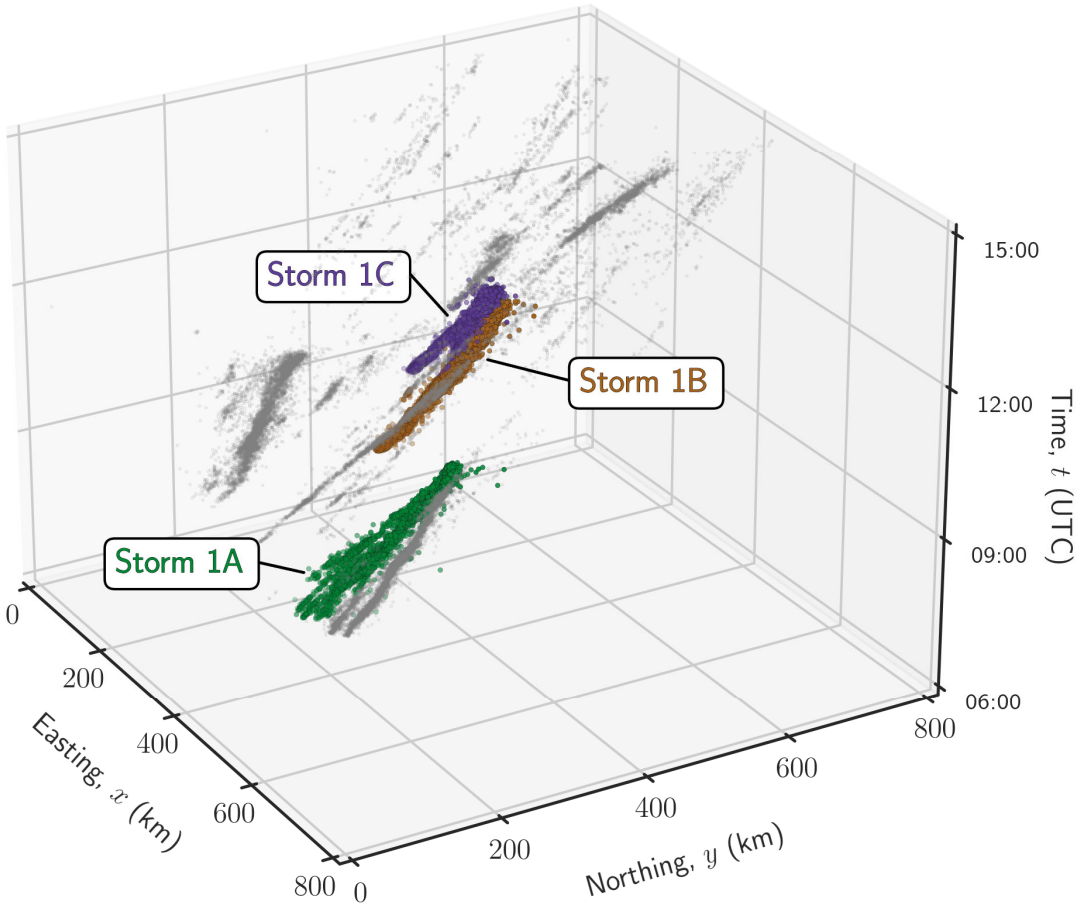

Figure 3.7. Spatial extent and propagation of the lightning strike occurrences as recorded on 28 June 2012 by the ATDnet system [**Gaffard *et al.*, 2008**] and projected onto the OSGB 1936 / British National Grid northing/easting coordinates. Lightning strikes are assigned to the three identified thunderstorms (Storms 1A, 1B, 1C) corresponding to the relevant synoptical analysis [**Clark and Webb, 2013**]. All the remaining unassigned strikes are coloured as per the legend.

Figure 3.7 shows the separation of the individual thunderstorms from the unassigned lightning strikes. Storm 1B had the most detailed synoptic analysis with additional information and figures of the observed Sleaford tornado and associated hail. It is noted that there is a discrepancy between the level of detail of synoptic analysis and additional information between the storms. Storms 1B and 1C are close to each other in time, and there are subsets of unassigned lightning strikes close to the identified thunderstorms in the spatio-temporal domain. This introduces a level of ambiguity in assigning the lightning strikes to individual thunderstorms.

To indicate the relative position of the assigned lightning strikes to the region's geography and their propagation, six one-hour intervals of the assigned lightning strikes are shown in **Figure 3.8**. The lightning strikes are separated by their assignment to a particular individual thunderstorm and are shown relative to the locations of the two Doppler radar stations and the observed tornado in Sleaford. These six intervals were chosen to cover the full temporal range of the lightning strikes as produced by the three thunderstorms.

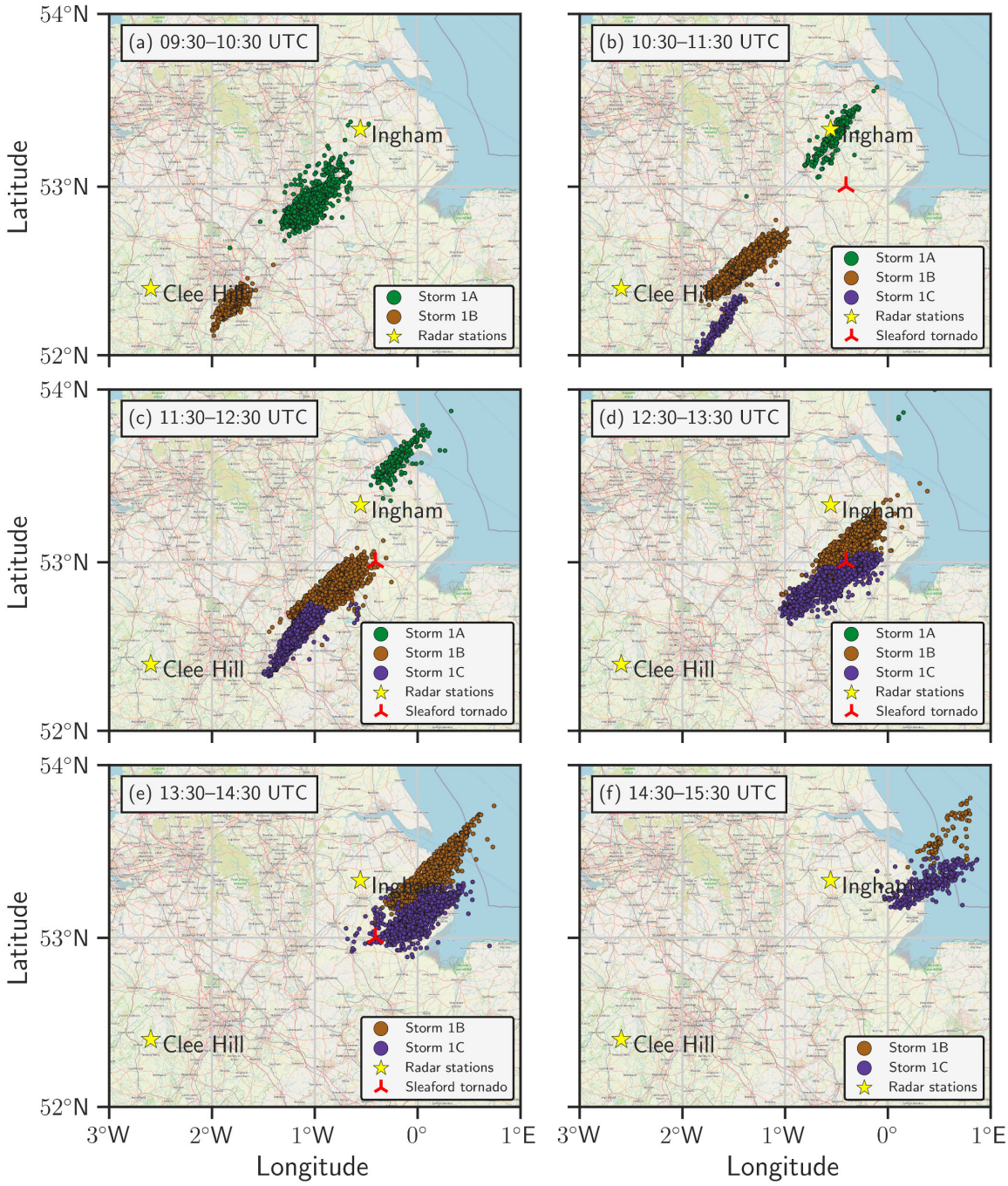

Figure 3.8. Spatial extent for 28 June 2012 of the lightning strike occurrences as recorded by the ATDnet system [**Gaffard *et al.*, 2008**] over times intervals corresponding to the relevant synoptical analysis (Clark & Webb, 2013): (**a**) 09:30–10:30, (**b**) 10:30–11:30, (**c**) 11:30–12:30, (**d**) 12:30–13:30, **(e)** 13:30–14:30 and (**f**) 14:30–15:30 UTC. Only lightning strikes assigned to any of the three selected thunderstorms (Storms 1A, 1B, 1C) are shown. Locations of relevant Doppler radar stations and the observed location of the Sleaford tornado are added as per the legend for direct comparison with synoptic analysis.

As seen from **Figure 3.8**, all three thunderstorms (Storms 1A, 1B, 1C), when viewed in the projected two-dimensional spatial domain, cross the Midlands in a SW–NE direction. The locations of the Clee Hill and Ingham radar stations and the Sleaford tornado allow for direct comparison with relevant figures in the synoptic analysis.

To characterise the lightning strikes in the spatio-temporal domain for an individual thunderstorm, a two-dimensional spatial count of lightning strikes per 1 km^2 can be investigated. This is displayed in **Figure 3.9** for Storm 1A. Violin and box plots of the marginal distributions of northing and easting positions of the lightning strikes are added to the y-axis and x-axis, respectively. These plots provide information on the shape of the empirical distribution, quartiles, and the 5^{th} and 95^{th} percentiles. For other storms, these figures can be seen in **Appendix I**.

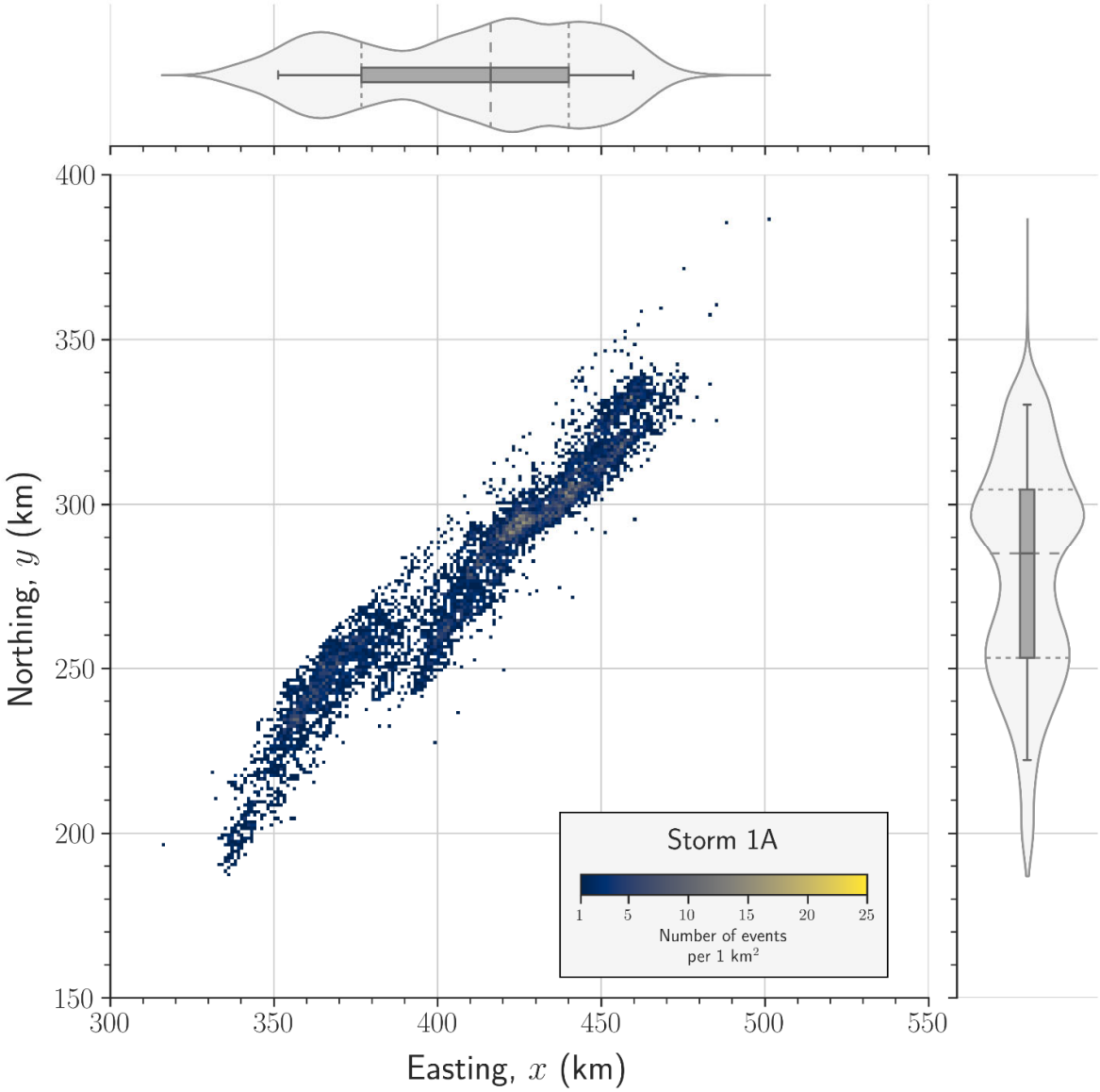

Figure 3.9. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for Storm 1A (28 June 2012). Only bins with more than 1 lightning strike are coloured. Note that the northing and easting extents are the same, but the axes limits differ. The marginal distribution is indicated by overlapping boxplots and violin plots for both the easting and northing values. Boxplots indicate the quartiles and have whiskers for the 5th and 95th percentiles, while the violin plots include all values.

In **Figure 3.9**, the lightning strikes assigned to Storm 1A cover the same spatial scale of close to 200 km in both northing and easting directions. Strike spatial count does not exceed 20–25 strikes/km^2, and areas of such lightning intensity are located further along the propagation track of the storm. The marginal distributions show a decrease in lightning intensity as there is a separation between two areas of increased count values at x = 390 km and y = 275 km.

3.3.2.3. Case study 2: Three structurally distinct severe thunderstorms across northern England on 1 July 2015

The second case study in **Figure 3.3b** is based on a synoptic analysis of three severe thunderstorms that crossed the United Kingdom on 1 July 2015 [**Lewis and Silkstone, 2017**], which are presented as the real-world point event dataset example in **Section 3.2.1**. The three thunderstorms are again identified by their approximate time and location of initiation, propagation path in descriptive terms and snapshots of the location around Doppler radar stations. These three thunderstorms are denoted as Storm 2A, 2B and 2C, corresponding to the three storms identified in their synoptic analysis. In this case study, there is no additional information from occurrences of other natural hazards.

For this second case study, N_L = 12,521 strikes are recorded and projected onto the OSGB 1936 (EPSG:27700) northing/easting coordinates. Identical to the first case study, planes are drawn to separate the lightning strike dataset in the three-dimensional spatio-temporal space. Points on these planes are selected based on the synoptic analysis and the information from only the Doppler radar stations as no other natural hazard were noted in this case study. A three-dimensional plot of the assigned lightning strikes and all unassigned strikes recorded between 12:00 and 24:00 UTC is displayed in **Figure 3.10.** Additional perspectives of the same plot in the three-dimensional space can be found in **Appendix Figure H2**.

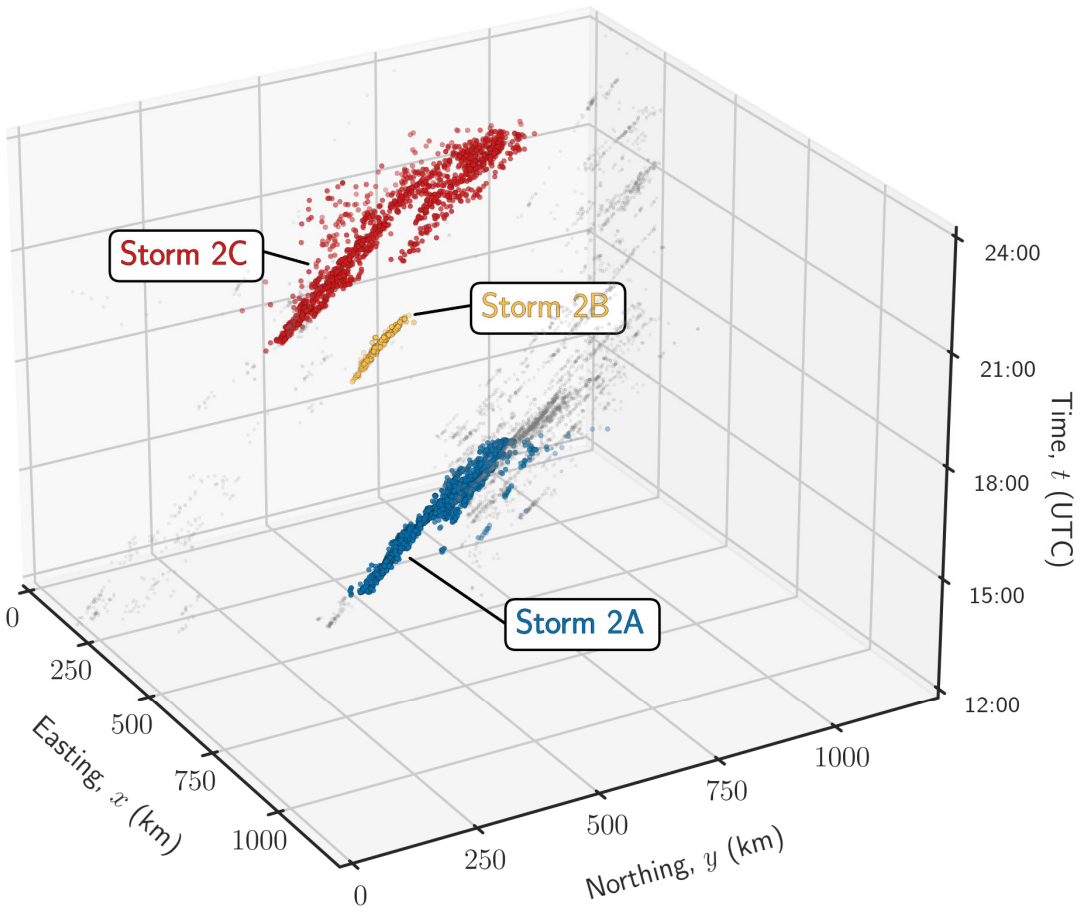

Figure 3.10. Spatial extent and propagation of the lightning strike occurrences as recorded on 1 July 2015 by the ATDnet system [**Gaffard *et al.*, 2008**] and projected onto the OSGB 1936 / British National Grid northing/easting coordinates. Lightning strikes assigned to the identified thunderstorms (Storms 2A, 2B, 2C) correspond to the relevant synoptical analysis [**Lewis and Silkstone, 2017**] and all the remaining unassigned strikes are coloured as per the legend.

In **Figure 3.10**, the separation of the three individual thunderstorms on 1 July 2015 from the unassigned lightning strikes is clearer than the lightning strike clusters on 28 June 2012, as shown in **Figure 3.4**. There are fewer unassigned lightning strikes for 1 July 2015 and the temporal separation is greater than in the first case study, for 28 June 2012. For this second case study, all three storms had the same level of descriptive and radar information available and does not add ambiguity to the assignment of individual lightning strikes to thunderstorms.

For this second case study, the six one-hour intervals of the lightning strikes as assigned to the individual thunderstorms relative to the locations of the three Doppler radar stations are shown

in **Figure 3.11**. These six intervals were chosen to cover the full temporal range of the lightning strikes as produced by the three thunderstorms, but are not continuous due to the significant temporal gap between Storms 2A and 2B.

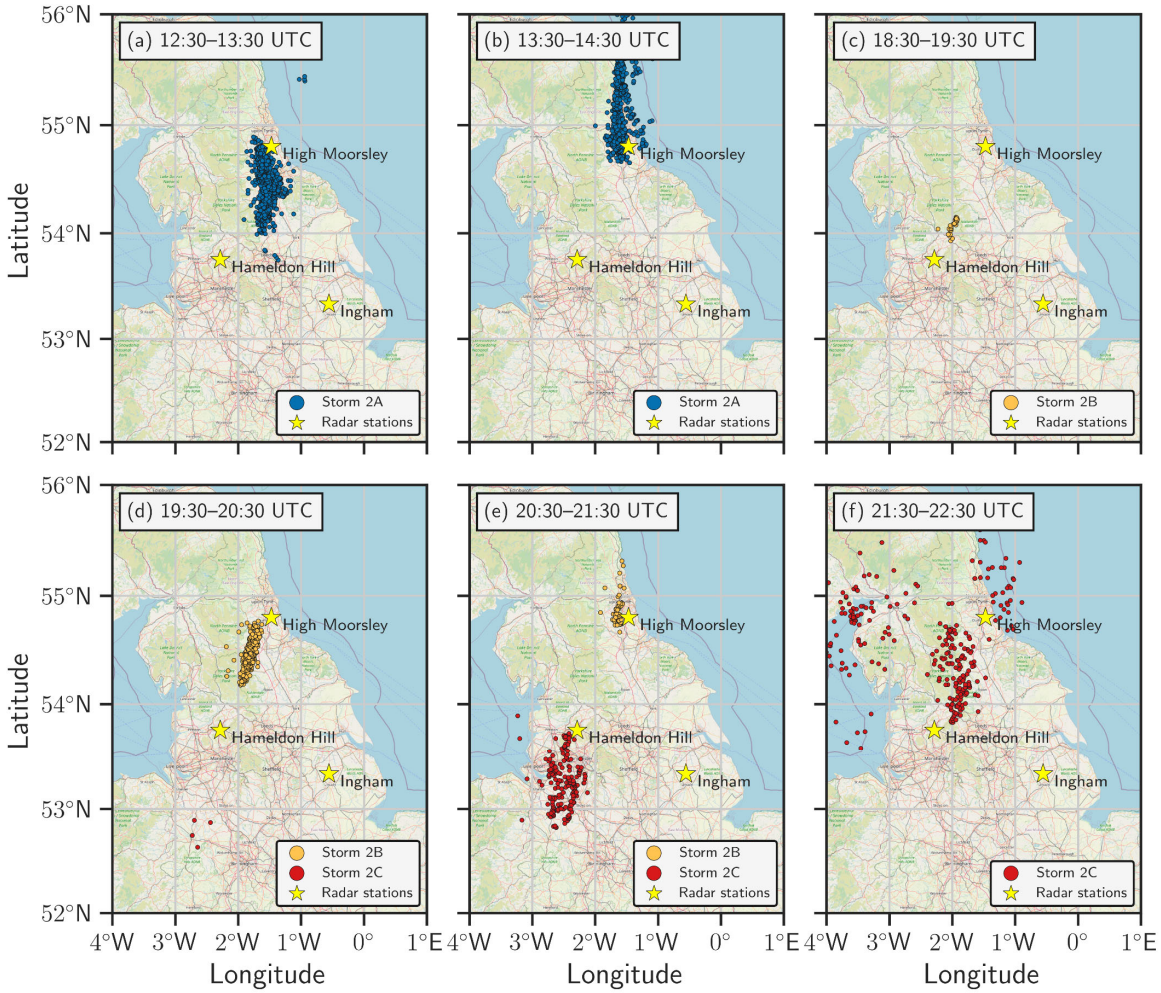

Figure 3.11. Spatial extent for 1 July 2015 of the lightning strike occurrences as recorded by the ATDnet system [**Gaffard *et al.*, 2008**] over times intervals corresponding to the relevant synoptical analysis [**Lewis and Silkstone, 2017**] **(a)** 12:30–13:30, **(b)** 13:30–14:30**, (c)** 18:30–19:30, **(d)** 19:30–20:30, **(e)** 20:30–21:30 and **(f)** 21:30–22:30 UTC. Note that not every one-hour interval is plotted. Only lightning strikes assigned to any of the three selected thunderstorms (Storms 2A, 2B, 2C) are shown. Locations of relevant Doppler radar stations are added as per the legend for direct comparison with synoptic analysis.

All three thunderstorms from 1 July 2015, when viewed in the projected two-dimensional spatial domain, cross northern England in a south-to-north direction. Locations of the Hameldon Hill, High Moorsley and Ingham radar stations allow for direct comparison with relevant figures in the synoptic analysis. The two-dimensional spatial count of lightning strikes

per 1 km^2 for Storm 2A is displayed in **Figure 3.12**. The same plots for Storm 2B and Strom 2C can be found in **Appendix I**.

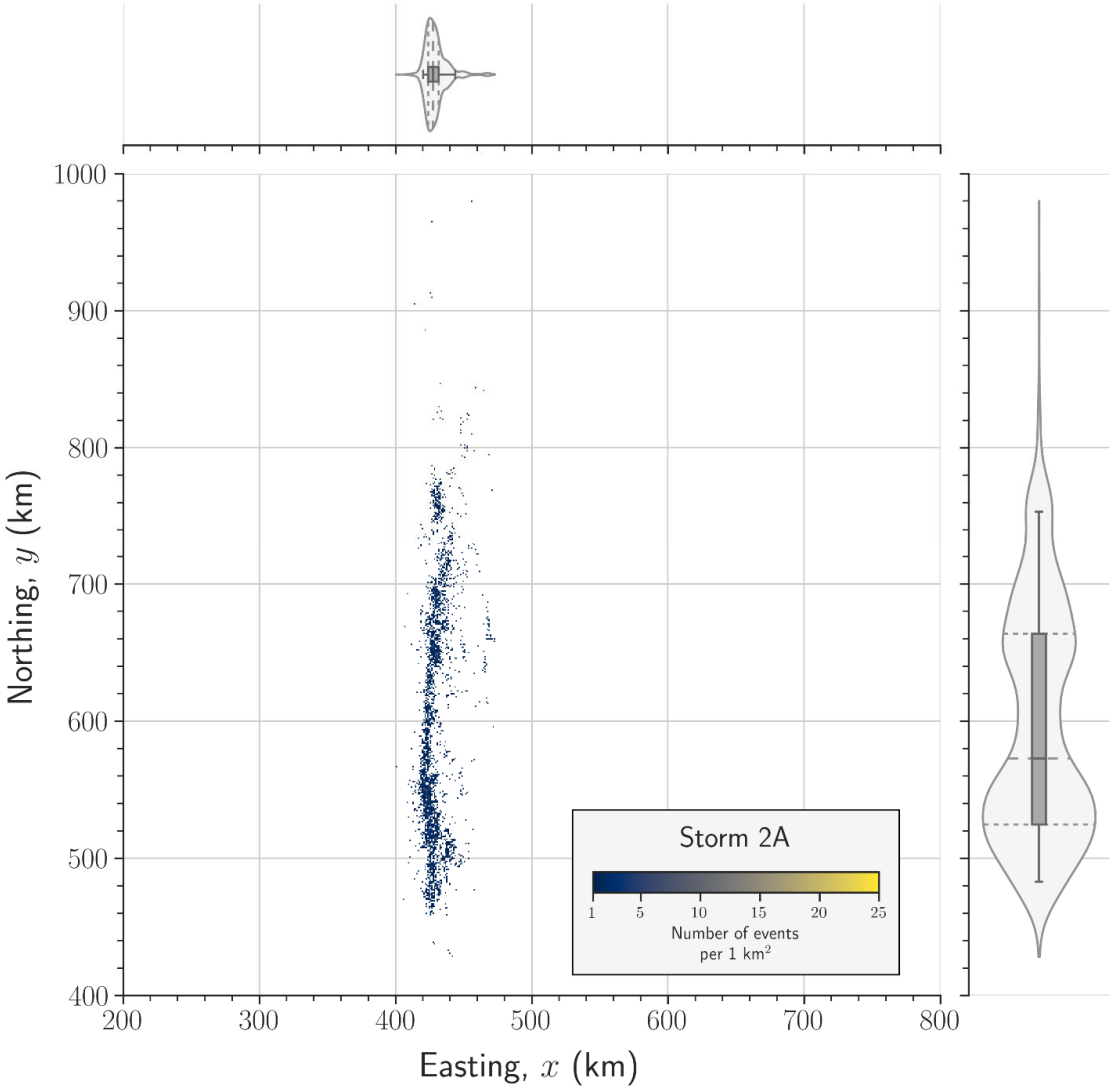

Figure 3.12. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for thunderstorm 2A (1 July 2015). Only bins with more than 1 lightning strike are coloured. Note that the northing and easting extents are the same, but the axes limits differ. The marginal distribution is indicated by overlapping boxplots and violin plots for both the easting and northing values. Boxplots indicate the quartiles, and have whiskers for the 5th and 95th percentiles, while the violin plots include all values.

As Storm 2A travels in the S-N direction, the lightning strikes assigned to Storm 2A cover a significantly larger scale in the northing direction (from 420 km to 980 km), while in the easting direction – from 400 km to 480 km. Areas of relatively higher lightning intensity are located at the beginning of the propagation track of the storm. The marginal distribution of the northing coordinates shows a separation between two areas of increased count values at y = 600 km.

To gain an overview of the spatial and temporal extents for the six identified thunderstorms, a summary table from the marginal distributions of the northing and easting coordinates and the times of occurrence is presented in **Table 3.2**. The occurrence time values are presented as hours starting at 00:00 UTC on 28 June 2012 for the first case study and 00:00 UTC on 1 July 2015 for the second.

Table 3.2. Ranges for spatial and temporal coordinates for the six thunderstorms on 28 June 2012 (Storms 1A, 1B, 1C) and 1 July 2015 (Storms 2A, 2B, 2C).

Date	Storm	Extent of easting values ($x_{max} - x_{min}$) (km)	Easting 5th and 95th percentiles x_5, x_{95} (km)	Extent of northing values ($y_{max} - y_{min}$) (km)	Northing 5th and 95th percentiles y_5, y_{95} (km)	Extent of time of occurrence values ($t_{max} - t_{min}$) (h)	Time of occurrence 5th and 95th percentiles t_5, t_{95} (h)
28	1A	186	351, 460	199	222, 330	4.4	7.0, 9.74
June	1B	153	419, 517	139	273, 363	3.2	10.5, 13.3
2012	1C	136	451, 528	90	293, 356	1.8	12.1, 13.7
1	2A	73	420, 444	562	483, 753	3.2	12.8, 15.5
July	2B	41	403, 424	115	477, 547	1.6	18.5, 19.6
2015	2C	322	296, 455	543	359, 768	4.7	19.9, 23.8

3.3.3. Methods to characterise lightning strike parameters in time and space

This section first introduces the physical variables used as characteristics of lightning strikes (**Section 3.3.3.1**). Three physical variables are then selected for characterisation and inclusion in a spatio-temporal model framework: lightning strike inter-event occurrence time (**Section 3.3.3.2**), lightning strike cluster movement speed (**Section 3.3.3.3**) and spatial spread (**Section 3.3.3.4**). Each of these variable subsections first presents the methodologies used to characterise the variable and then details the approach taken to include the variable into a spatio-temporal point event model framework. Characterisation results for all three variables are presented and analysed in **Section 3.3.4**. The modelling approach is detailed, analysed and examples of synthetic point event datasets representative of lightning strikes are shown in **Section 3.3.5**.

3.3.3.1. Lightning variables

Lightning as a natural hazard has been characterised by various methodologies, depending on the data type and purpose of the particular study. Characteristic variables depend on whether lightning data are analysed as a physical phenomenon (i.e. lightning flash) or as a point event in time and space. Each approach has distinct variables that can be characterised and used for various modelling approaches.

Lightning as a physical phenomenon has extensive research on the underlying physics that determine the initiation, propagation and phenomena related to lightning [**Dwyer and Uman, 2014**; **Rakov, 2013**]. From direct current measurement, the most common variables related to this approach to lightning characterisation include peak current, waveshape, the velocity of the return stroke and the total flash charge [**Chowdhuri *et al.*, 2005**]. Using high-speed cameras, characterisation focuses on continuing current duration, time intervals between strokes, number of strokes per flash, total flash duration, flash multiplicity and inter-stroke intervals [**Ballarotti *et al.*, 2012**; **Saba *et al.*, 2006**]. When analysing lightning strikes as point events recorded by the lightning location systems (described in **Section 3.3.2.1**), the relevant variables are not directly physical. The focus in this thesis is instead on the statistical properties of the event's process (identified at the individual storm) more than on physical characteristics.

Variables such as lightning event (either as lightning flash or strike) count, density per month and/or per area unit, distribution per time unit, mean inter-stroke intervals, and multiplicity have been used to characterise lightning point events. Spatial extent for these variables varies from global [**Mezuman *et al.*, 2014**], continental (e.g. Europe [**Enno *et al.*, 2020**; **Anderson and Klugmann, 2014**]), regional (e.g. eastern Mediterranean [**Yair *et al.*, 2014**], Mediterranean [**Kotroni and Lagouvardos, 2016**], Iberian peninsula grid [**Soriano *et al.*, 2005**]) to national (e.g. Austria [**Schulz *et al.*, 2005**], Poland [**Taszarek *et al.*, 2015**]) scales. In these studies, the spatial domain is gridded by either latitude/longitude (e.g. 0.1° × 0.1° grid [**Kotroni and Lagouvardos, 2016**], 0.2° × 0.2° grid [**Enno *et al.*, 2020**; **Anderson and Klugmann, 2014**; **Soriano *et al.*, 2005**]) or directly on Cartesian grid cells (e.g. 10 km × 10 km cells [**Taszarek *et al.*, 2015**]). Temporal scales for these variables range from hourly, diurnal, monthly, seasonal to yearly characterisation.

These variables are also characterised not only per spatial and/or temporal unit, but also by introducing additional lightning-specific nomenclature, stratification of lightning events or by

considering their relationship with other meteorological variables. For example, the variables can be characterised per thunderstorm day (i.e. at least two flashes occurring within 20 km from the centre of a grid cell from 00:00 to 23:59 UTC [**Enno *et al.*, 2020**]). The lightning events can be stratified by their occurrence over land or sea areas [**Mezuman *et al.*, 2014**]. Variables can also be characterised by their relationship to sea surface temperature [**Kotroni and Lagouvardos, 2016**].

This thesis selects lightning strike variables to characterise the storm behaviour in both spatial and temporal domains. Initial data stratification occurs by grouping the lightning strike data into individual thunderstorms (**Section 3.3.2.2** and **Section 3.3.2.3**). For each thunderstorm, the behaviour in the temporal domain is first characterised by analysing the lightning strike inter-event time variable $\Delta\tau$ (**Section 3.3.3.2**). In the spatial domain, the lightning strike spread is analysed around an estimated thunderstorm movement track. This spread is characterised by easting and northing distances (Δx and Δy) from the movement track (**Section 3.3.3.4**). In the combined spatio-temporal domain, the estimated movement speed v of the individual thunderstorms producing the lightning strikes is characterised (**Section 3.3.3.3**).

3.3.3.2. Variable Δτ: Lightning strike inter-event time

The first physical variable that is characterised and modelled is the inter-event occurrence time $\Delta\tau$, with

$$\Delta\tau_i = t_{i+1} - t_i \text{ for } i = 1, \dots, N-1 \tag{3.1}$$

where t is time, and i an index on t representing discrete instances where events (lightning strikes) occur, noting that these events are not equally spaced in time. For each thunderstorm, the characterisation process of the inter-event time data consists of descriptive analysis of inter-event occurrence time $\Delta\tau$ distribution. Distributions are analysed for each thunderstorm's complete lightning strike dataset and in one-hour time intervals for their duration (**Section 3.3.4.1**),.

To include inter-event occurrence time as an input parameter for the spatio-temporal lightning strike model (**Section 3.3.5.1**), the inter-event occurrence time $\Delta\tau$ data are processed in two steps:

Step 1. **Sampling from an empirical cumulative distribution function eCDF:** synthetic inter-event time occurrence time values $\Delta\tau$ are generated to create a synthetic dataset without any temporal structure.

Step 2. **Re-indexing sampled inter-event times $\Delta\tau$**: temporal structure is added to the synthetic data by using autocorrelation function $C(\Delta k)$ in inter-event occurrence times $\Delta\tau$ arranged in order of occurrence k from the individual thunderstorm datasets.

Inter-event time step 1 follows a methodology for non-parametric sampling from an empirical cumulative distribution [**Ivezić *et al.*, 2014**]. To generate a set of sampled inter-event occurrence time $\Delta\tau$ values from a real-world dataset, the Step 1 sampling methodology consists of three individual operations:

1.1 **Sorting**: the set of inter-event occurrence times $\Delta\tau$ is sorted, and the rank of each time value is used to approximate the empirical cumulative distribution.

1.2 **Inverting cumulative distribution**: the empirical cumulative distribution eCDF has the axes inverted to give the inverse empirical cumulative distribution.

1.3 **Spline fitting**: a cubic spline fit to the inverse distribution is performed. As result, a uniformly sampled x-axis value maps to a y-axis value, approximating the observed inter-event occurrence time probability distribution function.

Inter-event time step 2 introduces temporal structure in the synthetic inter-event occurrence time $\Delta\tau$ dataset from the individual thunderstorms. Such information on the temporal structure can be given by applying the autocorrelation function to the N-1 inter-event occurrence times $\Delta\tau$. The dataset from each individual thunderstorm is set in the order of occurrence k and the formulation of the autocorrelation function $C(\Delta k)$ [**Priestley, 1981**] for a given lag Δk is

$$C(\Delta k) = \frac{1}{\sigma_{\Delta\tau}^2}\frac{1}{N-1-\Delta k}\sum_{k=1}^{N-1-\Delta k}(\Delta\tau_{k+\Delta k} - \overline{\Delta\tau})(\Delta\tau_k - \overline{\Delta\tau}) \qquad (3.2)$$

After Step 1, a synthetic inter-event occurrence time dataset T_S is generated with the same length as any of the individual thunderstorm inter-event occurrence time datasets [T_{1A}, …, T_{2C}]. The methodology used in Step 2 for these two datasets consists of three individual operations:

2.1 **Sorting**: the inter-event occurrence time datasets T_S and corresponding thunderstorm dataset are sorted by their value and ranked.

2.2 **Re-indexing**: the synthetic dataset T_S has its inter-event occurrence times re-indexed to replicate the same ranking of inter-event occurrence times as the thunderstorm dataset.

2.3 **Autocorrelation comparison**: autocorrelation function $C(\Delta k)$ is applied to both datasets to ensure that the temporal structure of the thunderstorm data remains in the synthetic dataset T_S.

3.3.3.3. Variable *v*: Lightning strike cluster movement speed

The second physical variable that can be characterised and modelled is the assumed constant movement speed v of the lightning strikes as a point-event cluster (i.e. representative of an individual thunderstorm). The approach to estimating the movement speed of the cluster is to perform a least-squares plane and linear fits to the lightning strike data. The lightning strike occurrence times t are assumed to depend on the easting x and northing y values as the thunderstorms producing these strikes move through the area and time of the study. Such a relationship between the dimensions can be expressed as

$$\begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \qquad \text{or} \qquad \mathbf{A\beta} = \mathbf{t} \tag{3.3}$$

where $\mathbf{A}$ is the matrix of the available spatial data (easting x and northing y values) as the first two columns. Vector $\boldsymbol{\beta}$ contains the three plane parameters that will be estimated by the best plane-fit. Vector $\mathbf{t}$ contains all the occurrence time t values. Finding the plane parameters in vector $\boldsymbol{\beta}$ is an over-determined problem (the number of equations outnumbers the unknowns), so the left pseudo-inverse of the matrix $\mathbf{A}$ is calculated. The least-squares approach then minimises the vertical (i.e. temporal) distance between the lightning strikes and the plane, returning the best plane-fit that can be thought of as the cross-section of the thunderstorm producing these strikes (**Section 3.3.4.2**).

As the best-fit plane coefficients are estimated from the least-squares fit, the movement speed v can be estimated as the inverse of the dip angle as given by the following

$$v = \frac{1}{\sqrt{a^2+b^2}} \tag{3.4}$$

where a and b are the the plane parameters from vector $\boldsymbol{\beta}$ from **Equation 3.3**. To further characterise the movement speed v after using the least-squares fit, the movement speed is returned with a goodness-of-fit calculation for each thunderstorm. This calculation consists of finding the root square mean error (RMSE) of the residuals between vector $\mathbf{t}$ and $\mathbf{A}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^*$ is the resulting parameter vector as found by the least-squares plane-fit**.**

A least-squares linear fit can also be used to estimate the movement speed v. Such approach (applied for aftershock identification in seismology [**Zaliapin and Ben-Zion, 2013**], tornado touchdown location clustering and movement speed calculation [**Malamud *et al.*, 2016**] and storm movement speed [**Finke, 1999**]) works in the spatio-temporal lag space $\{\Delta s, \Delta t\}$, where the spatial lag Δs and the temporal lag Δt between any two lightning strikes $i = 1, \ldots, N$ and $j = 1, \ldots, N–1$ is the Euclidean distance and the time difference between them respectively

$$\Delta s_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3.5}$$

$$\Delta t_{ij} = |t_i - t_j| \tag{3.6}$$

As $\Delta s_{ij} = \Delta s_{ji}$ and $\Delta t_{ij} = \Delta t_{ji}$, the duplicate data points are removed. It is assumed that in the spatio-temporal lag space $\{\Delta s, \Delta t\}$, the least-squares linear fit with an intercept set to pass through the origin returns the movement speed v from

$$\begin{pmatrix} \Delta t_1 \\ \vdots \\ \Delta t_{\frac{N(N-1)}{2}} \end{pmatrix} v = \begin{pmatrix} \Delta s_1 \\ \vdots \\ \Delta s_{\frac{N(N-1)}{2}} \end{pmatrix} \quad \text{or} \quad \mathbf{T}v = \mathbf{S} \tag{3.7}$$

where $\mathbf{T}$ is the vector of all temporal lags and $\mathbf{S}$ is the vector of all spatial lags for N data points. The procedure that follows is the same as for **Equation 4.3**. As input for the spatio-temporal lightning strike model (**Section 3.3.5.3**), the movement speed v is used as a direct input parameter and assumed constant.

3.3.3.4. Variables Δx and Δy: the spatial spread of the lightning strike as a cluster

The third physical variable that is characterised and modelled is the spatial spread of the lightning strike as a cluster belonging to a particular thunderstorm. To characterise this variable, it is necessary to return an estimate of the track of the thunderstorm to have a reference against which to calculate the spatial spread. The spatial spread in the data can be characterised and modelled by the following two variables: easting and northing distances to the track (Δx and Δy). The calculation of these variables can be summarised in two steps:

1. **Application of uniform filter**: uniform filters are applied to easting x and northing y values in order of occurrence k to return a movement track.

2. **Calculation of easting and northing distances**: for each order of occurrence k, the easting, and northing distances to the track (Δx and Δy) are calculated.

For spatial spread step 1, the movement track is assumed to be non-linear. It is estimated by applying a uniform linear filter on the easting/northing (x, y) coordinates (in the OSGB 1936 / British National Grid) of the lightning strike dataset in the order of occurrence k instead of real-time. The filter is used with a window size of Δw. When window size causes the average calculation at a given k to exceed the length of the northing and easting data vectors, the approach uses duplicates of the first or the last values accordingly. For spatial spread step 2, at each order of occurrence k, the calculation of (Δx, Δy) can be directly done with the easting/northing (x, y) and the estimated movement track coordinates (**Section 3.3.4.3**).

To include the easting and northing distances to the track (Δx and Δy) variables in the spatio-temporal lightning strike model, a methodology for fitting and sampling a bivariate normal distribution is implemented [**Ivezić *et al.*, 2014**]. This methodology is robust with respect to outliers in the data. This is important due to the approach's simplicity in estimating the movement track. To return the parameters of the bivariate normal distribution, it (i) uses the median instead of the mean and (ii) uses the interquartile range to estimate variances.

Once the distribution parameters are found, easting and northing distances to the track (Δx and Δy) are sampled directly from this distribution for implementation in the spatio-temporal lightning strike model (**Section 3.3.5.2**).

3.3.4. Results of variable characterisation

This section presents the results of characterising the three lightning storm variables (inter-event times, movement speed, and distance to movement track) discussed in **Section 3.3.3**. These variable characterisations are for the six lightning strike storms (28 June 2012: Storms 1A, 1B, 1C; 1 July 2015: Storms 2A, 2B, 2C) as detailed in **Section 3.3.2**.

3.3.4.1. Lightning strike inter-event times

It is observed for most of the storms that the beginning and end are visually different in time, and therefore, continuous time intervals from the lightning strike datasets assigned to individual thunderstorms are manually removed. A subset of each storm's strike dataset was used to characterise the lightning inter-event times. The purpose was to remove those values that either (i) contained relatively large inter-event time values or (ii) had longer time periods between consecutive lightning. The original dataset of inter-event times $\Delta\tau$ without further processing is shown first in real-time. If the time of occurrence of the first lightning strike is t_1, the time axis is shifted to indicate the time since first strike t'

$$t' = t - t_1 \tag{3.8}$$

After the removal of time periods that were deemed visually different in time, the end inter-event time $\Delta\tau$ dataset for further characterisation is shown in order of occurrence k. This processing step is illustrated for storms 1A and 2C in **Figure 3.13** (all storms are shown in **Appendix Figure J1**).

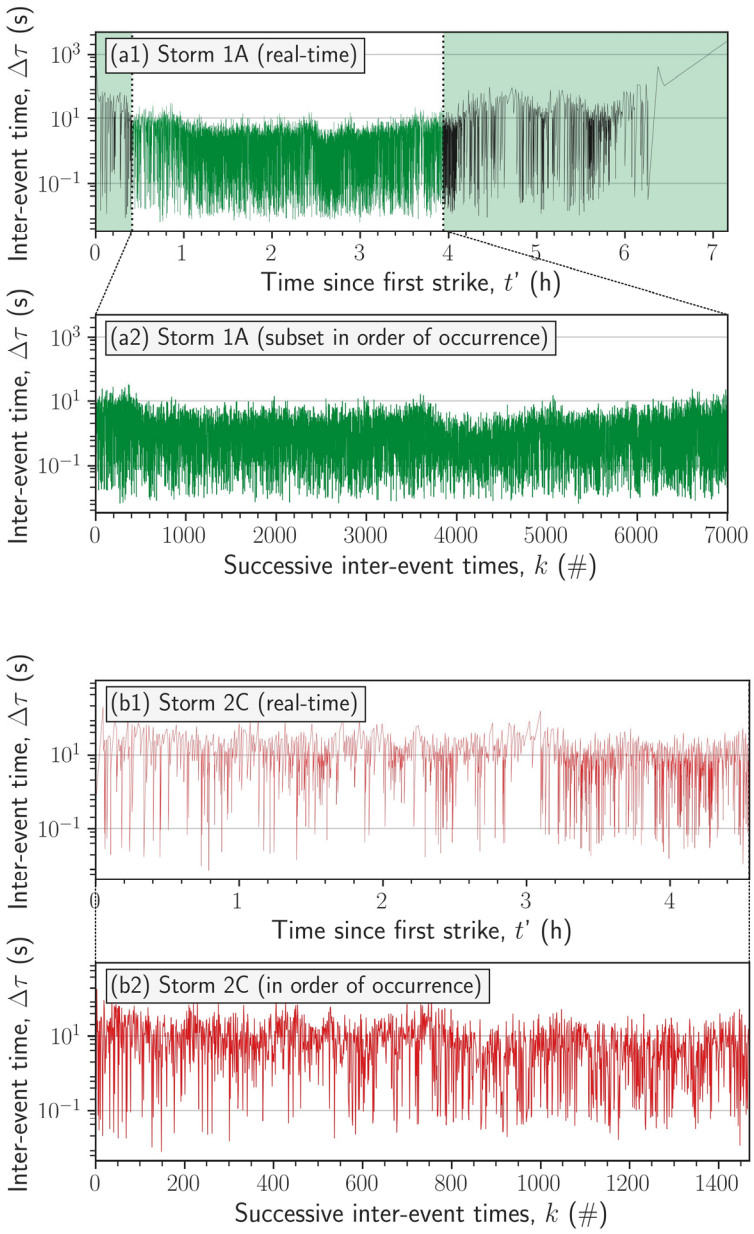

Figure 3.13. Two subplots showing the removal of inter-event time $\Delta\tau$ values in real-time t (removal of the black line in upper subfigure) and the subsequent inter-event time-series in order of occurrence k (lower subfigure) for thunderstorms (**a1**) 1A (real-time, 7,955 values), (**a2**) 1A (subset in order of occurrence, 7,000 values), (**b1**) 2C (real-time, 1,468 values), (**b2**) 2C (subset in order of occurrence, 1,468 values).

In **Figure 3.13**, the selection of a portion of the time series is shown. For Storm 1A in **Figure 3.13 (a1)**, there is an increase in inter-event time values for the first 0.2 hours and after the fourth hour of storm duration. There are also several longer time periods between consecutive inter-event times after the fourth hour. These two time periods were removed to leave a lightning strike subset of n = 7,000 inter-event times for further characterisation. For storm 2C

in **Figure 3.13 (b1)**, there were no clear time periods of increased inter-event times or extended time periods between inter-event times. No time periods were removed for storm 2C.

After the data processing to establish light strike subsets with less ambiguity in their assignment to a particular thunderstorm, the distribution of the inter-event times can be analysed. Violin plots for the inter-event time $\Delta\tau$ subsets of all six thunderstorms are shown in **Figure 3.14**.

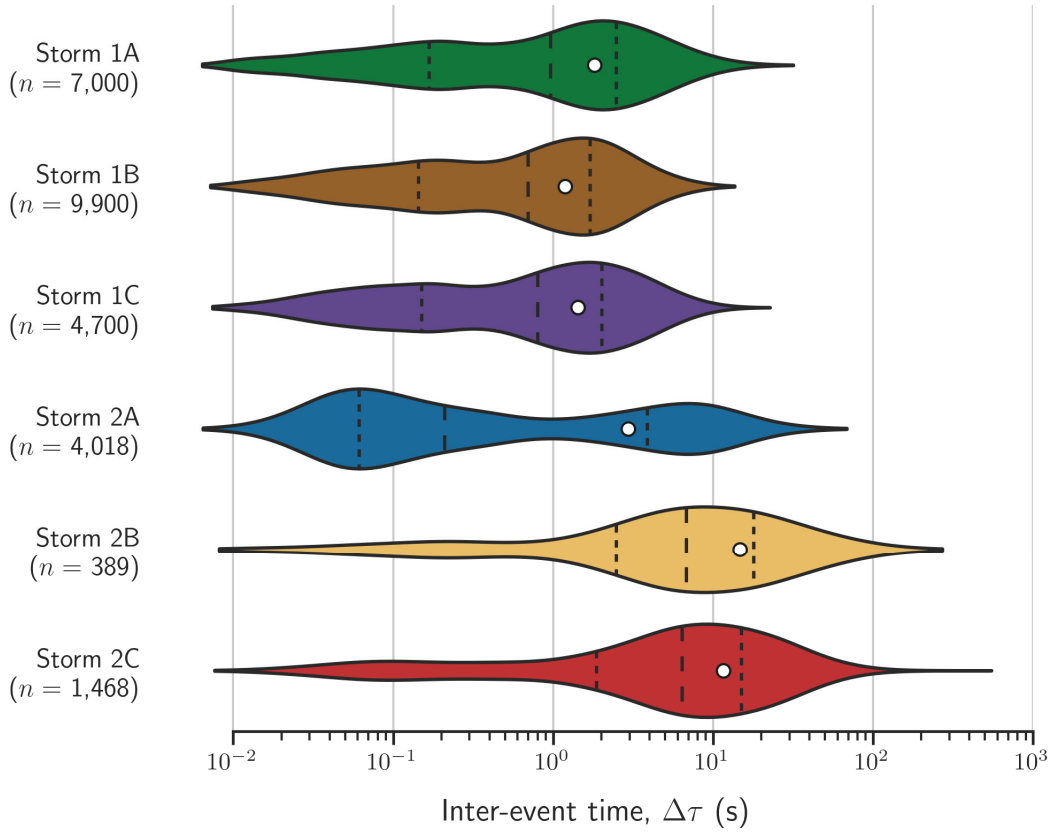

Figure 3.14. Violin plots of inter-event time $\Delta\tau$ distributions for all six selected thunderstorms. For each thunderstorm, the total number of observations after removing inter-event time values (**Figure 3.13**) of the time-series is indicated by n, dotted lines denote the quartiles, and the white dot represents the mean.

In **Figure 3.14**, the inter-event times for all six thunderstorms cover four orders of magnitude, ranging from below 0.01 s to over 100 s. For storms 1A, 1B and 1C, the empirical distributions cover similar ranges of 0.01 s to 20–30 s. They show two peaks, a smaller one for values around 0.2 s and a larger one around 1–3 s, and their quartiles are all within the same order of magnitude over all three storms.

The empirical distributions show more variability for storms 2A, 2B, and 2C. They include larger inter-event time values (more than 100 s) and show varying distribution shapes. Storm 2A has a larger peak around 0.05 s and the smaller one at around 7–8 s, which differs from the other five storms. The empirical distributions for storms 2B and 2C are similar, with a clearer single peak centred at 8–10 s and covering the full range of inter-event times.

For each storm, the empirical distributions of inter-event times are also represented hourly for the full duration of each individual storm in **Figure 3.15**.

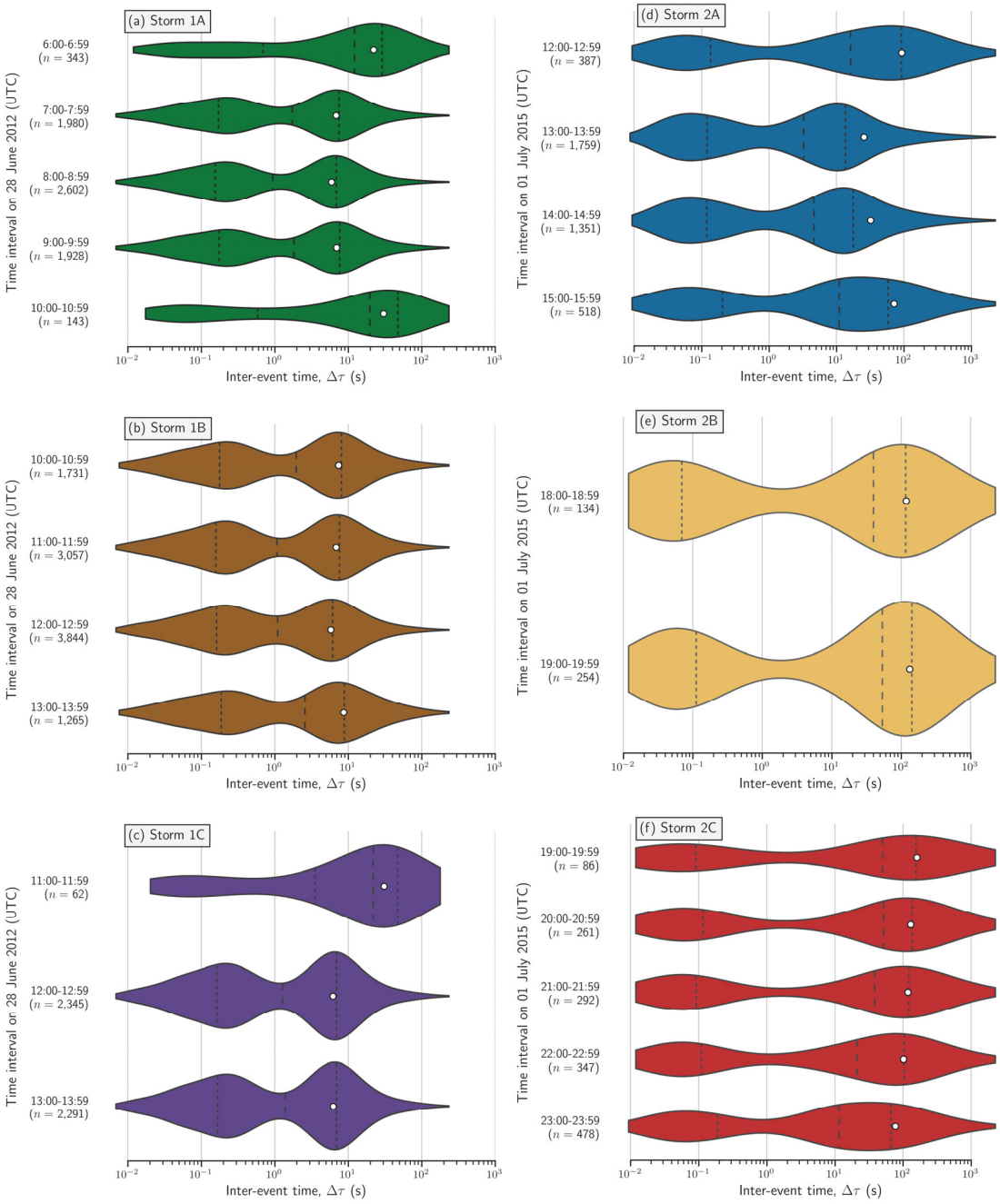

Figure 3.15. Violin plots of inter-event time $\Delta\tau$ distributions for each one-hour interval when lightning strikes were observed after cutting (**Figure 9**) of the time-series for Storm **(a)** 1A, **(b)** 1B, **(c)** 1C, **(d)** 2A, **(e)** 2B and **(f)** 2C. The total number of lightning strikes in each interval is indicated by n, dotted lines denote the quartiles, and the white dot represents the arithmetic mean.

In **Figure 3.15**, the empirical distributions, shown in one-hour intervals, allow for assessing each storm's duration and the changes to inter-event time distributions over this duration. Storms **(b)** 1B, **(e)** 2B and **(f)** 2C, while having varied durations, do not exhibit significant changes to the inter-event time distribution over one-hour intervals. Storms **(a)** 1A, **(c)** 1C and **(d)** 2A exhibit a more significant change for the first or last one-hour intervals. For these storms

and intervals, the larger peak in the distribution is seen at larger inter-event time values (e.g. mean values of 20–40 s for the first and last one-hour intervals compared to 6–8 s for others from storm 1A) with less significant smaller peak compared to other one-hour intervals.

Due to the assignment methodology for individual lightning strikes and the cutting of the datasets, it is not easy to analyse these variations in the first and last one-hour intervals as characteristics of the thunderstorms. For the storms without a noticeable difference, the cutting could have removed significant time periods from the storm that contained the more detailed beginning and end of the storm. These one-hour intervals are also chosen to cover one-hour intervals from the full hour. Changes to the starting time of the one-hour intervals and to the length of the interval time period could alter the distribution of inter-event times at the beginning and end of the storms.

3.3.4.2. Movement speed

For the movement speed variable v (**Section 3.3.3.3**), the results of the least-squares plane-fit to the lightning strikes assigned to storms 1B and 2A using the thunderstorm datasets as detailed in **Section 3.3.4.1** are shown in **Figure 3.16** (the best-fit plane solution is overlaid as a wireframe over the individual lightning strikes and all storms shown in **Appendix K**).

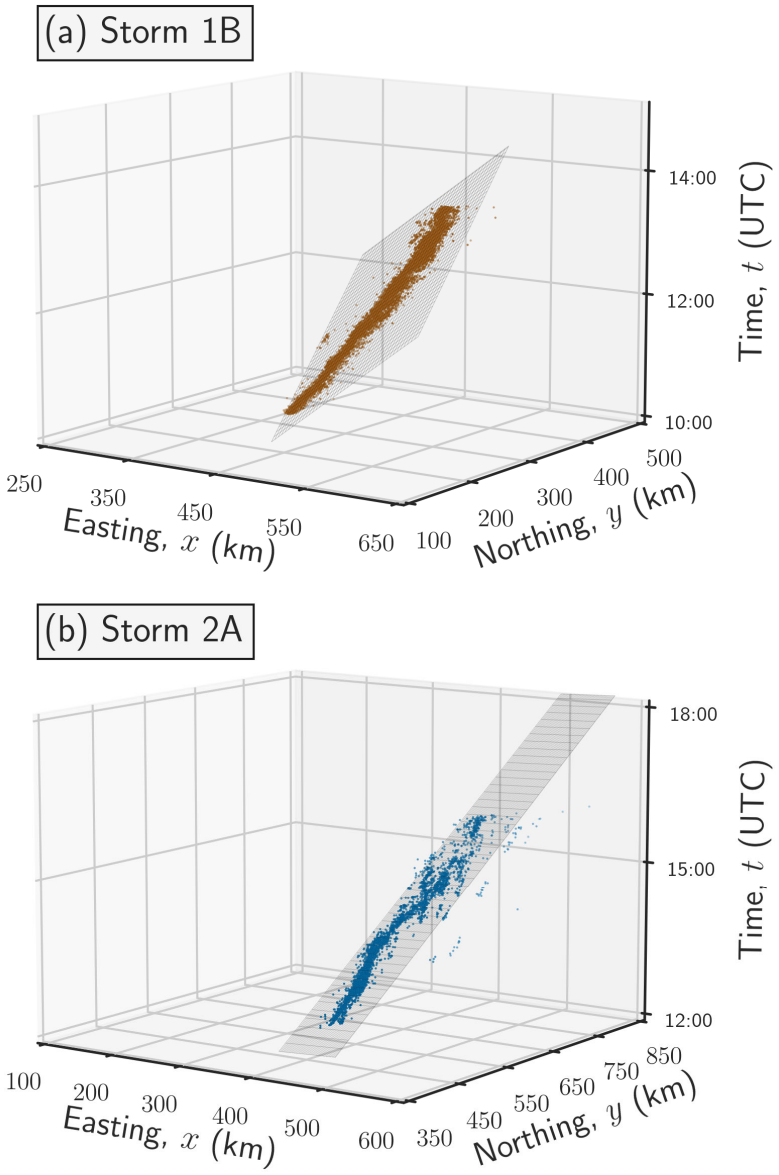

Figure 3.16. Least-squares plane-fit solution to the lightning strike datasets for **(a)** Storm 1B (28 June 2012, n = 9,901 events, v = 47 km h^{-1}), and **(b)** Storm 2A (1 July 2015, n = 4,019 events, v = 108 km h^{-1}). The colour scheme for lightning strikes of each storm is consistent with Figs. 2–5 and the best-fit plane is drawn as grey wireframes. Note that all three axes have different extents for each storm, while the ratio between the axes is constant.

Converting the lightning strike data into spatial and temporal lag space $\{\Delta s, \Delta t\}$, the resulting movement speed variable v estimates from the linear least-squares fits for storms 1A and 2C are shown in **Figures 3.17a** and **3.17b** respectively (all storms are shown in **Appendix L**).

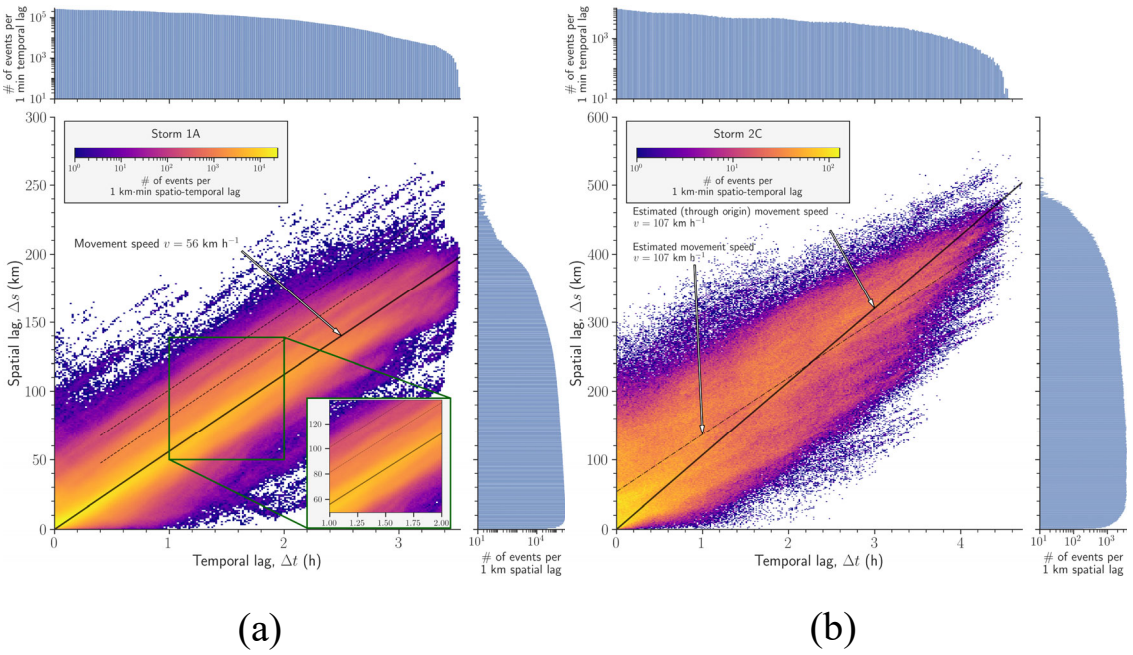

Figure 3.17. **(a)** Storm 1A (28 June 2012, n = 7,001 events, v = 56 km h^{-1}) and **(b)** storm 2C (1 July 2015, n = 1,469 events, v = 107 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The abrupt vertical end of lightning strike data in **(a)** at Δt = 3.45 h occurs due to the removal of inter-event time values (**Figure 3.13**). The solid black line indicates the best-fit linear line passing through the point of origin, while the dash-dotted line in **(b)** indicates the best-fit linear line. The two dashed lines in **(a)** that are parallel to the dark line, both set with identical v = 56 km h^{-1}, indicate other areas of increased count density (bandings of orange pixels as seen in the enlarged subfigure in the lower right corner). These lines are set manually to emphasise the visually observable structure in the data.

The results for the movement speed v calculations from the plane-fit (**Figure 3.16** and **Appendix L**) and linear-fit (**Figure 3.17** and **Appendix M**) and the RMSE of the residuals between lightning strike data and the best-fit plane-fit are indicated in **Table 3.3**.

Table 3.3. Results of least-squares plane-fit and linear-fit solutions to lightning strike data for six UK thunderstorms.

Date	Storm	Movement speed from best plane-fit, v (km h^{-1})	Movement speed from best linear-fit in lag space, v (km h^{-1})	RMSE of plane-fit residuals (h)
28 June 2012	1A	58	56	0.17
	1B	47	48	0.12
	1C	59	60	0.11
1 July 2015	2A	108	101	0.23
	2B	67	66	0.07
	2C	111	107	0.40

In **Table 3.3**, the calculated movement speeds from the two approaches vary between the two storm systems. Storms 1A, 1B and 1C return similar values in the range of 47–60 km h^{-1} and Storms 2A, 2B and 2C show greater variability and higher values, ranging from 66–111 km h^{-1}. The goodness-of-fit estimation for the best plane-fit solutions shows low variability between the storms of the first case study (1A, 1B and 1C). For the second case study (2A, 2B and 2C), the variation is higher with storm 2B having the lowest RMSE of 0.07 hours, while storm 2C has the highest at 0.40 hours.

In **Figure 3.17,** the least-squares linear-fit solution in the spatio-temporal lag space returns more detailed information on the structure of each thunderstorm relative to using just a least-squares plane-fit with a goodness-of-fit estimate (**Figure 3.16**). In the spatio-temporal lag space, the storm's movement speed is initially assumed to pass through the point of origin. When looking at the structure of the 2D count plots (**Figure 3.17a** for Storm 1A, **Figure 3.17b** for Storm 2C and **Appendix M** for other storms), additional structures can be seen in the data that either do not pass through the point of origin and move at a different movement speed.

Initial observation of the 2D count plots show bandings of pixels that indicate lines in the spatio-temporal lag space that do not pass through the point of origin, but have a similar gradient (i.e. movement speed). This would indicate that this larger lightning strike cluster (i.e. thunderstorm) consists of several smaller ones, each moving as part of the cluster with similar movement speeds (as seen by the dashed line examples in the enlarged subfigure for Storm 1A in **Figure 3.17a**).

It is also observed that bandings of pixels with different movement speeds within the same lightning strike dataset exist. The resulting movement speed value also depends on the assumption of the best-fit line passing through the origin (as seen by the dash-dotted and solid line examples for storm 2C in **Figure 3.17b**). Other storms (**Appendix M**) also show structures in the data that would either have different movement speeds based on this assumption or have pixel bandings that indicate smaller clusters in the data.

3.3.4.3. Distance to the movement track

To investigate the distance to the movement track variable (**Section 3.3.3.4**), 2D spatial counts of lightning strikes per 1 km^2 of easting and northing distances for Storm 1A are shown in **Figure 3.18** (all storms shown in **Appendix N**).

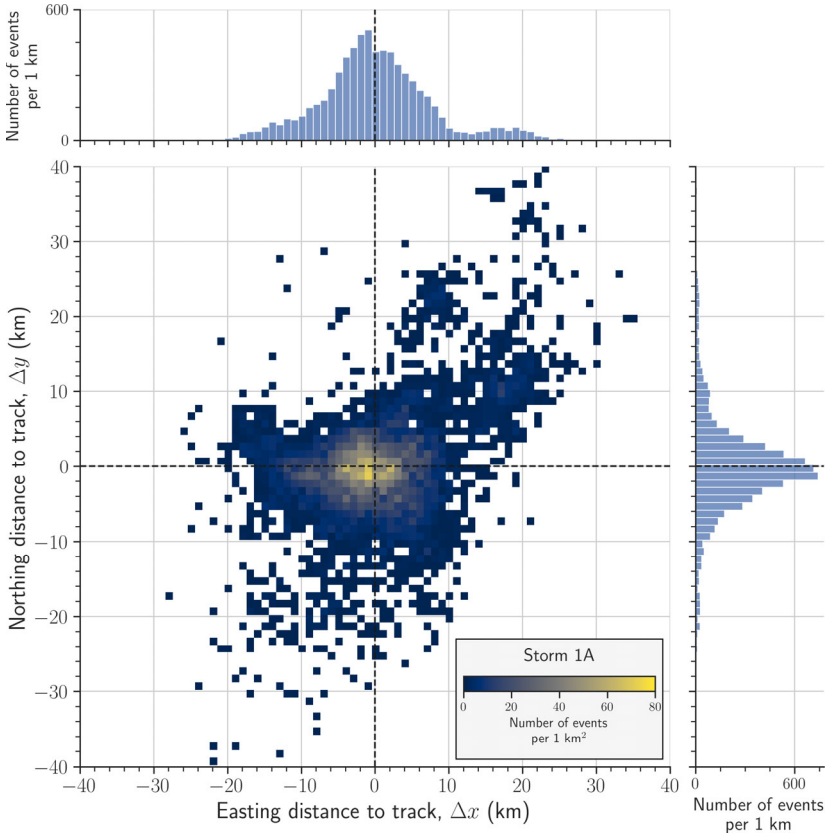

Figure 3.18. 2D spatial count of lightning strikes per 1 km^2 of the easting and northing distances to the movement track in natural time k for lightning strikes in thunderstorm 1A (28 June 2012) for window sizes $\Delta w = 500$ events. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [–40, 40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both northings and eastings.

In **Figure 3.18**, the 2D spatial and marginal bin counts for Storm 1A show that lightning strikes do not exceed 40 km distance to the movement track in either dimension. The 2D spatial count ranges from 0–80 lightning strikes per 1 km^2, while the marginal bin counts reach values of approximately 450 and 650 lightning strikes per 1 km in the easting and the northing domain respectively.

Highest values in both 2D spatial and marginal bin counts are observed around the origin ($x = 0$, $y = 0$) [km] point. The deviation in the 2D spatial and marginal bin counts from the origin is more pronounced in the positive (+,+) and negative (–,–) easting–northing distance quadrants.

A summary of the characterisation results for all three lightning strike variables, as presented in **Section 3.3.4**, is shown in **Table 3.4**.

Table 3.4. Summary of lightning strike characterisation results.

Storm	Movement speed, v (km h^{-1})	Inter-event time distribution		Distance to movement track		
		Quartiles $\Delta\tau_{25}$, $\Delta\tau_{50}$, $\Delta\tau_{75}$ (s)	Mean $\overline{\Delta\tau}$ (s)	Highest 2D bin count per 1 km^2	Range of easting distance Δx_{min}, Δx_{max} (km)	Range of northing distance Δy_{min}, Δy_{max} (km)
1A	58	0.17, 0.95, 2.45	1.79	73	–61, 35	–80, 48
1B	47	0.14, 0.69, 1.70	1.18	163	–41, 26	–34, 15
1C	59	0.16, 0.81, 2.04	1.44	58	–60, 36	–20, 20
2A	108	0.06, 0.21, 3.61	2.85	28	–43, 28	–226, 67
2B	67	2.42, 6.62, 17.05	13.02	7	–10, 25	–37, 35
2C	111	1.87, 6.41, 14.99	11.54	4	–124, 176	–136, 170

3.3.5. Modelling lightning strikes

This section presents an approach to modelling lightning strikes as point events in the line of [**Finke, 1999**], as mentioned in **Section 3.3.3.2**. The approach incorporates the three lightning storm variables (inter-event times, movement speed, distance to movement track) defined in **Section 3.3.3** and characterised in **Section 3.3.4** into a spatio-temporal model. The movement speed variable is set as a constant (including details from **Section 3.3.3.3**), while the approaches for incorporating inter-event time and distance to movement track are detailed in **Section**

3.3.5.1 and **Section 3.3.5.2,** respectively. The model details and procedure for creating a synthetic point event dataset are described in **Section 3.3.5.3,** with a discussion on the limitations and uncertainties of this approach detailed in **Section 3.3.5.4**.

3.3.5.1. Sampling inter-event times

The approach to creating a dataset of inter-event time to use in modelling lightning strikes (**Section 3.3.3.2**) involves specifying the real-world lightning strike dataset that will control two phases of the approach :

1. **Generation of an empirical distribution**: selecting real-world lightning strike data from which to generate an empirical distribution function eCDF of inter-event time values $\Delta\tau$.
2. **Introduction of autocorrelation**: using individual thunderstorms to establish the temporal structure in the data representative of lightning data.

The first phase of the inter-event time generation is done to sample inter-event times without any temporal structure. They are sampled from a larger multi-storm dataset to have values that are not representative of just a single thunderstorm producing these events, but of a particular category of thunderstorms. From the synoptic analyses [**Lewis and Silkstone, 2017**; **Clark and Webb, 2013**] on the selected six thunderstorms (discussed in **Section 3.3.2**), these thunderstorms are categorised as supercell (storms 1A, 1B, 1C, 2B and 2C) or single-cell (storm 2A). For example, the supercell thunderstorms can be selected and the individual datasets combined to create a supercell dataset from which to sample inter-event times.

A comparison between the cumulative distribution functions (CDFs) for all six individual thunderstorms and the common supercell dataset is shown in **Figure 3.19**. The inter-event times are sampled from these CDFs following the methodology described in **Section 3.3.3.2**.

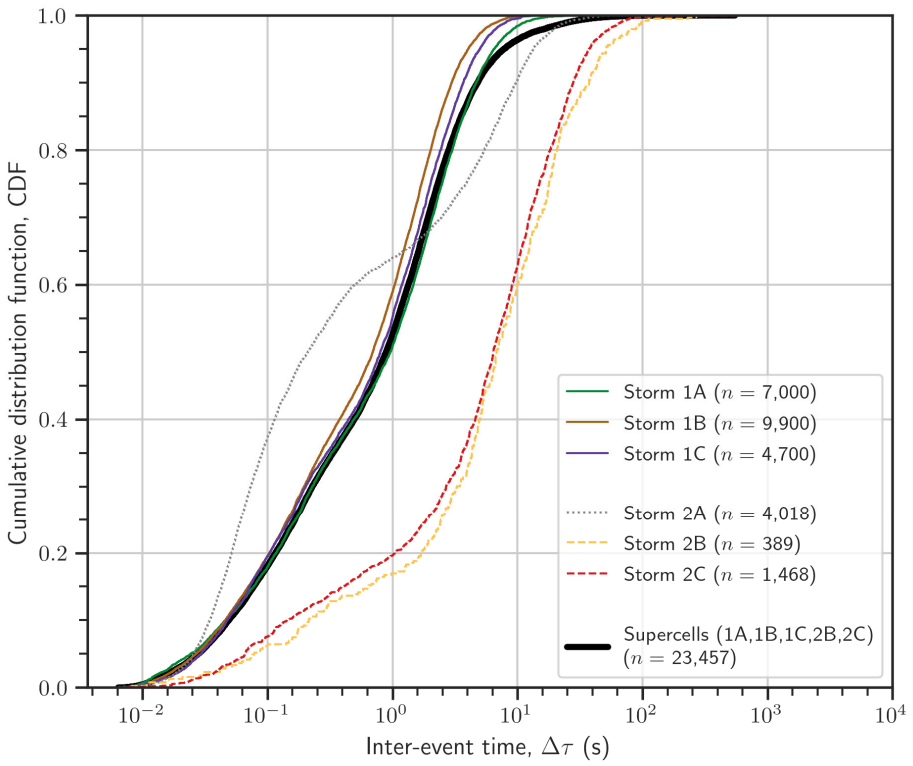

Figure 3.19. Cumulative distribution function (CDF) for the inter-event time $\Delta\tau$ for all six identified thunderstorms and the combined supercell thunderstorm dataset. Individual storms within the storm-system recorded on 28 June 2012 (Storms 1A, 1B and 1C) are indicated by solid lines, while those from storm-system recorded on 1 July 2015 (Storms 2A, 2B and 2C) are indicated by dashed lines. Storm 2A in indicated by a grey line due to being single-cell thunderstorm and is not added to the supercell dataset (black solid line).

Figure 3.19 shows the varying CDFs between supercell and single-cell (storm 2A as dashed grey line), showing two clearer distribution peaks for storm 2A. This is expected from the observed inter-event time distribution in **Figure 3.14** and **Figure 3.15**. The difference between the CDFs of the supercells from both storm-systems is also evident. For the storm system recorded on 28 June 2012 (storms 1A, 1B and 1C), the CDFs shows similar behaviour to the storm system recorded on 1 July 2015 (storms 2B and 2C) but have a distribution peak smaller by an order of magnitude.

Figure 3.19 also illustrates the contrast in the number of inter-event times *n* recorded during both storm systems. Storms 1A, 1B and 1C have a combined number of inter-event times n=21,500, while the combined supercell dataset has n=23,357 inter-event times. The supercell CDF closely resembles that of storms 1A, 1B and 1C with a heavier tail due to the presence of inter-event time data from storms 2B and 2C.

The second step of the inter-event time generation is sorting and re-indexing a sampled inter-event time dataset from the first phase (detailed in **Section 3.3.3.2**). This phase introduces temporal structure in the sampled inter-event time dataset and completes creating an inter-event time dataset representative of real-world lightning data. The results of this two-phase generation procedure for creating synthetic inter-event time datasets representative of storms 1A and 2C are shown in **Figure 3.20** (all storms are shown in **Appendix N**).

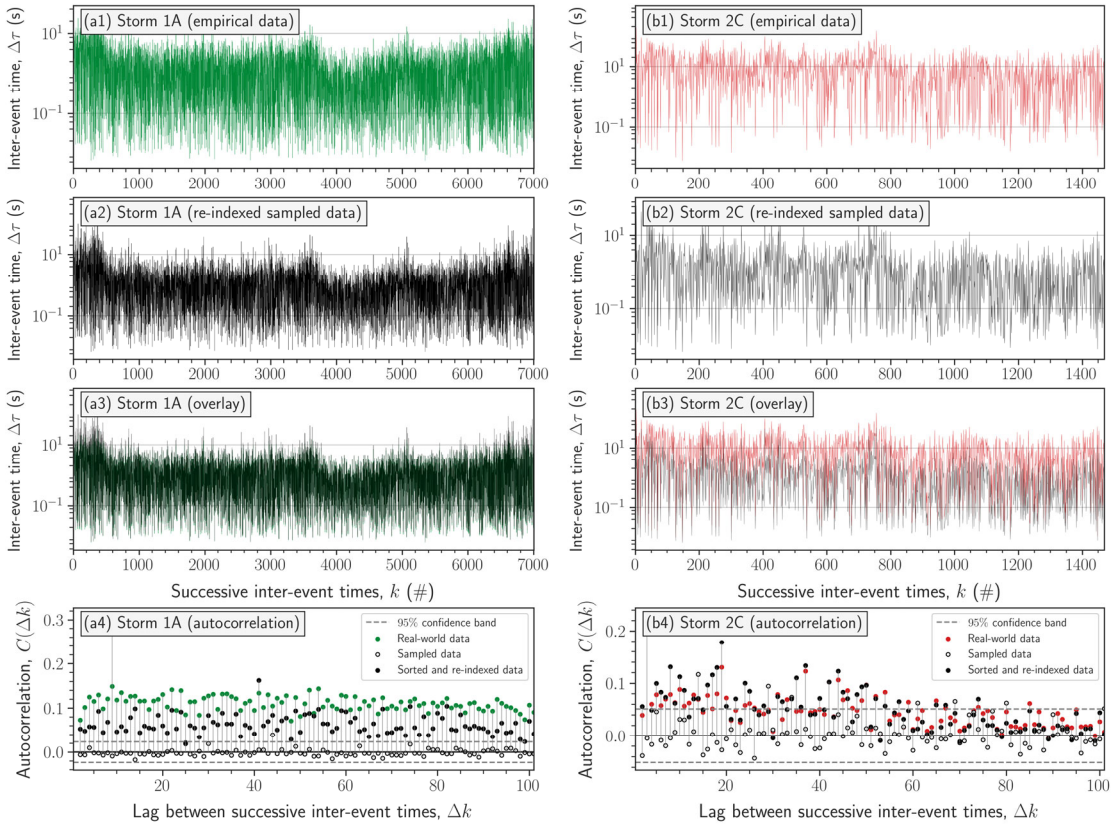

Figure 3.20. Sampling, sorting and re-indexing procedure based on the combined supercell inter-event time $\Delta\tau$ dataset and thunderstorm **(a#)** 1A, **(b#)** 2C. The first subfigure shows the original cut inter-event time time-series in natural time k. The second subfigure shows this sampled time-series being sorted and re-indexed based on the individual thunderstorm, where the indices of sorted time-series coincide between the two time-series. The third subfigure shows the overlay between the original time-series and the sorted and re-indexed one in natural time k. The fourth subfigure shows the overlay between the original time-series and the sorted and re-indexed one in natural time k.

In **Figure 3.20**, the sampling, sorting and re-indexing procedure has returned synthetic inter-event time datasets that are representative of supercell thunderstorms in value and representative of individual thunderstorms in their temporal structure. The difference between the inter-event time values between storms 1A, 1B, 1C and storms 2B and 2C is evident in (**b3**). The sampled, sorted and re-indexed dataset (grey line) have lower inter-event time values than original storm 2C (red line), which is expected from the discussion of **Figure 3.19**.

3.3.5.2. Sampling distances to the movement track

The approach to creating a dataset of distances to the movement track is detailed in **Section 4.3** and the results of this sampling methodology is shown for storm 1A in **Figure 3.21** (all storms are shown in **Appendix P**).

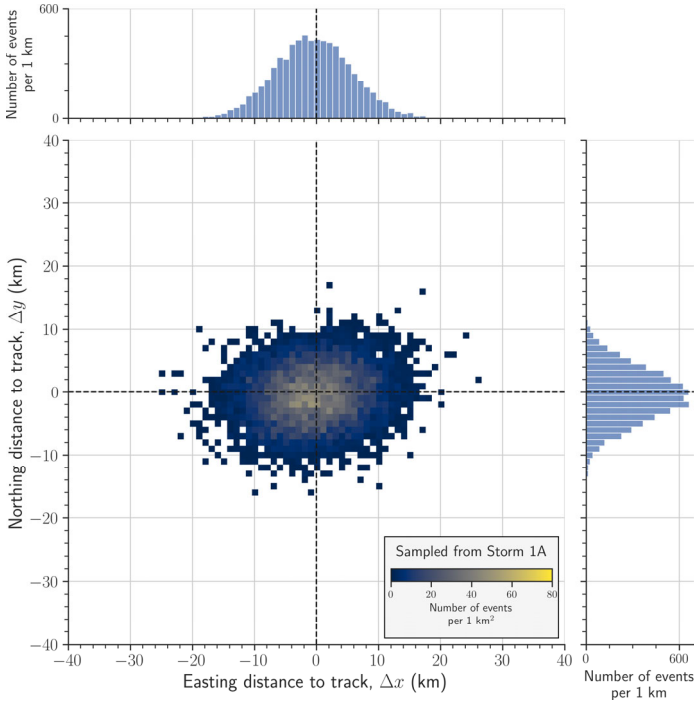

Figure 3.21. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for a bivariate Gaussian distribution based on storm 1A with window size $\Delta w = 500$ events. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [–40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions and the parameters of the bivariate Gaussian distribution are (μ_x=–0.46 km, μ_y=–0.61 km, σ_{xx}=37.25 km^2, σ_{yy}=17.83 km^2, σ_{xy}=2.33 km^2).

In **Figure 3.21**, the 2D and marginal counts of lightning strikes show clear contrast to **Figure 3.18**. This sampling approach, by using a bivariate Gaussian distribution, has not captured the minor peaks in northing and easting distance to track values. For marginal counts, both easting and northing dimensions reach similar values to the real-world data, but the 2D counts reach values of only 50–60 events per 1 km^2.

3.3.5.3. Creating a representative spatio-temporal model

A spatio-temporal model of point events produced by a moving point source is defined to generate synthetic lightning strike datasets. This model would incorporate lightning strike variables as constants or sampled datasets. For the sampled variables, the procedures for

generating inter-event times and distances to the movement track were detailed in **Section 3.3.5.1** and **Section 3.3.5.2,** respectively.

This spatio-temporal model generates a point event dataset representative of lightning strikes (point events) produced by a single thunderstorm (moving point source). When combining multiple runs of this model, the point event dataset would be representative of a storm-system. A single run of this model can be described as the following procedure:

- **Set constant variables:**
    - Initiation point (x_0, y_0, t_0).
    - Movement speed v.
    - Movement direction α.
- **Set sampled variables:**
    - Generate inter-event time dataset.
    - Generate distance to the movement track dataset.

1. **Place a point event**: around the initiation point, place a point event according to the first generated easting Δx_1 and northing Δy_1 distance to the movement track values.
2. **Move along the track**: starting from the initiation point, move in the movement direction with the set movement speed for a time period equal to the first inter-event time value $\Delta \tau_1$.
3. **Repeat steps 1 & 2 until exhausting the sampled variables.**

From this procedure and discussion from **Section 3.3.5.1** and **Section 3.3.5.2**, it is necessary to select a particular thunderstorm when creating a synthetic lightning strike dataset to set the sampled variables. A single run of the model to produce a lightning strike dataset representative of storm 1A is shown in **Figure 3.22a** and this run with an additional two runs of the model to form a three-storm-system shown in **Figure 3.22b**. For the two additional model runs, the bivariate Gaussian distribution for the distance to the movement track variable is set with parameters (μ_x=0, μ_y=0, σ_{xx}=10 km, σ_{yy}=10 km). The correlation parameter for the two additional runs is set at ρ=0 and ρ=0.9. No other of the constant or sampled variables are changed for these model runs.

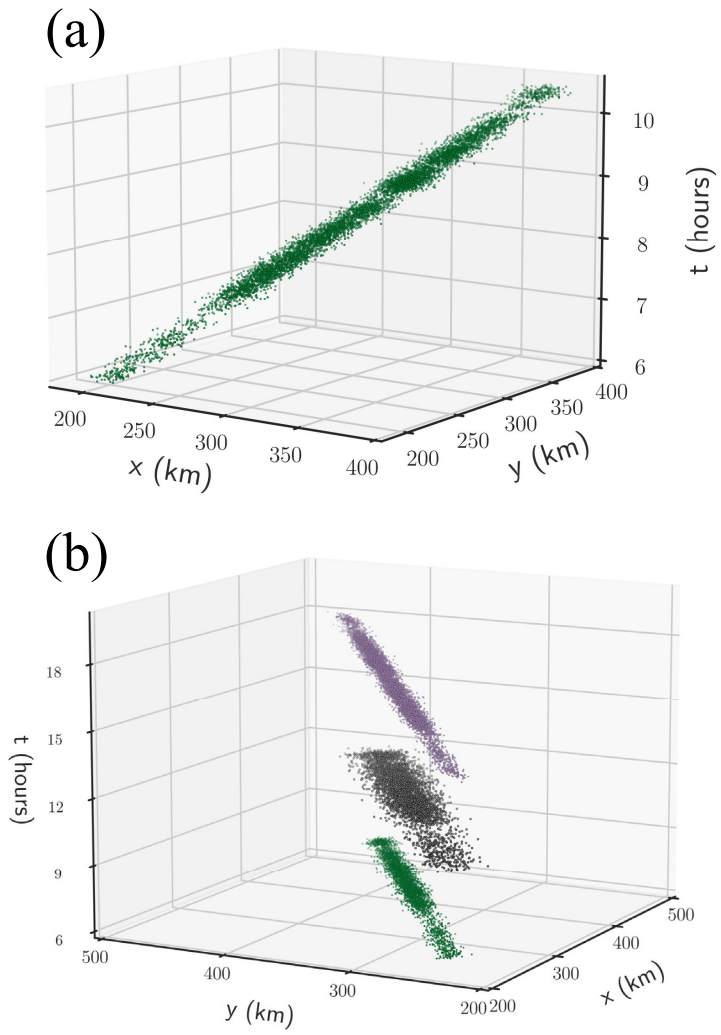

Figure 3.22. (a) A single run of the spatio-temporal model to produce a lightning strike dataset representative of storm 1A and a **(b)** synthetic three-storm-system with varying spatial spread parameters. In **(a)** initiation point is set at (x_0=200 km, y_0=200 km, t_0=6 h). The movement speed value is set at v = 58 km h^{-1} as per the characterisation of the movement speed variable (**Figure 3.13** and **Appendix Figure J1**). Inter-event times are sampled, sorted and re-indexed as per **Figure 3.20 (a1–a4)**. Easting and northing distances are sampled from the bivariate Gaussian distribution shown in **Figure 16**. The first (green) storm is identical to **(a)** with bivariate Gaussian distribution parameter set from parameter characterisation (μ_x=–0.46 km, μ_y=–0.61 km, σ_{xx}=37.25 km^2, σ_{yy}=17.83 km^2, σ_{xy}=2.33 km^2). The second (dark grey) storm is based on bivariate Gaussian distribution with parameters (μ_x=0, μ_y=0 , σ_{xx}=10 km, σ_{yy}=10 km, ρ=0). The third (purple) storm has parameters (μ_x=0, μ_y=0 , σ_{xx}=10 km, σ_{yy}=10 km, ρ=0.9). For all storms, the movement speed is kept the same at v=58 km h^{-1} and inter-event time distribution is representative of storm 1A.

In **Figure 3.22**, the flexibility of the model is demonstrated. A model run representative of a lightning strike dataset produced by a thunderstorm is compared with additional runs where only one lightning variable is changed to examine the visual effect of end-member values for the particular variable. Another example, where additional model runs are representative of

storms 1B and 1C are generated to create a synthetic storm-system representative of the first case study (**Section 3.3.2.2**), is shown in **Figure 3.23**.

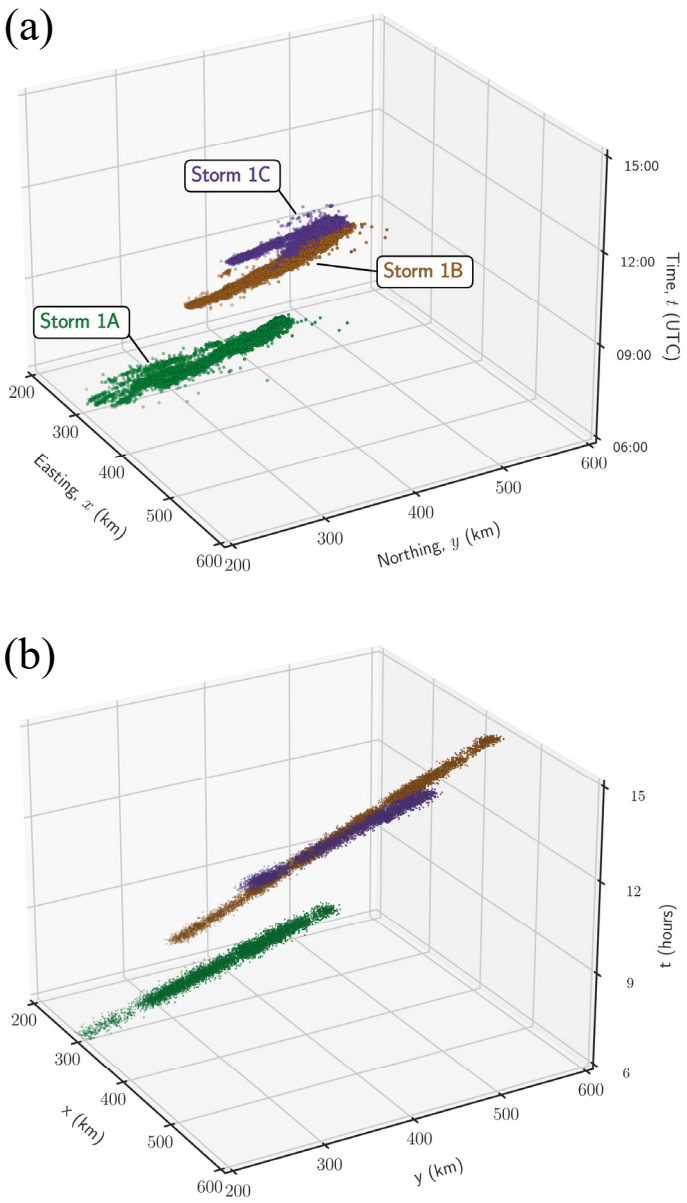

Figure 3.23. A synthetic three-storm-system **(b)** that consists of individual synthetic storms that would be representative of Storms 1A, 1B and 1C from the case study 1 in **(a)**. The movement speed values, re-indexed inter-event time values and spatial spread parameter distributions as selected as per Table 3 for each individual storm.

In **Figure 3.23**, the synthetic storm-system **(b)** is assigned both constant and sampled variables so as to be as representative of the lightning strike datasets from case study 1 **(a)**. The synthetic lighting strike datasets propagate further in the spatial and temporal domains than the real-world datasets they are based on. As this is case study 1, the effect of the heavy tail of the supercell inter-event time dataset has been discussed in **Figure 3.19**. The assumption of

straight-line propagation of the moving point source (thunderstorm) producing the individual point events is noted, which is a simplification as discussed in the modelling approach to the spatial spread of the lightning strikes (**Section 3.3.3.4**).

3.3.6. Limitations of the characterisation and modelling approach

In this study, the characterisation (**Section 3.3.4**) and modelling (**Section 3.3.5**) aspects of this study were based on two selected case studies (**Section 3.3.2**). Each case study described a storm-system consisting of three individual thunderstorms over a single day of the observed increase in storm activity. To characterise and model these thunderstorms, the lightning strike data for both days was gathered by the ATDnet LLS over the United Kingdom. There are no noted differences in data quality between the two dates and the initial uncertainty in lightning strike assignment results from the assignment procedure (detailed for Storm 1B in **Section 3.3.2.2**). The synoptic analyses for both case studies do not explicitly state the formation and termination points for each thunderstorm. There is also merging between the identified supercell thunderstorms and smaller cells, creating ambiguity in the assignment procedure. The pre-processing step of removing inter-event time values for the characterisation (**Section 3.3.4.1**) removes the ambiguity of exact formation and termination points.

In both case studies, five out of the six identified thunderstorms were categorised as supercells. In other characterisation studies (e.g. [**Enno *et al.*, 2020**; **Mezuman *et al.*, 2014**; **Anderson and Klugmann, 2014**]), the characterisation added to lightning climatology research without emphasis on particular thunderstorm type. In [**Anderson and Klugmann, 2014**], the diurnal flash density characterisation included case study 1 (**Section 3.3.2.2**), but without separation of characterisation results for individual thunderstorms. Although five out of the six thunderstorms in these case studies were supercells, there were clear differences in their characterisation results (e.g. empirical distribution differences in **Figure 3.14**). While applicable to these case studies, it is acknowledged that the characterisation results cannot be generalised to describe thunderstorms as a natural hazard.

To address this lack of generality for thunderstorm characterisation, the creation of a supercell inter-event time $\Delta\tau$ dataset (**Figure 3.19**) was done for modelling of the inter-event time variable $\Delta\tau$. The limitation for modelling this particular physical variable is the dataset size discrepancy between the case studies (i.e. first case study has n = 21,600, while the combined supercell dataset has n = 23,457 lightning strikes). The values sampled from this supercell

dataset are more representative of the first case study, while the larger inter-event time $\Delta\tau$ values of the second case study created a heavier tail for the combined distribution.

The characterisation results return similar value intervals as other studies that were focused on physical variable characterisation. Inter-event time values range from below 0.01 s to over 100 s for individual storms from both case studies. For storms 1A, 1B and 1C, the empirical distributions cover a range of 0.01 s to 20–30 s. Distributions show two clear peaks – a smaller one for values around 0.2 s and a larger one around 1–3 s and their quartiles are all within the same order of magnitude over all three storms. In [**Ballarotti *et al.*, 2012**], a study conducted over Vale do Paraíba, Brazil using high-speed cameras returned an arithmetic mean for the inter-stroke time interval of 92 ms. In [**Saba *et al.*, 2006**], over southern Brazil, a dataset of 608 inter-stroke time intervals had the arithmetic mean as 83 ms. Over different study areas, like the eastern Mediterranean and Israel [**Yair *et al.*, 2014**] and Arizona, the United States [**Stall *et al.*, 2009**], the mean inter-stroke interval was found to be is 93 ms and 98 or 84 ms (depending on whether strikes that used a pre-existing lightning channel or created new ground contacts) respectively. These results are in good agreement with the distribution of inter-event times found for storms 1A, 1B, 1C.

Movement speed values were found to be quite similar for storms 1A–1C, ranging between 47–60 km h^{-1} from the best plane-fit and the linear fit in lag space. For storms 2A and 2C, the two methodologies returned significantly higher movement speed values of 101–111 km h^{-1}. Storm 2B then had a value range of 66–67 km h^{-1}, similar to the first case study. Currently, there is no awareness of any publications that have estimated the movement speed of a thunderstorm from lightning as a point event dataset. Such estimates are usually based on two-dimensional radar reflectivity data.

As an example, the automated cell detection algorithm KONRAD [**Lang, 2001**] is used in [**Wapler and James, 2015**] to calculate the movement speed of thunderstorms over Central Europe spanning from 2007–2012. They used 29 categories to classify the observed thunderstorms based on regional synoptic characteristics over Central Europe. Over these defined thunderstorm categories, the median movement speeds (within each category) of all the observed thunderstorms varied from 21–63 km h^{-1}. When considered separately, the cell motions histograms show that individual cells reached movement speeds above 100 km h^{-1} relatively frequently (frequency of 0.02 for the 'Cyclonic Westerly' category), depending on the particular thunderstorm category.

Over North Dakota, the United States in [**Mohee and Miller, 2010**], the thunderstorm climatology was presented for a time period of 2002–2006. Radar reflectivity data were used together with automated surface observation systems. The thunderstorms were not categorised into various synoptic categories, but were instead stratified temporally based on the month. The movement speed values ranged from 15.4–20.4 m s^{-1} or 55–73 km h^{-1}. Both studies confirm that for the supercell thunderstorms in the first case study, their movement speed values fall within reasonable range found in other studies, while storm 2A and 2C have movement speed values still within a reasonable range, depending on their characteristics. Further expertise from a meteorology perspective would be necessary for any further assessment and analysis.

For the distance to the movement track variable, the bivariate Gaussian distribution fit to model the data was selected for simplicity, similar to [**Finke, 1999**]. Currently, there is no awareness of any publications estimating the distance to the movement track. The movement track estimates are usually done by temporal windows. In [**Strauss *et al.*, 2013**], additional precipitation data were used to estimate storm cells and their centroid locations over 10-minute intervals were used as point along the movement track. In [**Kunz *et al.*, 2020**], 26,000 severe convective storm tracks between 2005–2014 were estimated from two-dimensional radar reflectivity data. They estimated the storm tracks by using a modified version of the cell-tracking algorithm TRACE3D [**Handwerker, 2002**]. Such approaches were deemed as too complex as the objective of the characterisation process is to provide an estimate for the spatial spread parameter for the models.

The characterisation of the spatial spread variable (**Figure 3.18**) showed a non-trivial empirical distribution, but no evaluation of other potential distributions was performed on this data and is a potential future research direction. There was also no attempt to introduce additional complexity in the model by modelling the movement track as a more complex object than a straight line (**Figure 3.23**), although such an attempt was made to characterise the variable (**Section 3.3.3.4**).

The motivation for modelling lightning strikes was to create synthetic point event datasets generated by a moving point source representative of lightning strike data. Three physical variables were identified, which can be characterised to establish representative values for each variable (**Section 3.3.3**). Modelling and sampling approaches were proposed for the inter-event time and distance to the movement track variables (**Section 3.3.5**). No direct sampling of characterised variables was implemented in the spatio-temporal model of lightning strikes as

point events [**Finke, 1999**]. Inter-event times $\Delta\tau$ were determined by uniformly sampling occurrence times t and spatial spread was given by the variance of the two-dimensional Gaussian cloud. This study establishes a direct way to produce synthetic lightning strike data representative of the particular case studies.

The approach to designing the spatio-temporal model was selected to have complete control over the constant and sampled variables as the model can generate any point event datasets with a moving source, depending on the input parameters. The model also has clear procedures for propagating the moving point source (i.e. thunderstorm) and the placement of point events along this point source (i.e. lightning strikes). The simplicity of such a model is acknowledged, but this allows for it to be extended to other point event natural hazards produced by a moving point source and for easier assessment and evaluation of any applications of these datasets.

3.4. Simulation study of synthetic models and benchmark datasets

Before application of a particular methodology to a real-world dataset, benchmark datasets are established. This allows for the assessment of how particular methodologies perform on synthetic datasets with varying properties, dependent on the necessities of the application. In the context of this thesis, the benchmark datasets for point events consist of four distinct models that are detailed in **Section 3.4.1**. Simulation study for the point event data type consists of these models with variations in their input parameters. For trajectories, the two distinct models representing random and homogeneous trajectory datasets are detailed in **Section 3.4.2**. Results of the selected spatio-temporal clustering methodologies as applied to the datasets in the simulation study are shown in **Section 4.1** and **Section 4.2** for point events and trajectories respectively. This section focuses on the numerical implementation and visual presentation of the models and the selected benchmark datasets.

3.4.1. Point event benchmarks

For the point event data type, the four distinct models and the corresponding synthetic datasets are generated using various point process models [**González *et al.*, 2016**; **Wiegand and Moloney, 2004**] and the lightning strike modelling approach detailed in **Section 3.3.5**. Depending on the objective of the application, the point process determines the spatio-temporal patterns of the point events and the interaction between their spatial and temporal components. All synthetic datasets are created within a three-dimensional space-time rectangular cuboid

with its dimensions given by $\{X, Y, T\}$. X and Y represent the lengths of the two spatial dimensions, while T represents the length of the single temporal dimension. For simplicity, the parameters were chosen as $X = Y = 1000$ and $T = 24$, thus creating a three-dimensional space-time volume (i.e. study area), within which all synthetic point events are placed. The four distinct (i.e. reference) models are as follows:

1. **Homogeneous Poisson process** (Model 1): point events are distributed uniformly through the space-time cube to produce a homogeneous distribution of point events. The number of events generated is dependent only on the dimensions of the space-time cube and the rate of intensity (i.e. number of point events per unit volume) of the Poisson point process λ.
2. **Thomas (clustered) point process** (Model 2): point events are generated in two phases [**Diggle *et al.*, 1995**]:
    1. Generation of cluster centres: cluster centres (i.e. 'parent' events) are generated by a homogeneous Poisson point process with a given rate of intensity (i.e. number of cluster centres per unit volume) κ.
    2. Generation of inter-cluster events: around each centre, a number γ of assumed intercluster point events (i.e. 'children' events), are generated. The deviations in the coordinates of the 'children' events from the corresponding 'parent event' are identically and independently drawn Gaussian random variables with a constant standard deviation σ.

    The final point event dataset is composed of only the 'children' events. Using the Gaussian distribution allows for the 'children' point events to be clustered around their original 'parent' events.
3. **Reference single thunderstorm model** (Model 3): point events are generated according to the lightning strike modelling approach detailed in **Section 3.3.5**. A single model run is used in this reference model using Storm 1A as the storm that is used to introduce temporal structure in the inter-event times and sample distances to the movement track.
4. **Reference three thunderstorm model** (Model 4): point events are generated identical to Model 3, but the model is run three times using Storms 1A, 1B and 1C for the temporal structure of inter-event times and sampling distances to the movement track. The choice is made for the separation of the initiations points for the modelled storms.

All four models have differences in the necessary input parameters and their purpose in the simulation study. Model 1 and Model 2 are point processes, representing idealised conditions for a random and clustered point event dataset. Model 3 and Model 4 then introduce the characterised physical variables in the modelling framework, creating point events datasets that are 'representative' of lightning strikes as a natural hazard. For the Model 3 and Model 4 reference lightning strike, the necessary input parameters were detailed in **Section 3.3.5**. For Model 1 and Model 2, the aforementioned input parameters required for their generation are summarised in **Table 3.5**.

Table 3.5. Variables and their definitions for **Model 1** and **Model 2** dataset generation.

Variable	Definition
X	Length of the first spatial dimension
Y	Length of the second spatial dimension
T	Length of the temporal dimension
x	Spatial coordinate along $[0, X]$
y	Spatial coordinate along $[0, Y]$
t	Temporal coordinate along $[0, T]$
N	Total number of point events
λ	Rate of intensity for Model 1
κ	Rate of intensity for the generation of 'parent' events in Model 2
σ	Standard deviation of a Gaussian distribution for Model 2
γ	Number of 'children' events per 'parent' event for Model 2

For implementing Model 1 and Model 2, the implementation algorithms are detailed in **Appendix F** as **Algorithm 1**. In this thesis, the regular [**Taylor *et al.*, 2001**] point event dataset was initially included to better understand the implementation and interpretation of individual clustering methodologies and certain results in **Section 4** will be referenced to this dataset in the appendices. For implementing Model 3 and Model 4, this simulation study also included additional datasets that are variations in the input parameters to both models.

For Model 3 as the reference single thunderstorm model, there are three additional models that have a varied movement speed v, varied distance to the movement track distribution covariance

and varied both parameters simultaneously. These models are denoted as Model 3.1, Model 3.2, and Model 3.3, respectively, and all have the same number of point events $N = 3{,}000$. Their inter-event time temporal structure is based on Storm 1A and have the same initiation point set at (250 km, 250 km, 6 h) in the aforementioned $X = Y = 1000$ km and $T = 24$ h space-time volume as the study area of this simulation study. All the selected values for the model variables for Model 3 and its variations are listed in **Table 3.6**.

Table 3.6. Variables and their values for the generation of **Model 3** and its variations.

Model variable	**Model 3**	**Model 3.1**	**Model 3.2**	**Model 3.3**
Movement speed v, km h^{-1}	50	150	50	150
Number of points events per thunderstorm N	3000			
Inter-event time distribution	Temporal structure based on events 500–3500 from Storm 1A			
Initiation point (km, km, h)	(250,250,6)			
Movement direction relative to the easting axis	45°			
Covariance matrix (Cov) of Normal distribution around moving sources	From Storm 1A		$10 \cdot Cov$ from Storm 1A	

The resulting four models have the distinct variables listed in **Table 3.6**, but each initiation of an individual model will produce a different dataset due to the sampling procedures for the inter-event time distribution and distance to the movement track. Throughout the simulation study, the same initiation for each model is used for the application of the various spatio-temporal clustering methodologies in **Section 4**. These initiations for the reference single thunderstorm model and its three variants are shown in **Figure 3.24**.

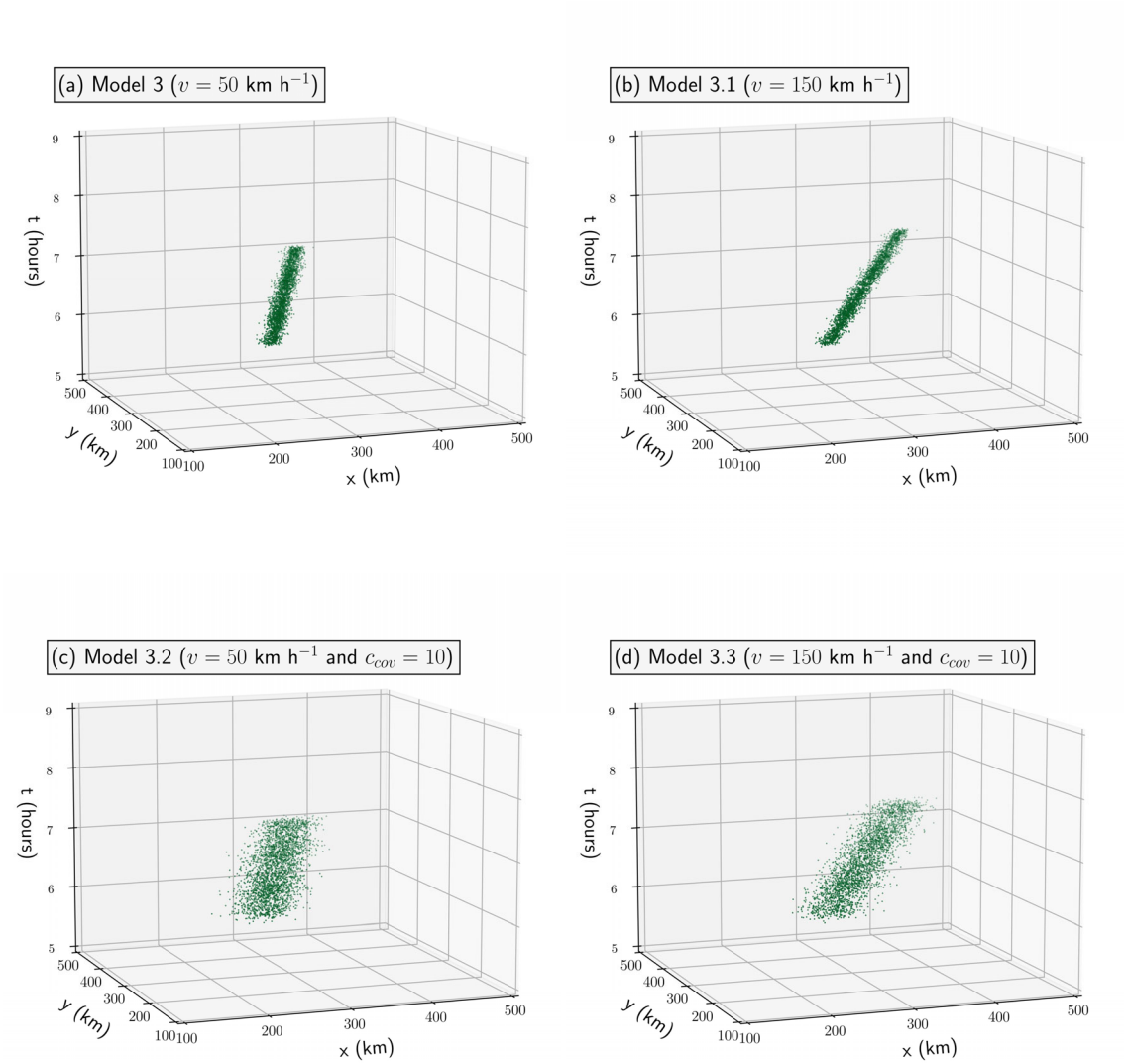

Figure 3.24. The selected initiations of **(a)** the reference single thunderstorm model (**Model 3**), **(b) Model 3.1** with the increased movement speed *v*, **(c) Model 3.2** with the increased spatial spread for the distance to the movement track and **(d) Model 3.3** which has both increased movement speed and spatial spread. All models are displayed on the same space-time volume and each dataset consists of N = 3,000 point events.

For Model 4 as the reference three thunderstorm model, there are also three additional models, each with variations in the model variables. As there are three thunderstorm sources in each model, a given model variable can be varied twice within the same model (i.e. having a smaller movement speed *v* value for one source and larger for the other). These models are denoted in the simulation study as Model 4.1, Model 4.2, and Model 4.3 and all have the same number of total point events N = 10,000 with the distribution of point events between the three thunderstorms in each model set as (3000, 5000, 2000). Denoting each of the three storms within a model as (S1, S2, S3), all four models have the same initiation points set at (250 km,

250 km, 6 h), (250 km, 250 km, 6.5 h) and (250 km, 250 km, 7 h). They all have the movement direction relative to the easting axis set at 45° and for each model, the individual storm inter-event time distributions are based on Storms 1A, 1B and 1C for S1, S2 and S3, respectively. These models are initiated and contained within the aforementioned $X = Y = 1000$ km and $T = 24$ h space-time volume. All the selected values for the model variables that are varied for Model 4 and the respective variations are listed in **Table 3.7**.

Table 3.7. Variables and their values for the generation of **Model 4** and its variations.

Model variable	**Model 4**			**Model 4.1**			**Model 4.2**			**Model 4.3**		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Movement speed v, km h^{-1}	50	55	60	150	55	0	50	55	60	150	55	0
Covariance matrix (*Cov*) of Normal distribution around moving sources from the corresponding storm (S1 –Storm 1A, S2 – Storm 1B, S3 – Storm 1C)	*Cov*	*Cov*	*Cov*	*Cov*	*Cov*	*Cov*	$0.1 \cdot$ *Cov*	*Cov*	$10 \cdot$ *Cov*	$0.1 \cdot$ *Cov*	*Cov*	$10 \cdot$ *Cov*

For Model 3, each initiation of an individual model will produce a different dataset due to the sampling procedures for the inter-event time distribution and distance to the movement track. The initiations used in the simulation study for the reference three thunderstorm model and its three variants are shown in **Figure 3.25**.

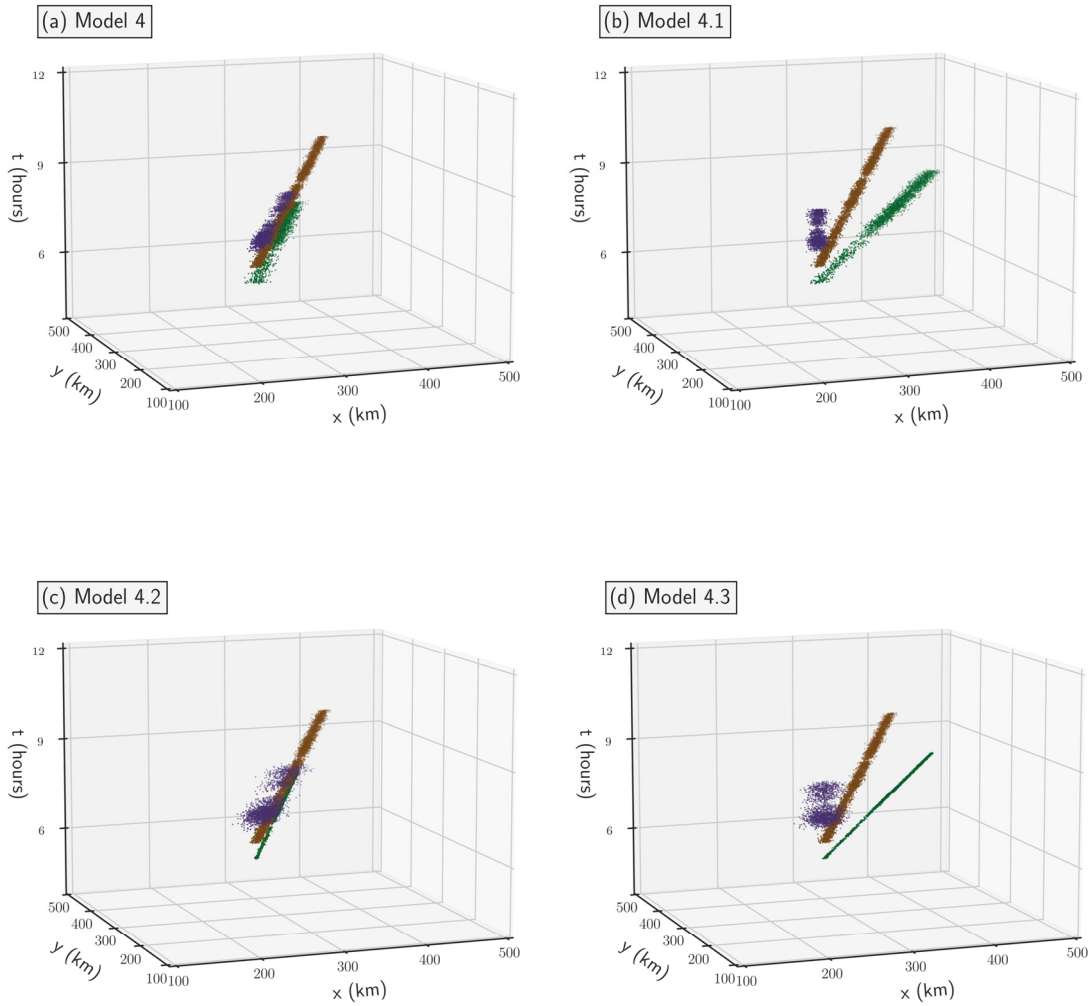

Figure 3.25. The selected initiations of **(a)** the reference three thunderstorm model (**Model 4**), **(b) Model 4.1** with varied movement speed v, **(c) Model 4.2** with varied spatial spread for the distance to the movement track and **(d) Model 4.3** which has variations in both the movement speed and spatial spread. All models are displayed on the same space-time volume and each dataset consists of $N = 10{,}000$ total point events point events distributed as (3000, 5000, 2000) point events between the three storms.

3.4.2. Trajectory benchmarks

To establish dataset benchmarks that would allow assessing spatio-temporal clustering methodologies on idealised trajectory datasets, datasets are restricted to a three-dimensional space-time cube as $\{X = 20, Y = 20, T = 20\}$ (but allowed to propagate outside of it with time). Similarly to using the synthetic point event datasets in Model 1 and Model 2, a distinction between homogeneous and clustered datasets is made.

Two different approaches have been implemented computationally, based on how the initial positions (hereafter designated *IPs*) of the trajectories at time $t = 0$ are initiated. After initiation of the IPs, trajectories are developed as Brownian motion with drift. The process for selecting the standard deviation and the drift parameters for the Brownian motion differentiates between the two datasets. A detailed description of the implementation for both datasets is given in **Appendix F** as **Algorithm 2**. Summaries for the implementation of both datasets are as follows:

- **Homogeneous trajectories**: homogeneous trajectory dataset is generated in two phases:
    1. **Generation of homogeneous IPs**: the IPs are generated by a two-dimensional homogeneous Poisson process, thus distributing them uniformly throughout the two-dimensional space at time $t = 0$.
    2. **Selection of Brownian motion parameters**: intervals for the drift $[\mu_0, \mu_1]$ and standard deviation $[\sigma_0, \sigma_1]$ values are defined. From these intervals, drift μ and standard deviation σ values are uniformly drawn from the intervals for each individual trajectory, which is then developed as Brownian motion with drift.
- **Clustered trajectories**: clustered trajectory dataset is generated in two phases:
    1. **Generation of clustered IPs**: the IPs of trajectories are generated by a two-dimensional Thomas (clustered) point process. At $t = 0$, the IPs are clustered throughout the two-dimensional space with each set of 'children' events assigned a cluster membership number c.
    2. **Selection of Brownian motion parameters**: intervals for the drift $[\mu_0, \mu_1]$ and standard deviation $[\sigma_0, \sigma_1]$ values are defined. From these intervals, drift μ an standard deviation σ values are uniformly drawn from the intervals for each cluster of trajectories (as determined by the assigned cluster number after the

Thomas point process). Therefore, all trajectories that were assigned the same cluster number c for their IPs are generated by the same set of Brownian motion parameters.

The variables relevant for the homogeneous Poisson and the Thomas (clustered) point processes necessary for generating these datasets were already detailed in **Table 3.5**. The additional variables and their definitions for generating the homogeneous and clustered trajectory datasets are listed in **Table 3.8**.

Table 3.8. Variables and their definitions for synthetic trajectory dataset generation.

Variable	Definition
X	Length of the first spatial dimension
Y	Length of the second spatial dimension
T	Length of the temporal dimension
x	Spatial coordinate along $[0, X]$
y	Spatial coordinate along $[0, Y]$
t	Temporal coordinate along $[0, T]$
N	Total number of trajectories
$[\sigma_0, \sigma_1]$	Interval of standard deviation values
$[\mu_0, \mu_1]$	Interval of drift parameter values
c	Cluster membership number
μ	Selected drift parameter value
σ	Selected standard deviation parameter value
n_t	Number of time steps taken between $[0, T]$

For a space-time cube of dimensions $\{X = 20, Y = 20, T = 20\}$ with $n_t = 50$ time steps to be taken for a range of temporal coordinates $t = [0, T = 20]$, the following parameters were selected to generate the two synthetic benchmark datasets:

- **Homogeneous trajectories**: number of IPs per unit area $\lambda = 0.5$, intervals for the drift $[\mu_0, \mu_1] = [-0.1, 0.1]$ and standard deviation $[\sigma_0, \sigma_1] = [0.1, 0.2]$ parameter.
- **Clustered trajectories**: number of trajectory IP cluster centres per unit area $\kappa = 0.01$, number of inter-cluster IPs $\gamma = 50$ with their deviations drawn as identically and

independently distributed Gaussian random variables with standard deviation $\sigma = 0.5$, intervals for the drift $[\mu_0, \mu_1] = [-0.1, 0.1]$ and standard deviation $[\sigma_0, \sigma_1] = [0.1, 0.2]$.

These parameter choices generated $N = 200$ trajectories for each dataset and four IP clusters for the clustered trajectory dataset (50 trajectories in each cluster). Both datasets are illustrated in the space-time cube in **Figure 3.26** with added plots of transformed two-dimensional space {x', y'}, where the IPs of all trajectories are centred on {x' = 0, y' = 0} and the trajectories are collapsed onto the two spatial dimensions. Such a transformation allows to illustrate the propagation of trajectories and how the difference in the approach for assigning the drift parameter μ affects the propagation direction.

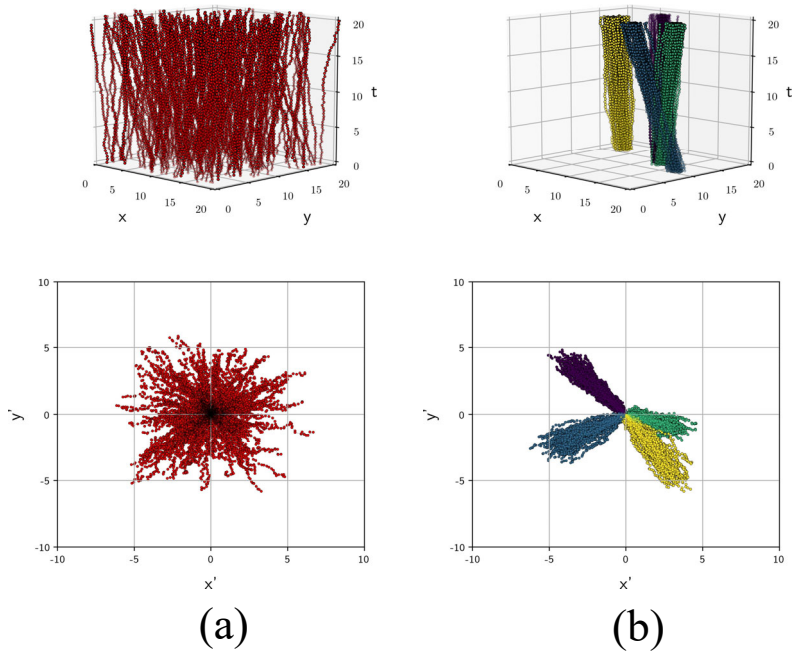

Figure 3.26. The two synthetic datasets used for spatio-temporal clustering analysis of trajectory data. The initial positions of the trajectories at $t = 0$ are chosen based on either 2D **(a)** homogeneous Poisson (homogeneous trajectory dataset) or **(b)** Thomas (clustered trajectory dataset) point process and trajectories are generated as Brownian motion with drift. Both datasets contain $N = 200$ trajectories with input parameters set as $\lambda = 0.5$, $\kappa = 0.01$, $\gamma = 50$, $\sigma = 0.5$, $n_t = 50$, $[\sigma_0, \sigma_1] = [0.1, 0.2]$ and $[\mu_0, \mu_1] = [0.1, 0.1]$. Upper subfigures show the propagation of trajectories in the space-time cube, while lower subfigures show the propagation directions once trajectories are collapsed onto the two spatial dimensions and centred on $x' = 0$, $y' = 0$. Each cluster in **(b)** is indicated by a different colour for all the trajectories within it.

3.5. Chapter conclusions

In this research chapter, the real-world datasets and the simulation study (Model 1 and Model 2 as synthetic benchmarks and Model 3 and Model 4 as synthetic models representative of single and three thunderstorm systems) were presented in **Section 3.2** and **Section 3.4,** respectively.

Three physical variables were used as input parameters to produce these representative lightning strike models: inter-event time, movement speed and distance to the movement track. These models are an extension of work in [**Finke, 1999**] by explicitly including lightning strike and thunderstorm physical variables into the model framework. The approach to designing the spatio-temporal model allows for complete control over the variables and generates runs that can represent both lightning strikes and non-physical conditions. These variations in the model parameters provide a clear differentiator between different sets of model input parameters as they are related to the physical variables.

The values and ranges of these variables were based on the characterisation results in **Section 3.3**. Together with the synthetic benchmarks, the simulation study allows to evaluate spatio-temporal clustering methodologies in **Chapter 4**. The characterisation frameworks returned physical variable values consistent with value ranges found in other studies that used auxiliary datasets and more complex characterisation methodologies.

In other studies working with lightning as point events recorded by long-range lightning location systems similar to the ATDnet, the characterisation framework focused on statistical properties of the lightning strike dataset (e.g. flash density [**Enno *et al.*, 2020**; **Anderson and Klugmann, 2014**]) over various spatial and/or temporal scales). From the three physical variables characterised in this thesis, the inter-event time distribution has been studied extensively over various regions and storm characteristics [**Yair *et al.*, 2014**; **Ballarotti *et al.*, 2012**; **Stall *et al.*, 2009**; **Saba *et al.*, 2006**]. The characterisation research in **Section 3.3.4** of this thesis extends the existing literature on how to characterise movement speed and distance to the movement track as physical variables based on point event lightning strike data. Both physical variables have only been characterised based on auxiliary data or different representations of lightning.

Two further research directions can be envisioned for this chapter. First, this chapter's most significant source of uncertainty stems from the assignment of lightning strikes to individual thunderstorms. Further research can be done to perform this assignment procedure in a more complex and detailed manner by incorporating not just synoptic analyses and auxiliary radar relative positions, but other types of meteorological observations and hazard datasets. Second, further meteorological perspective is necessary to interpret the characterisation results for the physical variables. This is particularly true for the second case study of Storms 2A–2C where the synoptic analysis authors noted the structural differences between the individual storms.

4. Application of the spatio-temporal clustering methodologies to real-world and synthetic natural hazard datasets

Chapter summary

This research chapter presents the spatio-temporal clustering results for the real-world and synthetic datasets selected for the simulation study. For point events, the spatio-temporal clustering methodology framework (**Section 2.5**) of combining global and local clustering measures was applied to the four spatio-temporal point event models in the simulation study and their variations (**Section 3.4.1**) and the real-world lightning strike dataset (**Section 3.2.1**). The methodology framework is presented through individual methodology results and the potential connections and use of methodologies to improve the performance and/or clustering accuracy of each other are discussed. The variations in the single storm (reference Model 3 with $N = 3{,}000$ events) and three storm lightning strike (reference Model 4 with $N = 10{,}000$ events total, distributed as {3000, 5000, 2000} between the storms) allow for the examination of the sensitivity of individual clustering methodologies to the various changes in the input parameters (including the lightning strike and thunderstorm physical variables characterised in **Section 3.3.4**) for the models. The methodology framework is evaluated for its limitations for application to point events produced by a moving source. For trajectories, the density-based clustering methodology Trajectory-OPTICS is applied to the synthetic homogeneous and clustered datasets (**Section 3.4.2**) and the selected real-world (**Section 3.2.2**) trajectory data. Initial clustering results are analysed and evaluated for both data types with their limitations presented and discussed. This chapter aims to establish the methodology framework results and to give context and reasoning for the proposed improvements to aspects of the methodology framework as presented and analysed in **Chapter 5**.

This chapter aims to answer the following questions:

Q4.1. If presented with a point event dataset produced by a moving source, what framework (i.e. combination and sequence of spatio-temporal clustering methodologies) would provide the most computationally efficient and complete clustering analysis?

Q4.2. How do the spatio-temporal clustering methodologies, whether global or local measures, compare between themselves in terms of their clustering accuracy and general purpose of application to point events produced by a moving source?

Q4.3. How sensitive are individual methodologies to changes in input parameters (in particular, the characterised physical variables) of the lightning strike spatio-temporal point event model?

4.1. Introduction

In this chapter, the spatio-temporal clustering methodologies presented in **Chapter 2** are applied to the real-world datasets from **Section 3.2** and the synthetic benchmark datasets as part of the point event simulation study from **Section 3.4**. Each data type is considered separately in this section. **Section 4.2** presents the results of the simulation study for the point event datasets. The next three sections of this chapter present the clustering results and analyses for trajectory benchmark datasets (**Section 4.3**), real-world point event (lightning strike) dataset (**Section 4.4**) and real-world trajectories dataset (**Section 4.5**). The results are presented visually, analysed and the limitations of the methodologies are discussed. This justifies further improvements on the existing methodologies within this thesis and identifies further research directions in the field.

For the point event data types presented and analysed in this thesis, **Sections 4.2** and **4.4** present the global and local clustering measures and their application framework, discussed in **Section 2.5**. The ability to identify useful properties of the datasets of the clustering methodologies, whether individually or complementary, is examined. For trajectories, the analysis and discussion include only the Trajectory-OPTICS methodology as presented in **Section 2.3.2.1**. The evaluation of the strengths and weaknesses of this methodology is postponed to **Section 6.3.3** with the discussion of potential research directions in this field.

Similarly, evaluating the clustering results for the point event data type leads us to identify possible improvements to the methodologies individually or within the overall framework of the spatio-temporal cluster methodology application and analysis. The overall framework for the analysis is evaluated in **Section 4.2.4** for the clustering results when the framework is applied to the point event simulation study datasets (**Section 3.4.1**) in **Section 4.2**. Additional evaluation and further discussion follows in **Section 4.4.4** after the framework is applied to the real-world lightning dataset in **Section 4.4**. Following the evaluation of the results, for the point event data type, the proposed improvements to individual methodologies is shown and discussed in **Chapter 5**.

4.2. Results and analysis for point event simulation study

For the point event data type, three different methodologies were identified and presented in **Section 2.2.1** and form a common framework for the spatio-temporal analysis (**Section 2.5**): K-functions, kernel density estimation and the space-time scan statistic. In **Section 3.4.1**, four spatio-temporal point event models were presented to simulate spatio-temporal point event data, with the aim to assess the performance of the analysis framework on these synthetic examples. Two of these models were based on characterisation results for lightning as a natural hazard as presented in **Section 3.3**. The methodology framework can then be evaluated based on not only purely mathematical models, but also models representative of a natural hazard process.

The results of applying all three methodologies within a common framework to the datasets in the simulation study are presented in this section. The necessary input parameters and their selected values are given for each methodology. The clustering results for each dataset are presented and analysed separately within the context of the individual methodology. The K-function results as the global clustering measure are displayed in **Section 4.2.1**, the kernel density estimation results to identify particular space-time volumes of interest are shown in **Section 4.2.2** and the space-time scan statistic results to identify particular clusters are shown in **Section 4.2.3**. A discussion and evaluation of the clustering results of methodology application framework follows in **Section 4.2.4**.

4.2.1. K-function analysis

As discussed in **Section 2.4**, the global clustering measures as a methodology group allow to gain insight into the spatial and temporal ranges over which point events exhibit clustering behaviour [**Wiegand and Moloney, 2004**; **Diggle *et al.*, 1995**]. While not being able to identify individual clusters, a global clustering measure is a starting point for a point event clustering analysis framework (**Section 2.5**).

Using K-functions (**Section 2.3.1.1**) as a global clustering measure allows for the quantification of the spatial (s) and temporal (τ) ranges of the presence of space-time clustering. It can provide estimates of the input parameters for the other point event spatio-temporal clustering methodologies in this application framework (i.e. bandwidth selection for kernel density

estimation [**Tonini *et al.*, 2017**; **Nakaya and Yano, 2010**] and dimensions of the cylindrical scanning window for the space-time scan statistic).

Implementation details of the K-functions and the global clustering measure D are expanded in **Appendix B**. The global clustering measure D is the difference between the spatio-temporal K-function $K(s,\tau)$ and the product of the individual spatial $K(s)$ and temporal $K(\tau)$ K-functions. If this difference is positive, there is an interaction between the spatial and temporal dimensions of the process producing the point events. For no interaction, this difference is zero. To implement this methodology, spatial and temporal weights need to be assigned for spatial and temporal ranges outside the defined space-time volume. The weights are an important correction for edge effects as the expected number of events within given spatial and temporal ranges will be underestimated due to some of the point events being close to the boundaries.

The spatial weight value is assigned as the reciprocal of the proportion of the circumference of the circle formed around a given point event that lies within the space-time cube. The radius of the circle is the spatial range value. For temporal weight, a unit value is assigned if a line of length 2τ with its centre on a given point is within the space-time cube. Otherwise, the weight is doubled. Such simplification is justified [**Diggle *et al.*, 1995**; **Ripley, 1982**] as spatial edge effects being more pronounced influence the methodology results. Continuous temporal weights were one of the minor modifications evaluated during this thesis and left as a potential direction for further research (**Section 6.5.2**).

To gain initial intuition for the K-function methodology, it was applied to regular grid data. The dataset, its dimensions, bin counts and the K-function analysis results with evaluation, are shown in **Appendix O**. As the methodology was implemented manually, this ensured that the implementation worked as intended. For all the datasets in the simulation study (**Section 3.4**), the global clustering measure D is calculated. For Model 1 and Model 2, the synthetic datasets of $N = 1{,}000$ point events are placed in a $\{X = 20, Y = 20, T = 20\}$ space-time volume according to the point event generation algorithms. The spatial and temporal ranges are set as $s = \{0.1, 0.2, ..., 5\}$ and $\tau = \{0.1, 0.2, ..., 5\}$. For Model 1 (i.e. the homogeneous point event dataset), the results for the clustering measure D and its K-function components are given in **Figure 4.1**.

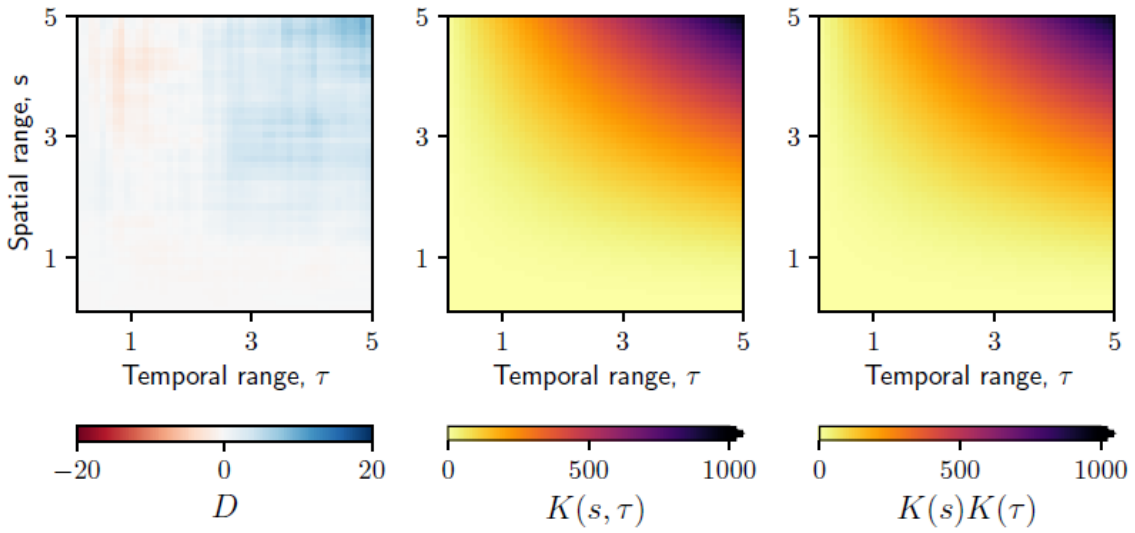

Figure 4.1. Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$, spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$ as applied to homogeneous point event data (**Model 1**). Their values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales.

Without the well-defined structure of the regular grid point event dataset (shown in **Appendix Figure O1**), the results of the clustering measure D and its components do not exhibit a clear structure for Model 1. The clustering measure D has values close to zero throughout spatial and temporal ranges with minor local variations that do not exhibit patterns. These clustering results depend on the particular initiation of the homogeneous Poisson process.

To establish how the clustering measure D values are distributed for the clustered point event dataset, the same spatial and temporal range values were applied to Model 2 on the same space-time volume and spatial and temporal range values. The same number of $N = 1{,}000$ point events were generated by setting $\kappa = 0.0025$, $\gamma = 50$ and $\sigma = 0.5$, according to **Table 3.5**. The results of all three measures for Model 2 are shown in **Figure 4.2**.

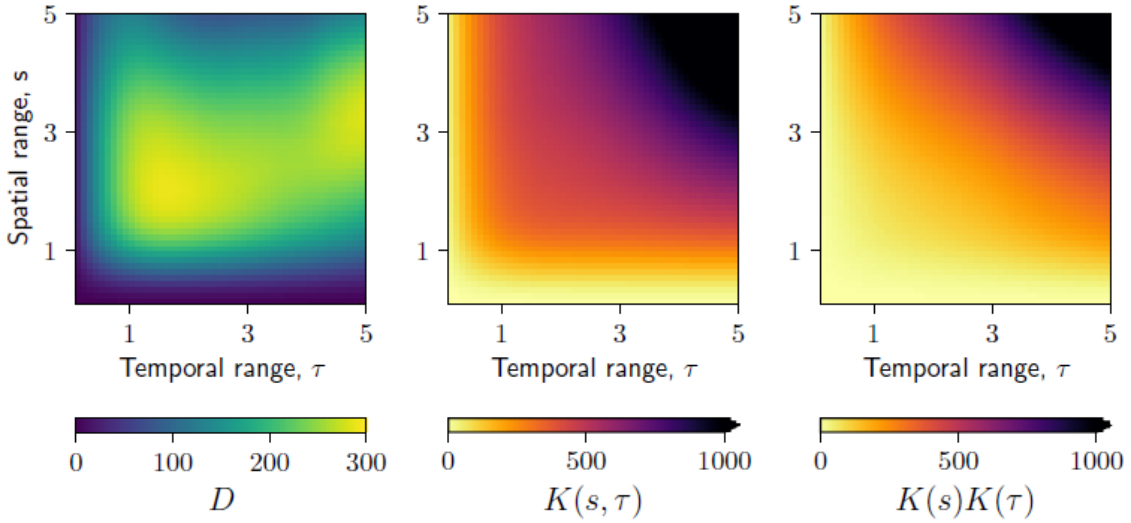

Figure 4.2. Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$, spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$ as applied to clustered point event data (**Model 2**). Their values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the different colour scales for the clustering measure D from other synthetic datasets.

In **Figure 4.2**, the clustering measure D for Model 2 exhibits contrasting behaviour compared to the homogeneous point event dataset generated by Model 1 in **Figure 4.1**. The distribution of the clustering measure D values has a clear structure, increasing almost two orders of magnitude compared to the homogeneous point event dataset. They reach their peak at ($s = 2$, $\tau = 1.6$), indicative of the procedure for generating the clustered point events (**Section 3.4.1**) as inter-cluster point events were independently drawn Gaussian random variables around each cluster centre with zero mean and a constant standard deviation set as $\sigma = 0.5$.

Spatial and temporal ranges for the peak D value confirm the extent to which the events would be expected to be separated from each other. The clustering measure decreases afterwards but increases again for larger spatial and temporal range values. As the number of 'parent' events or cluster centres were set to $\gamma = 50$ and the 'children' events are drawn from a Gaussian distribution with standard deviation set at $\sigma = 0.5$ for all dimensions around the 'parent' event. Any increases in the clustering measure D value after the peak is reached at ($s = 2$, $\tau = 1.6$) are likely due to some individual events encountering nearby clusters. It is noted that depending on the given initiation of Model 2, there might be variations in the exact locations (in terms of spatial and temporal range values) of the global and local peaks in clustering measure D values, but the overall behaviour would remain the same relative to any initiation of Model 1.

For Model 3 and its variations as listed in **Table 3.6**, the K-function analysis results are shown in **Figure 4.3**. It is noted that for Model 3 and Model 4 the space-time volume was expanded to {X = 1000 km, Y = 1000 km, T = 24 h} and the total number of point events within each model are N = 3,000 and N = 10,000, respectively. The spatial and temporal ranges are set as s = {2, 4, ..., 120} km and τ = {0.05, 0.1, ..., 2} h.

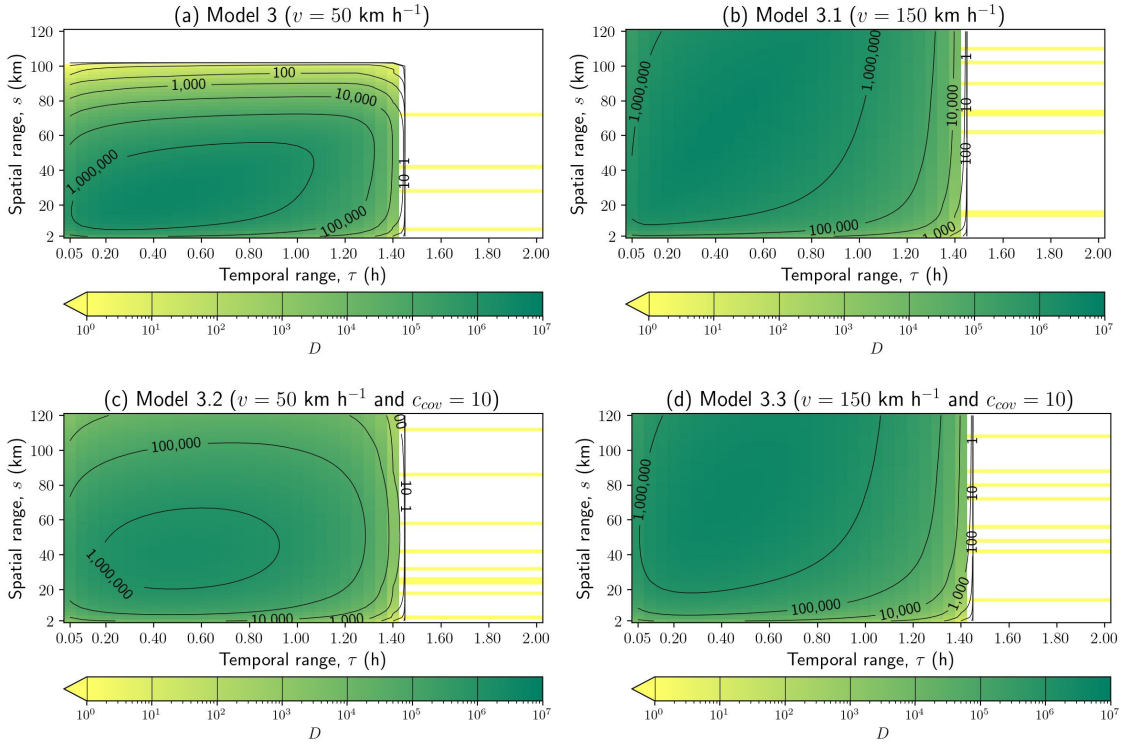

Figure 4.3. The clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to the **(a)** single thunderstorm reference model (**Model 3**) and its variations: **(b) Model 3.1**, **(c) Model 3.2** and **(d) Model 3.3**. The particular differences in the model variations are indicated in the titles of the subfigures. The D values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the different colour scales for the clustering measure D from previous figures. The colorscale is a log-scale and continues as 'yellow' for any D values below 1, and is 'white' for negative or non-existent values. The various D value thresholds are indicated by black contours.

In **Figure 4.3**, the clustering measure D values for Model 3 and its variations show the expected behaviour. The reference Model 3 (**Figure 4.3a**) shows similar behaviour to Model 2 (**Figure 4.2**), which is expected from a dataset, which, through its generation framework, is a clustered point event dataset. In **Figure 4.3a**, there is an initial increase in D values, reaching maximum values around (s = 40 km, τ = 0.55 h) and decreasing afterwards. Only individual synthetic lightning strikes contributing to the clustering results after (s = 100 km, τ = 1.4 h). The range of the D values as seen by the black contours extends over seven orders of magnitude. From

these results, it can be assessed that the spatio-temporal extent of the synthetic storm is close to 100 km and 1.4 h, which is reasonable given the input parameter for Model 3 (**Table 3.6**).

The variations to the reference Model 3 show contrasting behaviours relative to the reference model as expected. Increasing the movement speed of the lightning strike cluster from v = 50 km h^{-1} to v = 150 km h^{-1} for Model 3.1 (**Figure 4.3b**) extends the spatial range with relatively significant D values. Previously the D value contour D = 1,000,000 was observed up to s = 55 km (**Figure 4.3a**), but the model variation for Model 3.1 has extended this value contour past the maximum spatial range of s = 120 km. The behaviour of higher temporal ranges values remains unchanged as the inter-event times between the synthetic lightning strikes were not altered. The increased movement speed only governs the distance that the lightning strikes as a cluster (i.e. the thunderstorm producing them) cover given the total duration of the model.

When leaving the movement speed as the reference Model 3, but changing the covariance matrix factor for the distance to the movement track variable, Model 3.2 (**Figure 4.3c**) also exhibits clearly varied behaviour. Similarly to Model 3.1 (**Figure 4.3b**), the spatial range values shown in **Figure 4.3c** with relatively significant D values is extended, while simultaneously lowering the D values in the previously observed (**Figure 4.3b**) spatial and temporal ranges (i.e. within s = 100 km, τ = 1.4 h). The region enclosed by the highest noted contour D = 1,000,000 is now smaller, starting at (s = 35 km, τ = 0.2 h) compared to (s = 10 km, τ = 0.05 h) for Model 3.1 (**Figure 4.3b**). While the unchanged movement speed implies no change in spatial separation between events in the spatial distance component 'in the movement direction', the increased distance to the track (i.e. perpendicular to the movement direction) lowers D values for the spatial and temporal range region up to (s = 60 km, τ = 1.0 h).

The combined increase in the movement speed and the distance to the movement track in Model 3.3 (**Figure 4.3d**) exhibits behaviour expected following the analysis of the previous two models (Model 3.1 and 3.2). In **Figure 4.3d**, The D value contour D = 1,000,000 extends past the maximum spatial range of s = 120 km, while the region enclosed by the contour is slightly smaller (the contour starts at (s = 20 km, τ = 0.1 h) compared to (s = 10 km, τ = 0.05 h) for Model 3.1.

For Model 4 and its variations as listed in **Table 3.7**, the K-function analysis results are shown in **Figure 4.4**. The spatial and temporal ranges consist of two different sets. With the same

spatial and temporal spacing as for Model 3 and its variations as shown in **Figure 4.3**, the results in **Figure 4.4** are calculated for ranges s = {2, 4, ..., 100} km and τ = {0.05, 0.1, ..., 1.75} h. Two extensions are added to this region with larger spacing values. Spatial range is extended to s = {110, 115, ..., 300} km for τ = {0.1, 0.2, ..., 1.0} h. Temporal range is extended to τ = {1.85, 1.95, ..., 2.25} h for s = {5, 10, ..., 110} km. The remaining region of s = {110, 300} km and τ = {1.0, 2.25} h is left without results.

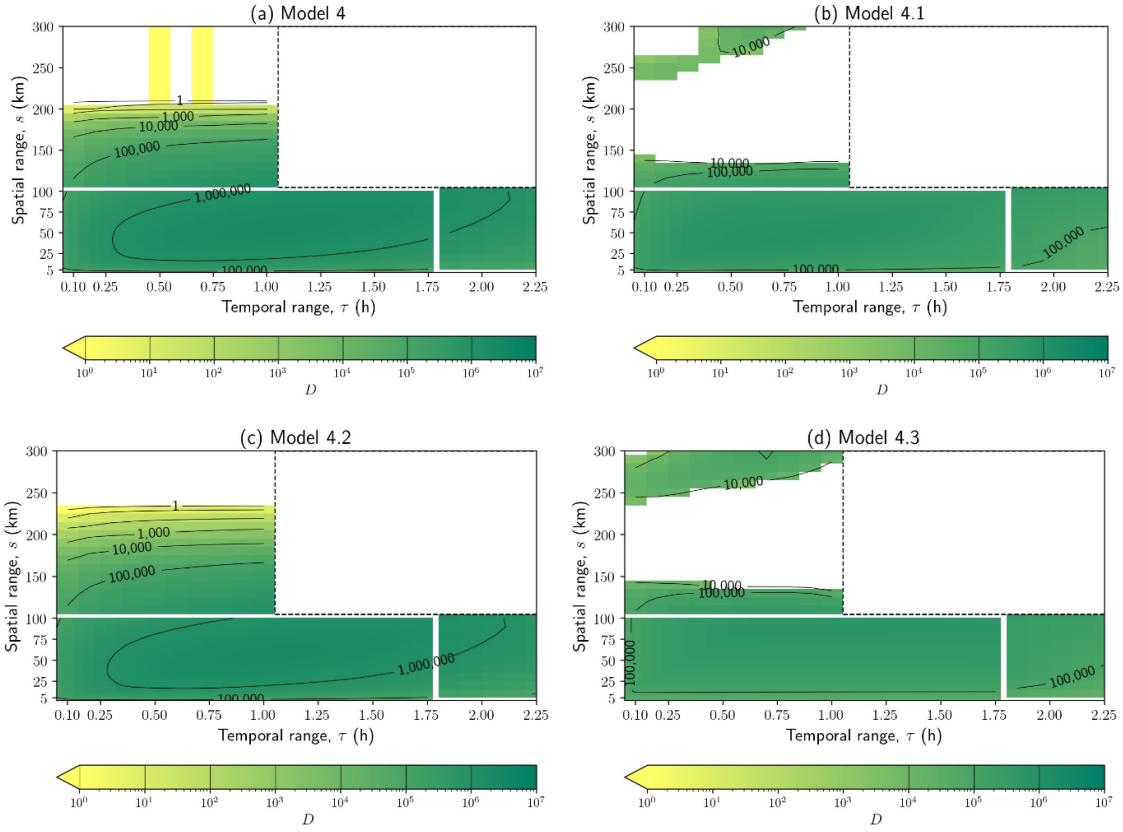

Figure 4.4. The clustering measure $D = K(s,\tau)–K(s)K(\tau)$ as applied to the **(a)** three thunderstorm reference model (**Model 4**) and its variations: **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. The D values are calculated for various combinations of spatial range s and temporal range τ values. The colour scale is a log-scale and continues as 'yellow' for any positive D values below 1, and is 'white' for negative. The various D value thresholds are indicated by black contours on the colour bar. The region of s = (105, 300) km and τ = (1.05, 2.25) h is left without results (marked by dashed lines).

The clustering measure D values for Model 4 and its variations in **Figure 4.4** show the expected behaviour. For the reference Model 4 (**Figure 4.4a**) and the model variation with changing covariance structure for the distance to the movement track value sampling (Model 4.2 in **Figure 4.4c**) both show similar behaviour to Model 3 (**Figure 4.3a**) for smaller ranges values. As the movement speeds of the three thunderstorm model runs are similar (v = {50,55,60} km h^{-1} from **Table 3.7**), the variations in the covariance matrix multiplication factor only change the extent of the space-time interaction in the spatial domain. In absolute values, the clustering

measure *D*, Models 4 and 4.2 (**Figure 4.4**) also reach maximum values within the same order of magnitude as Models 3 and 3.2 (**Figure 4.3**).

For the Model 4 variations with significantly varying movement speeds (Model 4.1 and 4.3) between the three thunderstorm model runs, this behaviour is unique compared to other Model 4 variations. When extended into the spatial range there is a range between $s = \{145, 240\}$ km for $\tau = 0.1$ h where clustering measure *D* reaches significant negative values (which are not displayed by the log-scaled colourbar). In **Figures 3.25b** and **3.25d**, the three-dimensional view of the three storms indicates the reason for this unique behaviour. In addition to the three-dimensional view of the Model 4 and its variations, the 2D spatial view and the 1D temporal point event counts per 0.1 h (as the temporal range value separation) are presented in **Figure 4.5** for Models 4 and 4.1 (for all other model variations, the results are shown in **Appendix P**).

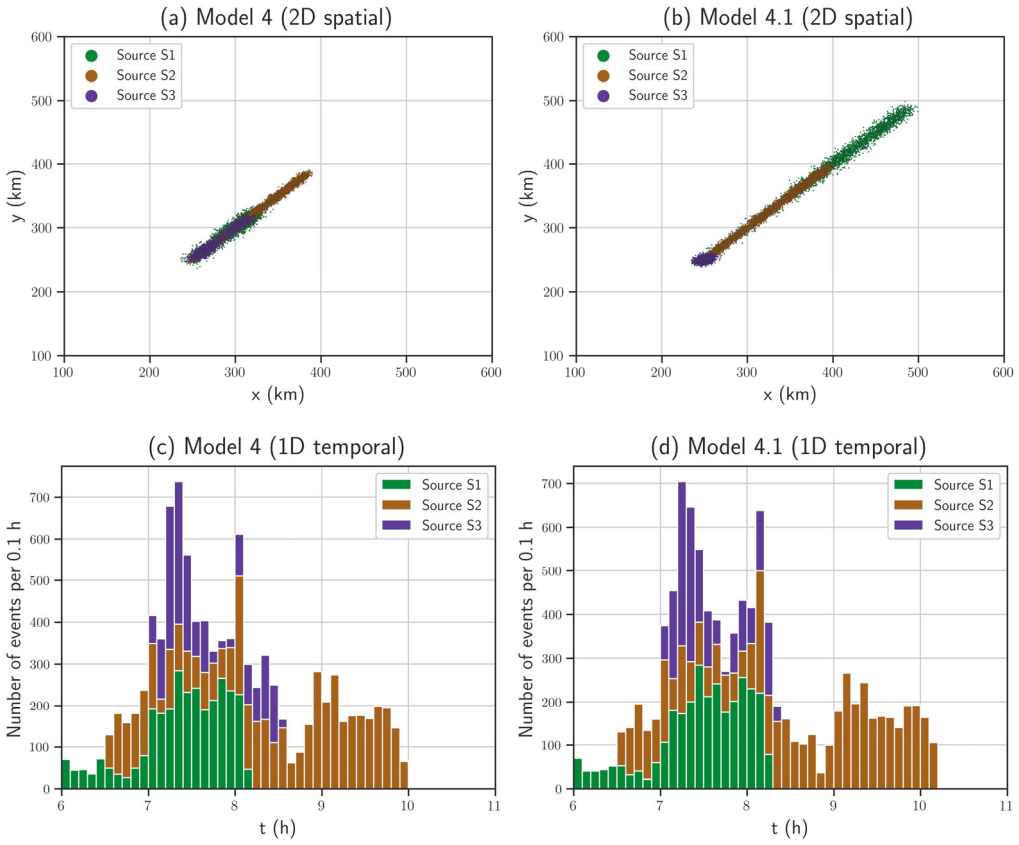

Figure 4.5. The collapsed datasets of the [**(a)**, **(c)**] reference model for three thunderstorms (**Model 4**) and its variation [**(b)**, **(d)**] **Model 4.1**. In [**(a)**, **(b)**] datasets are collapsed onto the 2D spatial domain as scatter points, while in [**(c)**, **(d)**] datasets are collapsed onto the temporal dimension with the number of events (stacked from all three sources) indicated per 0.1 h as the temporal range step in **Figure 4.4**. The colour scheme for the three storm sources (S1, S2, S3 as per **Table 3.7**) is the same as **Figure 3.25**.

Keeping the same notation for each individual storm as in **Table 3.7**, storm sources {S1,S2,S3} have their initiation points at the same spatial locations, but separated temporally by 0.5 h. The movement speed of S1 is v = 150 km h^{-1}, of S2 v = 55 km h^{-1} and of S3 v = 0 in Model 4.1, compared to v = {50,55,60} km h^{-1} for Model 4. Outside of space-time interaction within each individual storm, small temporal separation values would return clustering D values indicative of space-time interaction when there is an overlap between multiple storms as the spatial range is extended. When comparing the 1D temporal view of both models (**Figures 4.5c** and **4.5d**), there are only minor differences due to model initiations. Changing the movement speed, without changing the inter-event times, does not alter the duration of individual sources. The temporal K-function $K(\tau)$ would return similar values for varying movement speed v values.

The spatial K-function $K(s)$ is directly affected by varying movement speed v values as seen in **Figures 4.5a** and **4.5b**. S1 with v = 150 km h^{-1} (**Figure 4.5b**) covers a significantly larger distance for its duration compared to **Figure 4.5a** when S1 had v = 50 km h^{-1}. Now, for the same temporal range τ values, these point event pairs are further away. Point event pairs that are close in temporal range, will not be close in spatial range, resulting in significant $K(s)K(\tau)$ values (from contributions of other pairs), but zero contribution to the spatio-temporal K-function $K(s,\tau)$. This results in negative clustering measure D values. For spatial range values at s = {240, 245} for τ = 0.1, there is only overlap between point events of S1 and S3, resulting in the positive clustering measure D values (**Figure 4.4b**). The increased covariance matrix factor in **Figure 4.4d** then increases the extent to which these positive D values are observed.

As per the spatio-temporal clustering methodology application framework (**Section 2.5**), the spatial and temporal range values with maximum clustering measure D values $\{s_{MAX}, \tau_{MAX}\}$ as returned by the K-function analysis framework for all Model 3 and Model 4 variations are shown in **Table 4.1**. These values are used for selecting a set of potential kernel bandwidth values for the kernel density estimation methodology.

Table 4.1. Identified spatial and temporal ranges $\{s_{MAX}, \tau_{MAX}\}$ for maximum clustering measure D values for **Models 3, 4** and their variations (**Models 3.1, 3.2, 3.3, 4.1, 4.2, 4.3)**.

Model	$\{s_{MAX}, \tau_{MAX}\}$ (km, h)
3	{24, 0.45}
3.1	{60, 0.40}
3.2	{40, 0.50}
3.3	{74, 0.45}
4	{58, 1.05}
4.1	{44, 0.75}
4.2	{54, 1.00}
4.3	{100, 0.7}

4.2.2. Kernel density estimation

The kernel density estimation (**Section 2.3.1.2**) is a part of the spatio-temporal clustering methodology framework (**Section 2.5**) as it allows to identify space-time volumes that are of interest for identification of individual clusters by any local clustering measure. Global clustering measures (i.e. K-function analysis) provide the necessary estimation of the spatial and temporal scales over which spatio-temporal clustering is present.

To implement the kernel density estimation methodology, an optimal set of the bandwidth values for all three dimensions (two spatial and one temporal) has to be selected. All the sets that will be used for the leave-one-out density estimation (**Appendix C**) are generated based on the spatial and temporal ranges of clustering as found by the clustering measure D when using K-functions (**Section 4.2.1** and **Table 4.1**). All permutations of the bandwidth values are used as input for the leave-one-out density estimator.

While the implementation details are well-defined for the kernel density estimation, the exact visualisation and interpretation of spatio-temporal clustering results is not well-established in the scientific literature as there exist three distinct approaches for this problem [**Gao, 2015**; **Demšar and Virrantaus, 2010**; **Brunsdon *et al.*, 2007**]:

1. **Volume slicing**: the probability density volumes are displayed in temporal slices, showing cross-section of the space-time volume.

2. **Isosurface approach**: surfaces that share the same probability density value are traced through the space-time volume.
3. **Volume rendering**: assigns colour and transparency to voxels (space-time volume) according to the probability density values. By changing these voxel properties (i.e. they can emit, absorb or transmit light) and using different projections of the volume with different points of view, the high point event density volumes can be visually identified.

Of these three visualisation approaches, volume rendering would allow us to identify smaller space-time volumes that would have higher probability density values assigned to the voxels within them. For the synthetic point event datasets, the volume rendering was done first without any lower threshold on the displayed probability density values.

From the K-functions results for Model 1 (**Figure 4.1**) and Model 2 (**Figure 4.2**), for each of the three dimensions, the selected bandwidth values are {0.1, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5}. The optimal bandwidth values identified for the clustered point event dataset were {$h_x = 0.5$, $h_y = 0.5$, $h_t = 0.5$}, while for the homogeneous point event dataset, the leave-one-out density estimator returned {$h_x = 1.25$, $h_y = 1.5$, $h_t = 1.5$}. To calculate the probability density volume, points were placed in the same {$X = 20$, $Y = 20$, $T = 20$} space-time volume with a separation of 0.2 in all directions, representing centroids of {0.2, 0.2, 0.2} voxels. The volume rendering results for both datasets are shown in **Figure 4.6**.

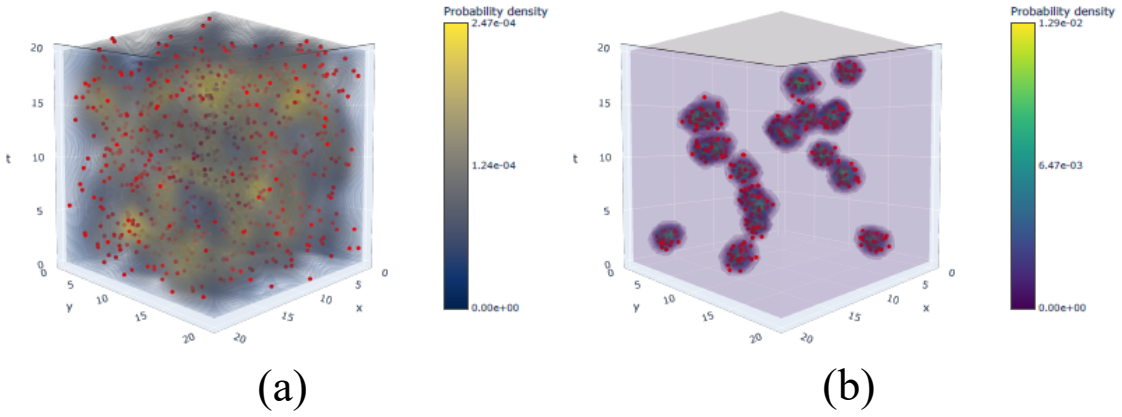

Figure 4.6. Volume rendering of probability density values as returned by the kernel density estimation methodology for **(a)** homogeneous (**Model 1**) and **(b)** clustered (**Model 2**) point event datasets. For each dataset, red points indicate the corresponding point events. The probability density value is given by the colour scale. Note the variations in the colour scale and its limits between both subplots.

The expected visual contrast between the volume renders of the probability density values for the homogeneous (Model 1) and the clustered (Model 2) point event datasets is confirmed. The clustered point event dataset (**Figure 4.6b**) has smaller volumes within the space-time cube of increased probability density. The density continues to increase closer to the cluster centres as generated by the Thomas (clustered) point process (**Section 3.4.1**). For the homogeneous point event dataset (**Figure 4.6a**), the maximum probability density values are relatively lower (as indicated by the upper limits of the colour scales) and do not form any discernible volumes of increased probability density.

To further illustrate the differences between the two datasets and how the kernel density estimation can be used to identify space-time volumes of relatively higher probability density, a lower threshold on the number of voxels displayed is introduced [**Hu *et al.*, 2018**; **Nakaya and Yano, 2010**; **Brunsdon *et al.*, 2007**]. The volume rendering is performed only on voxels above the 90th percentile of probability density values. Each voxel has a probability density value, so they can be arranged and only the ones with the highest probability density values remain. The same number of voxels remains for both datasets as they were placed with the same spatial and temporal separations. The resultant volume rendering of voxels with the highest 10% of probability density values is shown in **Figure 4.7**.

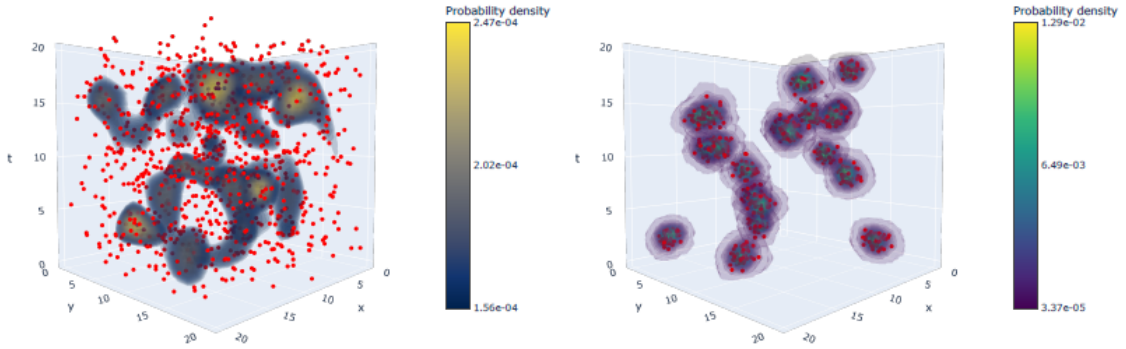

Figure 4.7. Volume rendering of voxels with the highest 10% of the probability density values as returned by the kernel density estimation methodology for **(a)** homogeneous (**Model 1**) and **(b)** clustered (**Model 2**) point event datasets. For each dataset, red points indicate the corresponding point events. The probability density value is given by the colour scale. Note the variations in the colour scale and its limits between both subplots.

When introducing a lower limit of the voxels used for volume rendering by selecting only those containing the highest 10% of probability density values, the contrast between the volumes of higher probability density is similar to what was seen in **Figure 4.6**. The highest 10% of the

probability density values for the clustered point event dataset (**Figure 4.7b**) form the expected quasi-spherical volumes over the space-time cube as generated from the Thomas point process with constant standard deviation parameter σ in all three dimensions (**Table 3.5**). For the homogeneous point event dataset (**Figure 4.7a**), the volumes of higher probability density do not form any coherent clusters.

For Model 3 and its variations, as listed in **Table 3.6**, the kernel density estimation methodology can be applied through the same approach, but the voxel dimensions are now changed due to the $\{X = 1000$ km, $Y = 1000$ km, $T = 24$ h$\}$ space-time volume. Using the K-function analysis results (**Figure 4.3** and **Table 4.1**), a reasonable set of bandwidth values was selected as {2, 5, 10, 20, 30, 40} km for the spatial kernels and {0.05, 0.1, 0.2, 0.3, 0.4} h for the temporal kernel. The leave-one-out density estimation for the optimal bandwidth values returned the set $\{h_x = 2$ km, $h_y = 2$ km, $h_t = 0.05$ h$\}$ for Model 3 and Model 3.1. For Model 3.2 and Model 3.3, the identified optimal bandwidths increased to $\{h_x = 10$ km, $h_y = 10$ km, $h_t = 0.1$ h$\}$. For calculating the probability density values, the voxel dimensions were selected as {5 km, 5 km, 0.05 h} covering the spatial intervals from 150–450 km and temporal intervals from 6–9 h. The choice of these values is based on the coverage of the full simulation study volume and the computational complexity for calculating and visualising the results. The volume rendering results for the reference Model 3 and its variations are shown in **Figure 4.8**.

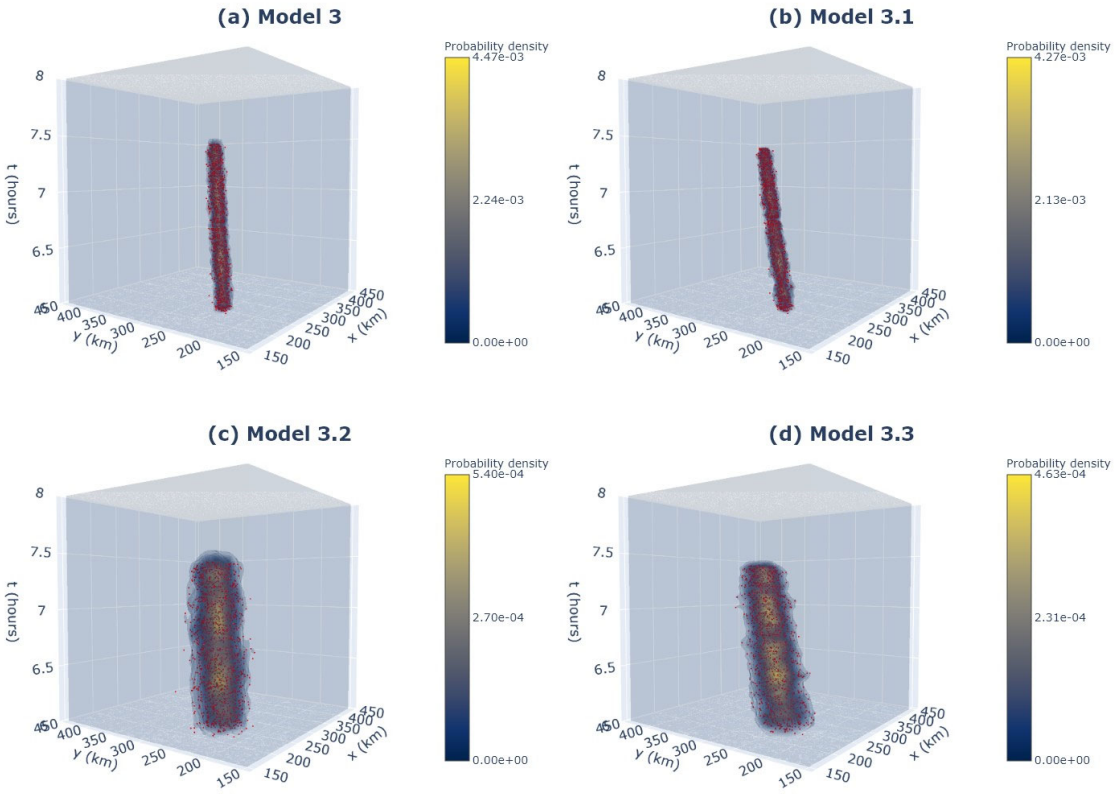

Figure 4.8. Volume rendering of probability density values as returned by the kernel density estimation methodology for the **(a)** single thunderstorm reference model (**Model 3**) and its variations: **(b) Model 3.1**, **(c) Model 3.2** and **(d) Model 3.3**. For each dataset, red points indicate the corresponding point events. The probability density value is given by the colour scale. Note the variations in the colour scale upper limits between all subplots.

Initial results when looking at all the probability density values in **Figure 4.8**, show the expected effect of increased distance from the movement track in Model 3.2 (**Figure 4.8c**) and Model 3.3 (**Figure 4.8d**). The optimal bandwidth values were higher for the more spatially disperse datasets and the maximum probability density values are lower compared to Model 3 and Model 3.1. When selecting only voxels containing the highest 10% of probability density values for visualisation as in **Figure 4.7**, the results did not indicate particular space-time volumes within the storm. A lower threshold of 0.1% of probability density values was selected. The procedure of finding optimal bandwidth values, instead of selecting $\{s_{MAX}, \tau_{MAX}\}$ directly from **Table 4.1** (as discussed in **Section 2.5**), allows us to identify additional higher density value structures within the dataset. The kernel density estimation results for Model 3 and its variations with the highest 0.1% of probability density values is shown in **Figure 4.9**.

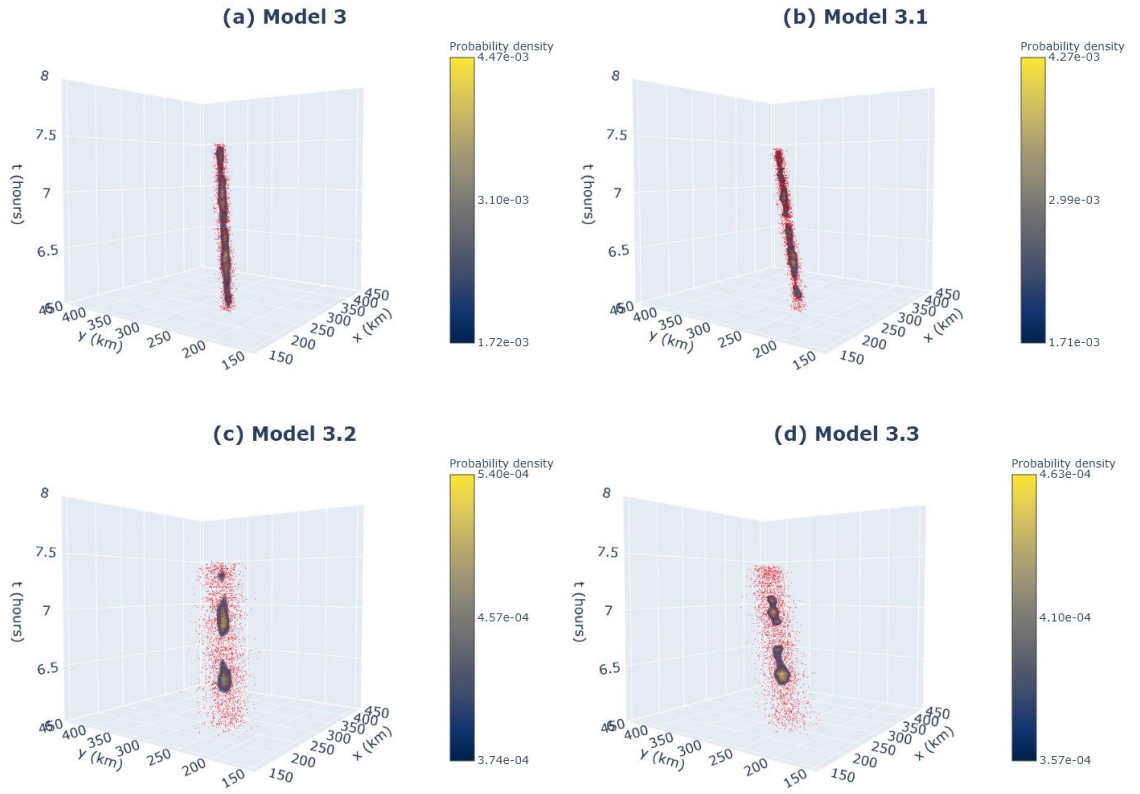

Figure 4.9. Volume rendering of voxels with highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the **(a)** single thunderstorm reference model (**Model 3**) and its variations: **(b) Model 3.1**, **(c) Model 3.2** and **(d) Model 3.3**. For each dataset, red points indicate the corresponding point events, but are decreased in size compared to **Figure 4.7**. The probability density value is given by the colour scale. Note the variations in the colour scale and its upper limits between all subplots and from **Figure 4.7**.

In **Figure 4.9**, the effect of the model parameters on the space-time volumes with the highest 0.1% of probability density values is clear. For Models 3.2 and 3.3, the value threshold allows to identify several space-time volumes within the model run that have a higher density of point events relative to the rest of the datasets. These volumes also extend further away from the movement track. For Models 3 and 3.1, the identified space-time volumes follow the track closer, while also indicating positions along the track with relatively more significant extent.

For Model 4 and its variations as listed in **Table 3.7**, the kernel density estimation methodology can be applied through the same approach as for Model 3 and its variations in **Figure 4.8**. The voxel dimensions were kept the same at {5 km, 5 km, 0.05 h}, covering the same spatial intervals from 150–450 km, but for an extent temporal interval of 6–11 h. Using the K-function analysis results (**Figure 4.4** and **Table 4.1**), a reasonable set of bandwidth values for these models was selected as {2, 5, 10, 30, 50, 70} km for the spatial kernels and {0.05, 0.1, 0.3, 0.5,

0.7} h for the temporal kernel. The leave-one-out density estimation for the optimal bandwidth values returned the same sets of bandwidths of $\{h_x = 2$ km, $h_y = 2$ km, $h_t = 0.05$ h$\}$ for Model 4 and Model 4.1 and $\{h_x = 10$ km, $h_y = 10$ km, $h_t = 0.1$ h$\}$ for Model 4.2 and Model 4.3. The volume rendering results for the reference Model 4 and its variations are shown in **Figure 4.10**.

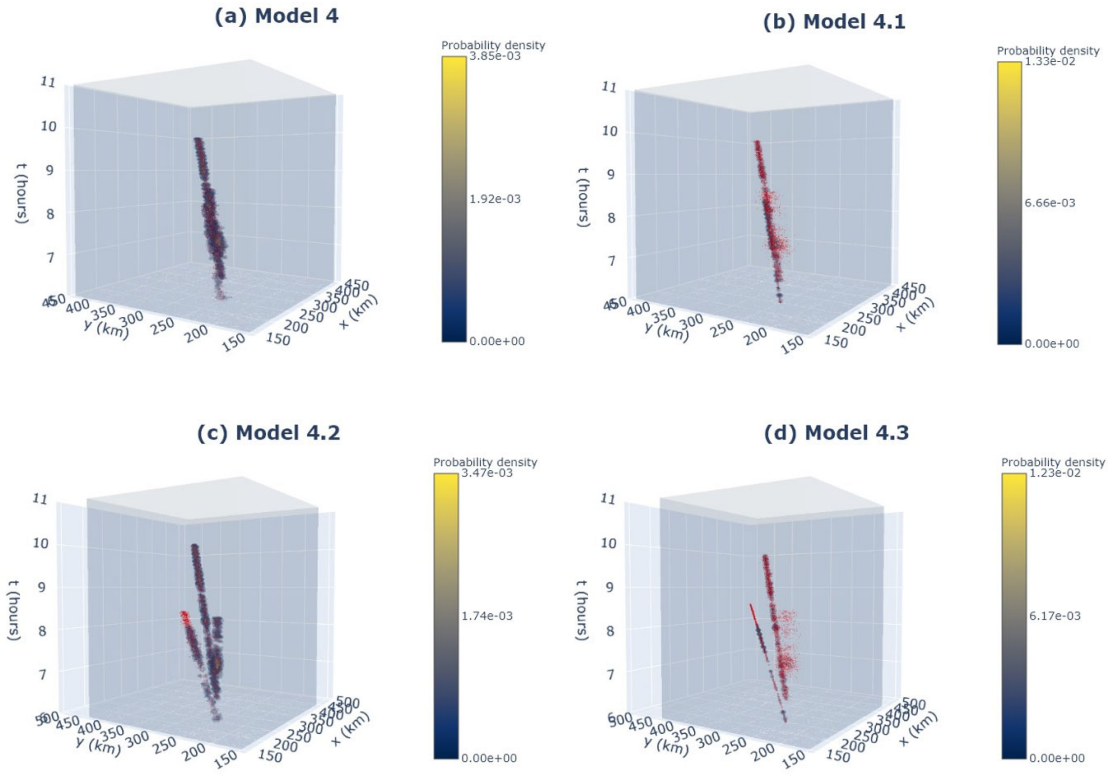

Figure 4.10. Volume rendering of probability density values as returned by the kernel density estimation methodology for the **(a)** three thunderstorm reference model (**Model 4**) and its variations: **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. For each dataset, red points indicate the corresponding point events. The probability density value is given by the colour scale. Note the variations in the colour scale upper limits between all subplots.

In **Figure 4.10**, when three storm sources have significantly different distance to the movement track value generation covariance matrix values (Model 4.1 in **Figure 4.10b** and Model 4.3 in **Figure 4.10d**), the kernel density estimation methodology is able to identify the space-time volumes associated with the sources that have the smaller values {S1, S2}. To further illustrate this, a lower threshold of 0.1% of probability density values was selected for Model 4 and its variations, identical to Model 3 in **Figure 4.9**. The kernel density estimation results for with the highest 0.1% of probability density values are shown in **Figure 4.11**.

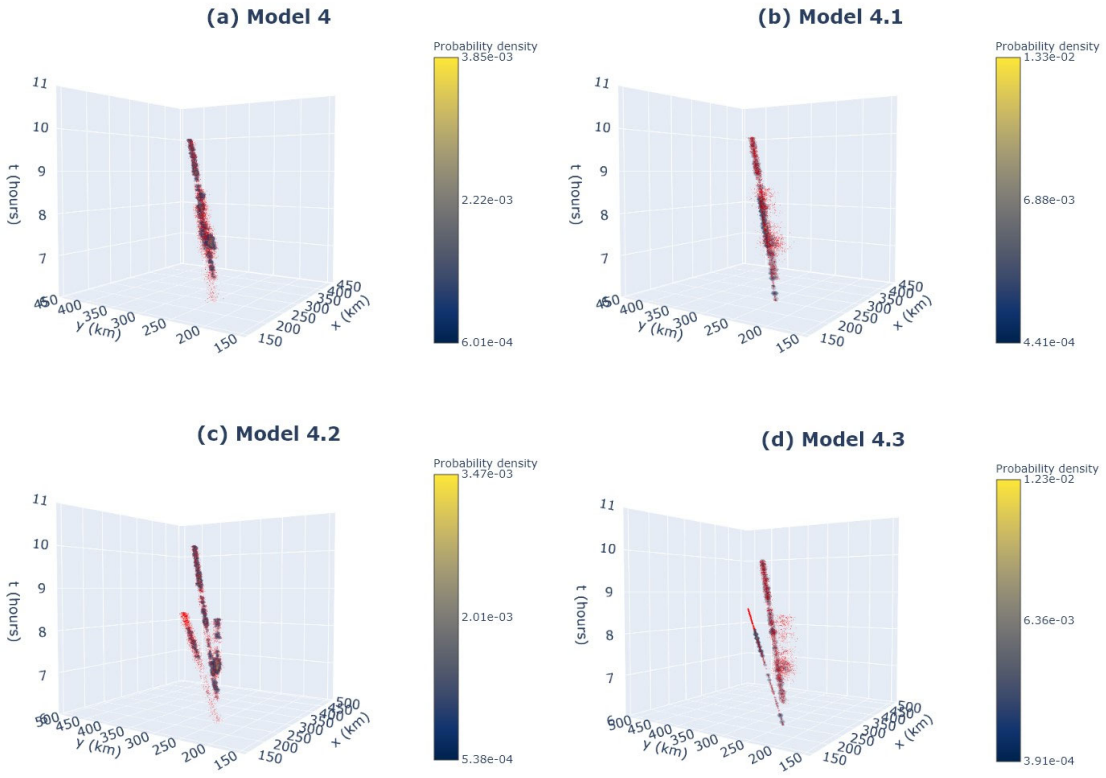

Figure 4.11. Volume rendering of voxels with highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the **(a)** reference three thunderstorm model (**Model 4**) and its variations: **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. For each dataset, red points indicate the corresponding point events, but are decreased in size compared to **Figure 4.10**. The probability density value is given by the colour scale. Note the variations in the colour scale and its upper limits between all subplots and from **Figure 4.10**.

In **Figure 4.11**, for Models 4 and 4.2, the value threshold allows to identify several space-time volumes within all three sources with higher density of point events relative to the rest of the datasets. These volumes also extend further away from the movement track. For Models 4.1 and 4.3, the identified space-time volumes follow the track closer, while also indicating positions along the track with relatively more significant extent, except for S3 with increased covariance matrix multiplication factor, resulting in no significant structure identified within the point events produced by S3 with this particular value threshold.

The kernel density estimation results for Models 1–4 showed clear ability to select a subset of the available voxels with higher probability values, given a particular value threshold. As all voxels are of identical space-time volume within the same model, these identified voxels can

be used as centroid locations for the space-time scan statistic (**Section 4.2.3**) as per the spatio-temporal clustering methodology application framework (**Section 2.5**).

4.2.3. Space-time scan statistic

For the elliptic space-time scan statistic methodology (**Section 2.3.1.3**), the necessary input parameters define the location, shape and size of the elliptic cylinder that is used as the scanning volume around the given centroid location. The complete set of such parameters is defined by the centroid location, length of the semi-major axis, shape of the elliptic base, angle between the semi-major axis and first spatial dimension and the height of the space-time cylinder.

Depending on the available computational resources, the level of computational complexity can be increased by considering extensive intervals for each of the parameters. To limit the set of necessary input parameter iterations that the methodology has to iterate through, the spatio-temporal clustering methodology application framework (**Section 2.5**) is applied. It consists of a sequence of methodologies that give additional insight into the clustering characteristics of the point event data and limits this set of scan-statistic input parameters.

The results of the K-function analysis (**Figures 4.1–4.5**) are able to identify the spatial and temporal extents where space-time interaction is present between the point events. These results are summarised in **Table 4.1** and are used to limit the height and the length of the semi-major axis parameters of the elliptic cylinder scanning volume. The results of the kernel density estimation for particular datasets (**Figures 4.6–4.11**) allow to focus on particular space-time volumes that show increased probability density values based on the presence of point events.

The initial set of centroid locations and size parameters of the elliptic cylinder can be adjusted quickly after the K-function and kernel density estimation results. As each permutation of the input parameters would have to be run individually, the computational complexity of the given methodology can be reduced. By only considering a subset of centroid locations, more computational resources can be used on analysing the scan statistic performance on other input parameters, enabling the scan statistic to cover a wider range of input parameters.

The elliptic space-time scan statistic clustering results for the homogeneous point event data (Model 1) are not presented. During the implementation of the space-time scan statistic, the homogeneous Poisson point process is selected as the background point process, invalidating

the use of homogeneous Poisson process as the comparable point process. Point event clusters can be identified for the regular grid data (**Appendix O**), but they are not useful for establishing any further benchmarks on the synthetic datasets. The space-time scan statistic is then initially applied to the clustered point event dataset (Model 2) to assess the ability of the methodology to identify individual clusters. Within the space-time cube $\{X = 20, Y = 20, T = 20\}$ used for generating Model 2, the necessary input parameters for the elliptic space-time scan statistic and the sets of their values selected are given in **Table 4.2**.

Table 4.2. Input parameters and their values for the space-time scan statistic for **Model 2**.

Input parameter	Input values
Centroid locations	Voxels from **Figure 4.7b**
Length of semi-major axis	{0.1, 0.2, …, 3}
Shape of elliptic base	{1, 1.5, …, 3}
Angle	$\{0, (1/6)\pi, \ldots, (5/6)\pi\}$
Height	{0.1, 0.2, …, 3}

From the given input parameter values in **Table 4.2**, permutations of the values are generated and kept constant for implementation of the space-time scan statistic. Various angles of the spatio-temporal domain are used for visualising the clustering results of the space-time scan statistic. All point events within the scanning volume are considered part of the cluster as identified by the methodology. The elliptic space-time scan statistic results for Model 2 are shown in **Figure 4.12**, and the input parameters and their values for defining the elliptic cylinder encompassing the identified cluster are listed in **Table 4.3**.

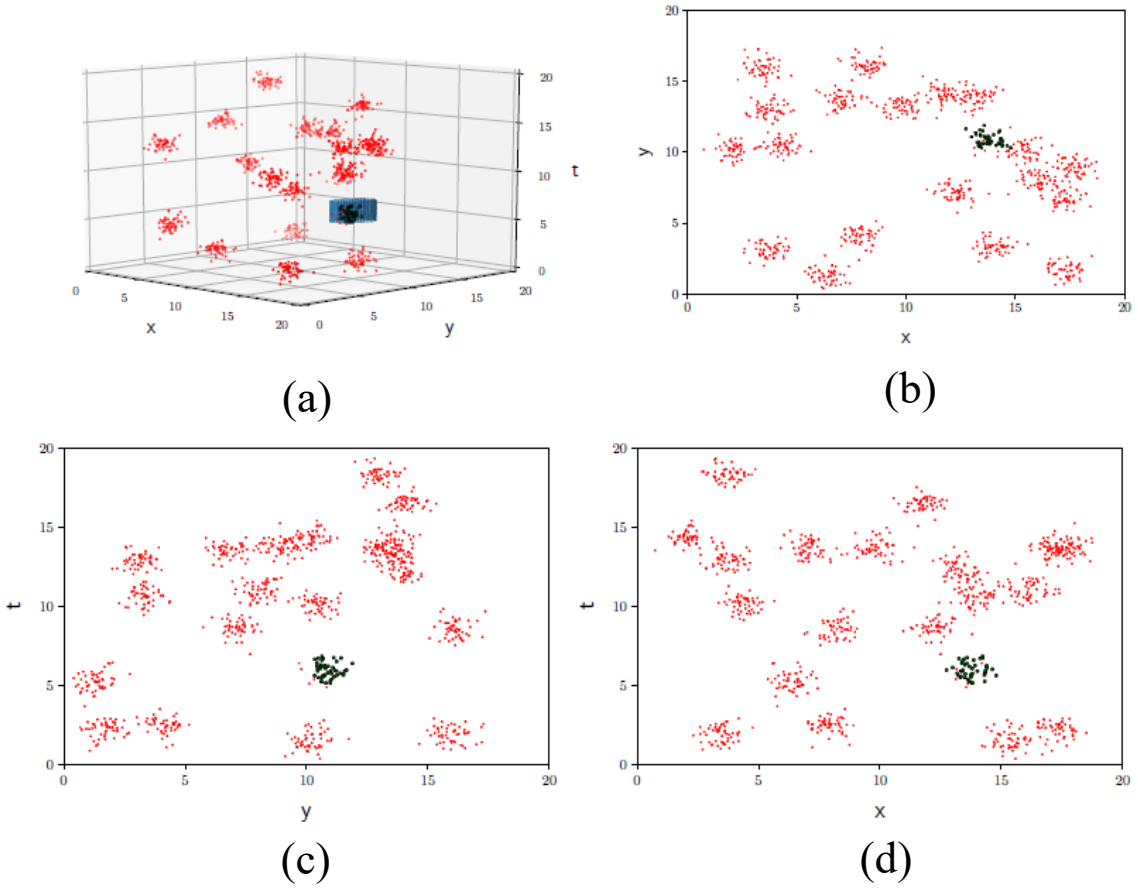

Figure 4.12. Various angles [**(a)** 3D view of the spatio-temporal domain, **(b)** 2D spatial view, **(c)** 2D *t*–*y* view, **(d)** 2D *t*–*x* view] of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to clustered point event data. In all subfigures, the events contained within the scan statistic are coloured green, while those outside in red. In **(a)**, the elliptic scan statistic is drawn in transparent light blue.

Table 4.3. Identified input parameter values by the elliptic space-time scan statistic on clustered point event data (**Model 2**).

Input parameter	Identified values
Centroid location	{13.8, 11.2, 6.2}
Length of the semi-major axis	1.6
Shape of elliptic base	2
Angle	$(2/3)\pi$
Height	2.2

For the clustered point event data in Model 2, the scan statistic has identified one of the point events clusters (**Figure 4.12a**). This is an expected result, but identifying a particular cluster in the clustered point event dataset depends on the centroid locations. From the 2D spatial view perspective (**Figure 4.12**), the centroid location of {13.8, 11.2, 6.2} fits this particular cluster.

If the centroid is located close to the initial cluster centre (i.e. 'parent' event as defined in **Section 3.4.1**), the scanning volume will encompass more of the point events for the same set of parameters. When assuming all cluster centres would coincide with the centroid locations, the scan statistic results would be expected to identify the cluster with the lowest standard deviation parameter σ value (**Table 3.5** and **Section 3.4.1**). Without noting each cluster's individual parameters, it is impossible to assess whether the space-time scan statistic has identified the cluster with the lowest σ value.

For Model 3 and its variations, the K-function analysis results of the spatial and temporal ranges values $\{s_{MAX}, \tau_{MAX}\}$ for maximum clustering measure D value are summarised in **Table 4.1**. The results of the kernel density estimation for these datasets with a probability density value threshold of 0.1% (**Figure 4.9**) indicate the space-time voxels that can be selected as centroids of the elliptic cylinder scanning volume. Within the extended space-time cube $\{X = 1000$ km, $Y = 1000$ km, $T = 24$ h$\}$ used for generating Models 3 and 4, the input parameters for the elliptic space-time scan statistic and the sets of their values selected are given in **Table 4.4**. The centroid locations are directly taken from the voxel locations in **Figure 4.9**, while the maximum values for the length of the semi-major axis and the height (from the centroid) of the elliptic cylinder scanning volume are taken as $\{s_{MAX}, \tau_{MAX}\}$ from **Table 4.1**. The step size for individual parameters is indicated and depends on the desired cluster dimension accuracy and computational resources.

Table 4.4. Input parameters and their values for the space-time scan statistic for single storm reference Model **3** and its variations (**Models 3.1, 3.2, 3.3**).

Input parameter	**Input values**			
	Model 3	**Model 3.1**	**Model 3.2**	**Model 3.3**
Centroid locations	From **Fig. 4.9a**	From **Fig. 4.9b**	From **Fig. 4.9c**	From **Fig. 4.9d**
Length of semi-major axis	{5, 10, …, 25}	{5, 10, …, 60}	{5, 10, …, 40}	{5, 10, …, 75}
Shape of elliptic base	{1, 1.5, …, 3, 3.5, 4}			
Angle	$\{0, (1/8)\pi, \ldots, (7/8)\pi\}$			
Height	{0.05, 0.1, …, 0.45}	{0.05, 0.1, …, 0.4}	{0.05, 0.1, …, 0.5}	{0.05, 0.1, …, 0.45}

From the given input parameter values in **Table 4.4**, permutations of the values are generated. Only the 3D view of the spatio-temporal domain is for visualising the clustering results of the space-time scan statistic for Model 3 and its variations. The elliptic space-time scan statistic results for Model 3 and its variations are shown in **Figure 4.13** and the input parameters and their values for defining the elliptic cylinder encompassing the identified cluster are listed for each individual model in **Table 4.5**.

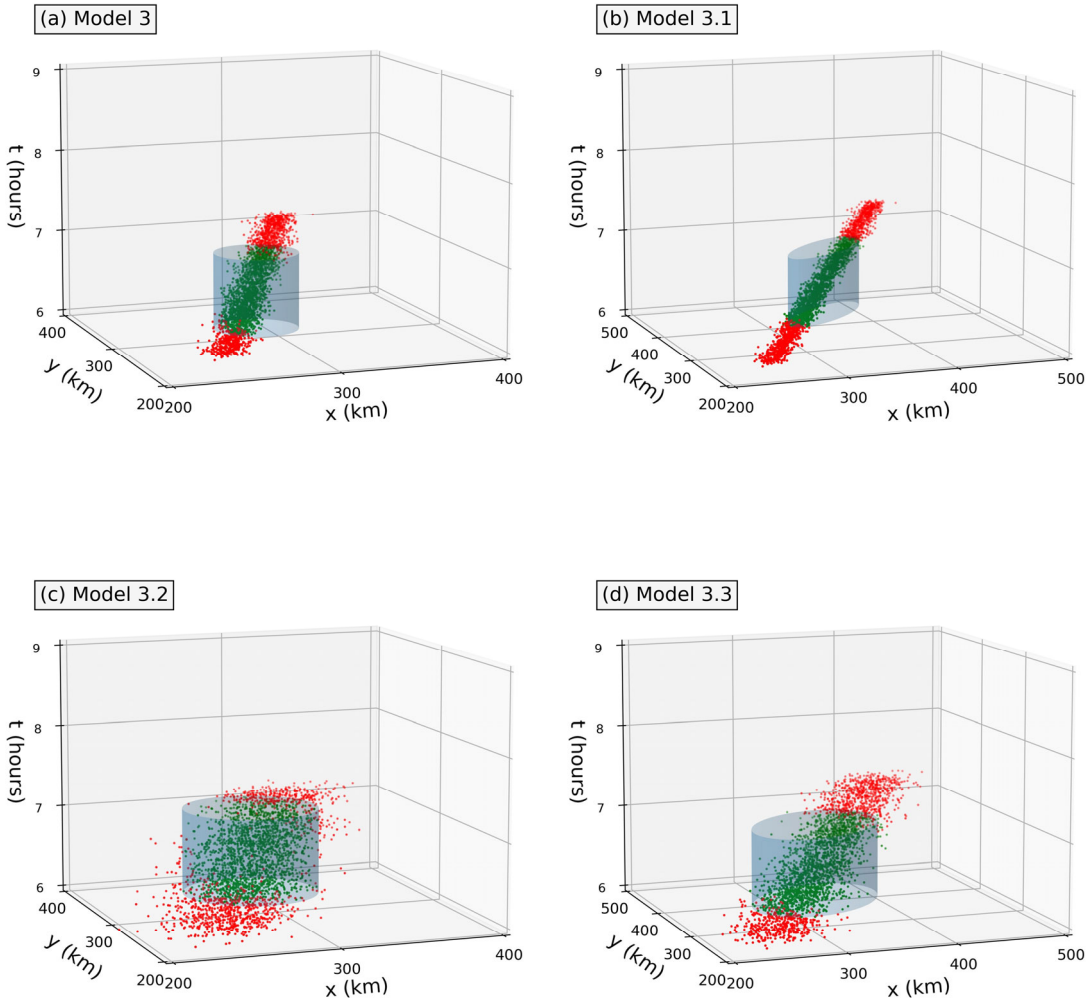

Figure 4.13. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **(a) Model 3 (b) Model 3.1**, **(c) Model 3.2**, and **(d) Model 3.2**. In all subfigures, the events contained within the space-time volume as identified by the scan statistic are coloured green, while those outside in red. The elliptic scanning volume is drawn in transparent light blue. Note the varying spatial limits between [**(a)**, **(c)**] and [**(b)**, **(d)**].

Table 4.5. Identified input parameter values by the elliptic space-time scan statistic on single storm reference **Model 3** and its variations (**Models 3.1, 3.2, 3.3**).

Input parameter	**Identified values**			
	Model 3	**Model 3.1**	**Model 3.2**	**Model 3.3**
Centroid location	{275, 270, 6.65}	{325, 325, 6.70}	{275, 280, 6.80}	{315, 320, 6.65}
Length of the semi-major axis	25 km	60 km	40 km	75 km
Shape of elliptic base	1	3	1	1.5
Angle	N/A	$(1/4)\pi$	N/A	$(1/4)\pi$
Height	0.45 h	0.40 h	0.50 h	0.45 h

The identified elliptic cylinders as space-time volumes (**Figure 4.13**) and their input parameter values (**Table 4.5**) for Model 3 and its variations indicate that the spatio-temporal clustering analysis framework has successfully identified local clusters of point events within the larger datasets. With comparing the identified input parameter values of the cylinders between the different models, certain limitations of the analysis framework and individual methodologies are apparent. Some limitations of the simulation study and the generation of synthetic single storm models can also be noted. Initial evaluation of these limitations is further continued in **Section 4.2.4** after a discussion on the results for Model 4 and its variations.

The two input parameters directly influenced by the K-function analysis results as per the framework – the length of the semi-major axis and the height are the maximum values of their respective input value ranges (**Table 4.4**). For these parameters, the identified $\{s_{MAX}, \tau_{MAX}\}$ values from **Table 4.1** were used as the value range maxima. The exact maxima selection is dependent on the particularities of the dataset and the aim of the clustering analysis – to identify smaller clusters with relatively higher point event density within a larger cluster or to identify separate larger clusters. For other natural hazards and fields of study, these spatial and temporal ranges of the scanning volume were restricted by the study volume (e.g. 50% [**Tonini *et al.*, 2022**; **Parente *et al.*, 2016**; **Kulldorff *et al.*, 2006**] and 10% [**McNally and Colver, 2008**] of the study area and duration).

The exact centroid location for the identified space-time volume would vary slightly with the particular initialisation of the model run, but the movement speed v changes (**Table 3.5**) between Model 3 and its variations introduces eccentricity in the base shape and an angle relative to the first spatial dimension (i.e. x dimension) of the scanning volume cylinder. For Models 3 and 3.2, the scanning volume is a cylinder as for the regular space-time scan statistic without any eccentricity. As movement speed increases for Models 3.1 and 3.3, the scanning volume aligns with the movement direction of the storm and eccentricity is introduced. This eccentricity is reduced when the distance to the movement track is increased between Models 3.1 and 3.3, halving the ratio between the ellipse axes.

For Model 4 and its variations, the K-function analysis results of the spatial and temporal ranges values $\{s_{MAX}, \tau_{MAX}\}$ for maximum clustering measure D value are summarised in **Table 4.1**. The results of the kernel density estimation for these datasets with a probability density value threshold of 0.1% (**Figure 4.11**) indicate the space-time voxels that can be selected as centroids of the elliptic cylinder scanning volume. The input parameters for the elliptic space-time scan statistic and the sets of their values selected are given in **Table 4.6** and the results shown in **Figure 4.14**. The centroid locations are directly taken from the voxel locations in **Figure 4.11**, while the maximum values for the length of the semi-major axis and the height (from the centroid) of the elliptic cylinder scanning volume are taken as $\{s_{MAX}, \tau_{MAX}\}$ from **Table 4.1**.

Table 4.6. Input parameters and their values for the space-time scan statistic for single storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**).

Input parameter	**Input values**			
	Model 4	**Model 4.1**	**Model 4.2**	**Model 4.3**
Centroid locations	From **Fig. 4.11a**	From **Fig. 4.11b**	From **Fig. 4.11c**	From **Fig. 4.11d**
Length of semi-major axis	{5, 10, …, 60}	{5, 10, …, 45}	{5, 10, …, 60}	{5, 10, …, 100}
Shape of elliptic base	{1, 1.5, …, 3, 3.5, 4}			
Angle	$\{0, \frac{\pi}{4}, \ldots, \frac{3\pi}{4}\}$			
Height	{0.05, 0.1, …, 1.05}	{0.05, 0.1, …, 0.75}	{0.05, 0.1, …, 1}	{0.05, 0.1, …, 0.7}

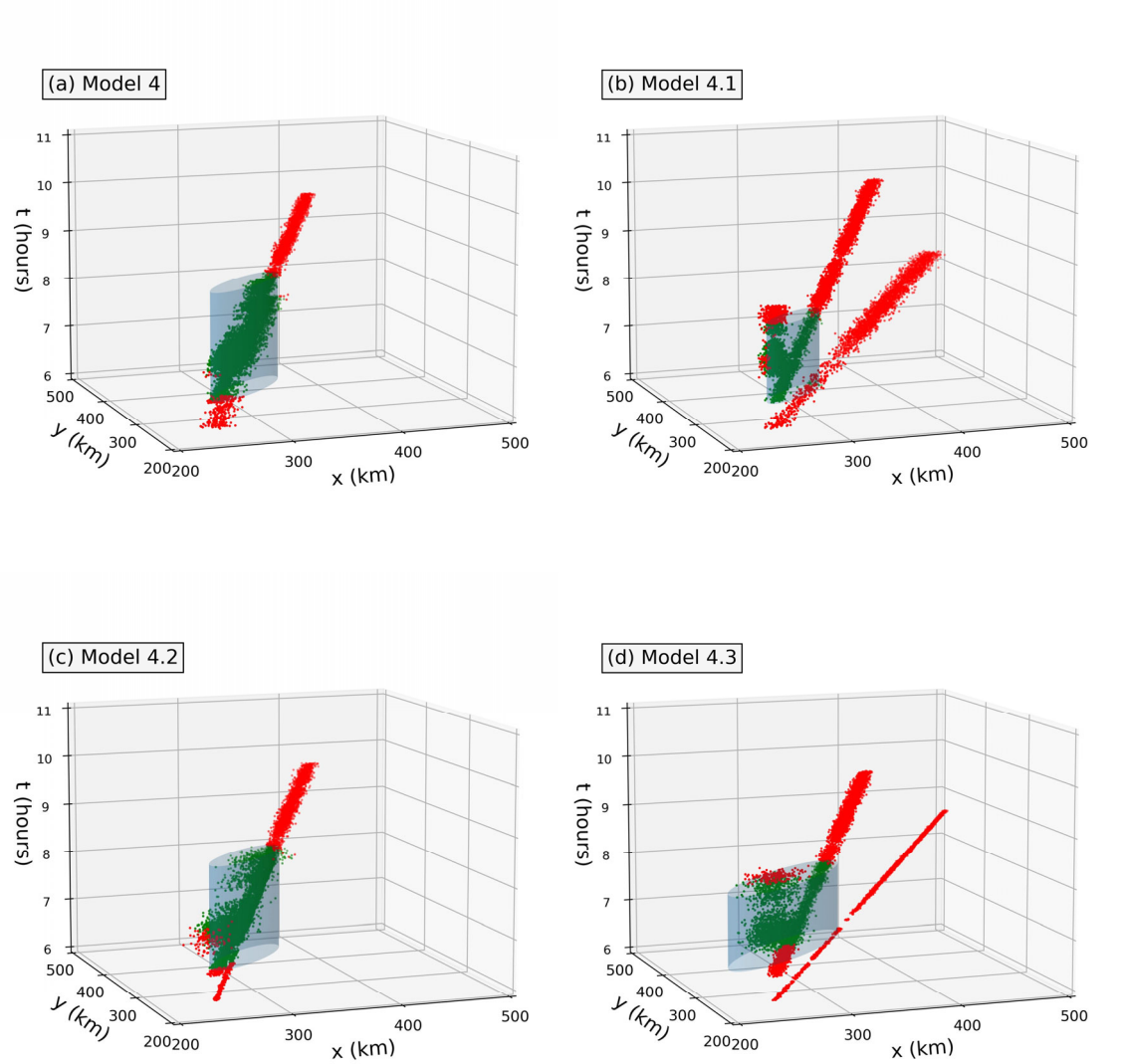

Figure 4.14. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **(a) Model 4**, **(b) Model 4.1**, **(c) Model 4.2**, and **(d) Model 4.2**. In all subfigures, the events contained within the space-time volume as identified by the scan statistic are coloured green, while those outside in red. The elliptic scanning volume is drawn in transparent light blue.

Table 4.7. Identified input parameter values by the elliptic space-time scan statistic on three storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**).

Input parameter	**Identified values**			
	Model 4	**Model 4.1**	**Model 4.2**	**Model 4.3**
Centroid location	{290, 290, 7.55}	{275, 275, 7.25}	{290, 290, 7.60}	{260, 255, 7.65}
Length of the semi-major axis	60 km	45 km	60 km	95 km
Shape of elliptic base	4	3.5	3.5	4
Angle	$\frac{\pi}{4}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$
Height	1.05 h	0.75 h	1.00 h	0.70 h

The identified elliptic cylinders as space-time volumes (**Figure 4.14**) and their input parameter values (**Table 4.7**) for Model 4 and its variations identify the already existing limitations when discussing the scan statistic results for Model 3 and its variations in **Figure 4.13**. The spatial and temporal extents of the scanning volume are the respective input parameter value range maxima (**Table 4.6**). Additional ways of restricting these exact input parameter ranges were mentioned in discussion after **Figure 4.13** and is continued in **Section 4.2.4**. The shape and the angle parameters of the identified elliptic cylinder are similar for all models, identifying the movement direction of the three storms as per the model generation framework (**Section 3.3.5**).

Compared to single storm Model 3 (and its variations), the main identified limitation of the elliptic scanning volume when applied to the three storm Model 4 (and its variations) is the identification of a cluster containing point events from multiple storms. The elliptic base of the scanning volume allows the methodology to identify clusters of point events with radially varying spatial extents from the volume centroid, but there is no parameter accounting for the movement of the point event source. When extending the scanning volume temporally, the initial separation of 0.5 h is not significant enough to limit the scanning volume (based on the selected height value range) from including multiple storm point events within the cluster. The scan-statistic methodology identified the space-time volume with the highest density of point events but failed to separate an individual storm as a local cluster. Discussion on potential improvements of the limitations presented within this section follows in **Section 4.2.4**.

4.2.4. Initial evaluation of the spatio-temporal clustering framework

In **Sections 4.2.1**–**4.2.3**, the spatio-temporal clustering methodology application and analysis framework (**Section 2.5**) was applied to the various synthetic point event datasets of the simulation study (**Section 3.4.1**). The objective of applying the framework was to provide insight into the global clustering scales, identify individual clusters and use the combination of these methodologies to aid the computational complexity and clustering performance. Certain limitations and immediate results have been discussed within each results section, but a more comprehensive evaluation is necessary to understand potential improvements to individual methodologies and the overall application framework.

The K-function analysis (**Section 4.2.1**) indicated the spatial and temporal ranges $\{s_{MAX}, \tau_{MAX}\}$ over which significant space-time interaction exists between point events. These values were used to inform the kernel density estimation methodology (**Section 4.2.2**) on calculating optimal bandwidth values and the scan statistic (**Section 4.2.3**) on selecting the scanning volume extent. For the kernel density estimation, the use of leave-one-out density estimation indicated that the optimal bandwidth values are significantly smaller than the identified $\{s_{MAX}, \tau_{MAX}\}$, which were used as the choice for bandwidth values in [**Tonini and Cama, 2019**].

For the scan statistic methodology, the identified scanning volumes were always at the maximum selected values for the spatial and temporal extents, indicating that identified space-time volumes were the most point-event dense. Limitations on these extents depending on the study volume have been used in a variety of studies [**Tonini *et al.*, 2022**; **Parente *et al.*, 2016**; **McNally and Colver, 2008**; **Kulldorff *et al.*, 2006**]. For the real-world dataset scan statistic application (**Section 4.4**), the identified $\{s_{MAX}, \tau_{MAX}\}$ values are halved, so that the elliptic base's major axis and the cylinder's full height correspond to the range of maximum space-time interaction. An improvement to K-function analysis methodology by standardising the results with the scanning volume to further address this issue is presented in **Section 5.2**.

The elliptic scan statistic results for Models 4 and its variations (**Figure 4.14**) indicated the inability of the methodology to identify point events clusters produced by a moving source. The identified elliptic cylinder space-time volumes were the most point-event dense, but contained events from multiple storms. Improvements to the scan statistic scanning volume dimensions and additional input parameters are presented in and discussed in **Section 5.3**.

4.3. Results and analysis for synthetic benchmark trajectory datasets

The benchmark datasets, as described and shown in **Section 3.4.2**, are used to apply the Trajectory-OPTICS density-based spatio-temporal clustering methodology (**Section 2.3.2.1** and **Appendix E**). For both the homogeneous and clustered synthetic trajectory datasets, input parameters for the first implementation step in the methodology are selected. These two parameters are the neighbourhood distance threshold ε and the minimum number of other trajectories within this distance n_{pts} (**Appendix Figure E1**). The number of assigned 'core' trajectories (i.e., central trajectories representing trajectory-dense regions) varies depending on the selected values – higher neighbourhood distance threshold ε and lower minimum number of other trajectories within this distance n_{pts} increase the number of core trajectories. In the Trajectory-OPTICS methodology, the set of core trajectories is used to measure the reachability distance to any given trajectory, which affects the ordered reachability distance plots. To evaluate the parameter ranges of $\{\varepsilon, n_{pts}\}$ for both datasets, the number of core trajectories n_{CT} as found for the various combinations of input parameter values is shown in **Figure 4.15**.

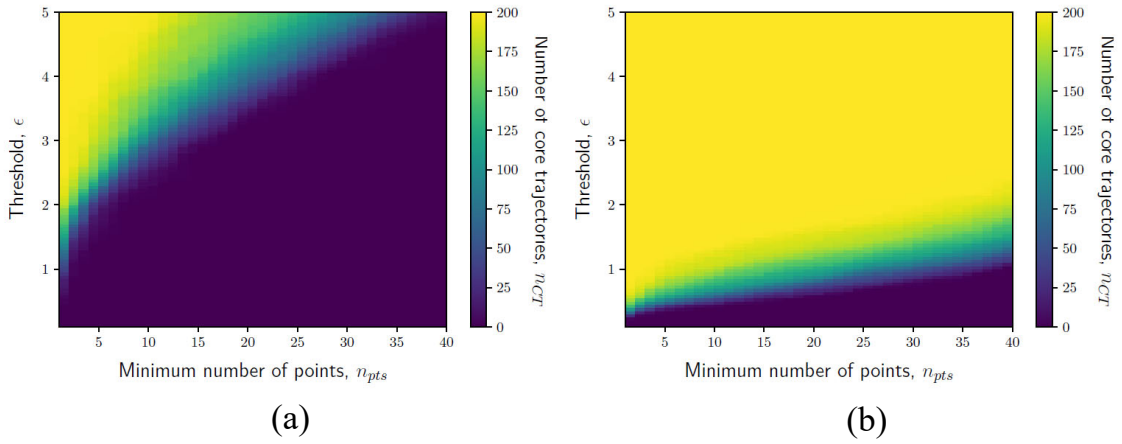

Figure 4.15. The number of core trajectories n_{CT} for various neighbourhood distance threshold ε and minimum number of trajectories within given threshold n_{pts} combinations for **(a)** homogeneous and **(b)** clustered trajectories. The number of core trajectories n_{CT} is indicated by the colour scale. There are a total of $N = 200$ trajectories in each dataset.

As the sets of selected input parameter $\{\varepsilon, n_{pts}\}$ ranges were the same for both synthetic datasets, **Figure 4.15** shows varying behaviours for the number of core trajectories n_{CT} . While for both homogeneous (**Figure 4.15a**) and clustered (**Figure 4.15b**) trajectories n_{CT} increases with ε, but decreases with n_{pts}, for the homogeneous trajectories larger ε values are required to assign the same number of core trajectories.

For clustered trajectories, almost all trajectories are assigned as 'core' with significantly smaller ε values. Such variation would be expected as the spatio-temporal distance (as defined in **Equation E.1**) can be understood as 'the average Euclidean distance over all time iterations of a given trajectory pair'. Each of the homogeneous trajectories has an independent set of drift and standard deviation parameters, while for clustered trajectories, these parameters are shared between all trajectories within a cluster. When measuring 'the average Euclidean distance over all time iterations', it would be expected to have more trajectories contained within the same distance threshold if these trajectories share the same drift and standard deviation parameters (unless $n_{pts} > 50$ as each individual trajectory cluster has 50 trajectories).

To continue the implementation of the Trajectory-OPTICS methodology, the sets of input parameters were selected as $\{\varepsilon = 5, n_{pts} = 10\}$ and $\{\varepsilon = 1, n_{pts} = 10\}$ for the homogeneous and clustered trajectory datasets respectively. This would assign most of the trajectories as 'core' and identify only the furthest trajectories as noise. With these input parameters, the resulting reachability distance plots are shown in **Figure 4.16**.

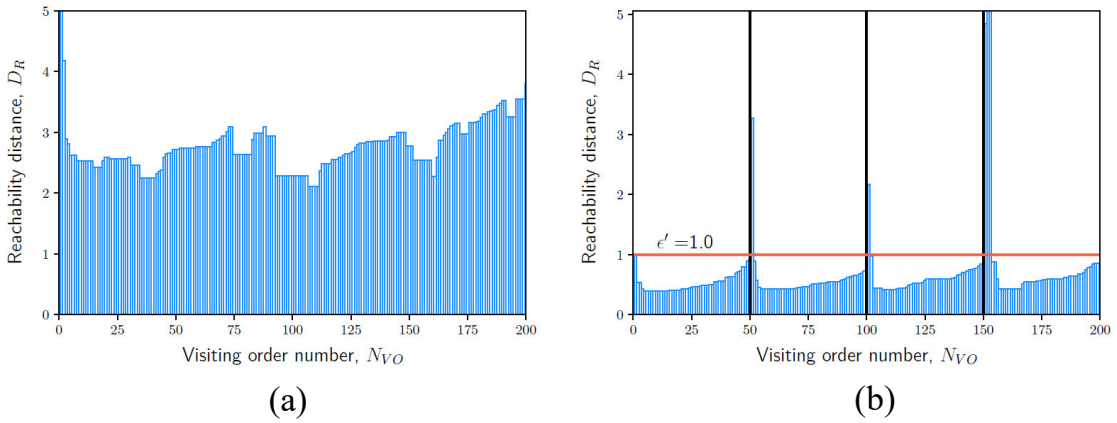

Figure 4.16. Reachability distance D_R plots for **(a)** homogeneous and **(b)** clustered trajectories as numbered by their visiting order number N_{VO}. The reachability distance threshold ε' is indicated by a red line and its value given above it. Black vertical lines indicate first trajectories for which $D_R \geq \varepsilon'$. There are a total of $N = 200$ trajectories in both datasets.

The reachability distance D_R plots indicate two contrasting behaviours as would be expected for these two synthetic trajectory dataset benchmarks. For the clustered trajectories (**Figure 4.16b**), the reachability distances exhibit the expected behaviour of four intervals along the visiting order number N_{VO} that have low D_R values and are separated by sharp peaks in D_R. The third parameter necessary to implement Trajectory-OPTICS is the reachability distance

threshold ε' (**Appendix E**), which separates the trajectories into individual trajectory clusters and 'noise' trajectories.

All trajectories with a D_R value larger than ε' relative to any previously visited core trajectories are then classified as noise, while the four intervals of $D_R < \varepsilon'$ are separated and assigned to clusters. From a purely visual inspection (separating local peaks in D_R values [**Nanni and Pedreschi, 2006**]) this reachability distance threshold was set to $\varepsilon' = 1$. The trajectories classified as noise also occur approximately when all the trajectories within each synthetically generated cluster are visited (i.e. at approximately $N_{VO} = 50$, $N_{VO} = 100$ and $N_{VO} = 150$).

The reachability distance values (**Figure 4.16a**) are relatively higher for the homogeneous trajectories than for the clustered trajectories, with no clear low-value intervals. While setting $\varepsilon' = (2.5, 3)$ would return several relatively small clusters, the lack of structure in the plot relative to the clustered data was expected. To confirm that the Trajectory-OPTICS algorithm for the clustered trajectories has identified the four generated clusters, the visiting order of all trajectories and the resultant division of the dataset into clusters is shown in **Figure 4.17**.

The clustering results of the Trajectory-OPTICS methodology (**Figure 4.17b**), when compared to the four generated clusters as seen in **Figure 3.26b**, confirm that the intervals of low D_R values identified in **Figure 4.16b** correspond to the four individual trajectory clusters. The plot of visiting order of trajectories (**Figure 4.17a**) confirms the expected behaviour of the methodology as all trajectories within each cluster are visited first and moving into a different cluster creates the peaks in D_R values observed in **Figure 4.16b**. Trajectories with D_R values above the set $\varepsilon' = 1.0$ threshold were categorised as noise.

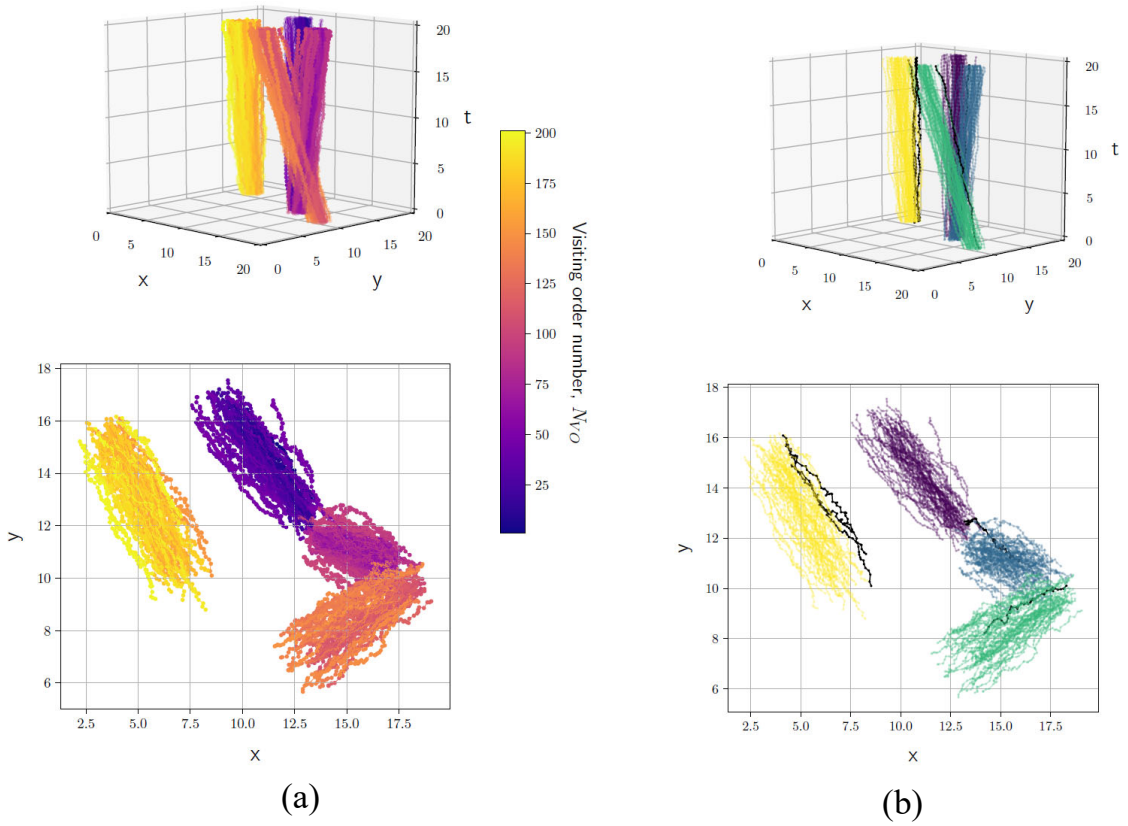

Figure 4.17. (a) The visiting order and **(b)** clustering results after the execution of the Trajectory-OPTICS methodology on clustered trajectories. Visiting order number N_{VO} in **(a)** is indicated by the colour scale. Each identified cluster in **(b)** is assigned a unique colour, while all noise trajectories ($D_R \geq \varepsilon$') are coloured in black. Upper subfigures show the propagation of the trajectories in the space-time cube, while lower subfigures illustrate the different propagation directions once trajectories are collapsed onto the two spatial dimensions. There are a total of $N = 200$ trajectories with 50 trajectories in each of the four generated clusters.

4.4. Results and analysis for real-world point event data

The spatio-temporal clustering methodology analysis framework (**Section 2.5**) was applied to various synthetic point event datasets as part of the simulation study (**Section 3.4.1**) and their clustering results were presented, analysed and evaluated (**Section 4.2**) with reference to the application framework. The results of applying all three methodologies within the common framework to the real-world lightning strike dataset presented in **Section 3.2.1** are shown and analysed in this section. Similarly to **Section 4.2**, the selected necessary input parameters are given for each methodology. The clustering results for each methodology are presented and analysed separately within the context of the individual methodology.

The K-function results as the global clustering measure are displayed in **Section 4.4.1**, the kernel density estimation results to identify particular space-time volumes of higher density values are shown in **Section 4.4.2** and the space-time scan statistic results to identify particular clusters are shown in **Section 4.4.3**. A discussion on the clustering results of the methodology framework follows in **Section 4.4.4** with final evaluation of the application framework after synthetic and real-world dataset applications.

4.4.1. K-function analysis

The performance benchmarks of the K-function analysis framework were established in **Section 4.2.1** for four different models and their variations. Within the common application framework, K-function analysis results inform the set of potential bandwidth values for the kernel density estimation and the extent of clustering present in the data, limiting the input parameter permutations for the space-time scan statistic. In this section, the clustering patterns of the K-function results from **Section 4.2** will be compared to the results of the methodology when applied to the real-world lightning strike location data (**Section 3.2.1**). Any particularities of the methodologies when changing their applications from synthetic to real-world datasets will be noted and justified.

To apply the K-functions to the real-world dataset, spatial and temporal range values, over which the global clustering measure is calculated, have to be set. For the real-world dataset, the lightning strikes recorded over 1 July 2015 are placed in the same {X = 1000 km, Y = 1000 km, T = 24 h} space-time volume as for Models 3 and 4, when the lightning strike spatial coordinates are projected onto the OSGB 1936 (EPSG:27700) northing/easting coordinates (**Section 3.3**). Considering the scale of the dataset from a visual assessment (**Figure 3.1**), the spatial and temporal ranges are set as {5, 10, ..., 400} km and τ = {0.1, 0.2, ..., 4.5} h. The results of the global clustering measure D and its components — the spatio-temporal K-function $K(s,\tau)$ and the product of the spatial and temporal K-functions $K(s)K(\tau)$ as applied to the real-world dataset are shown in **Figure 4.18**.

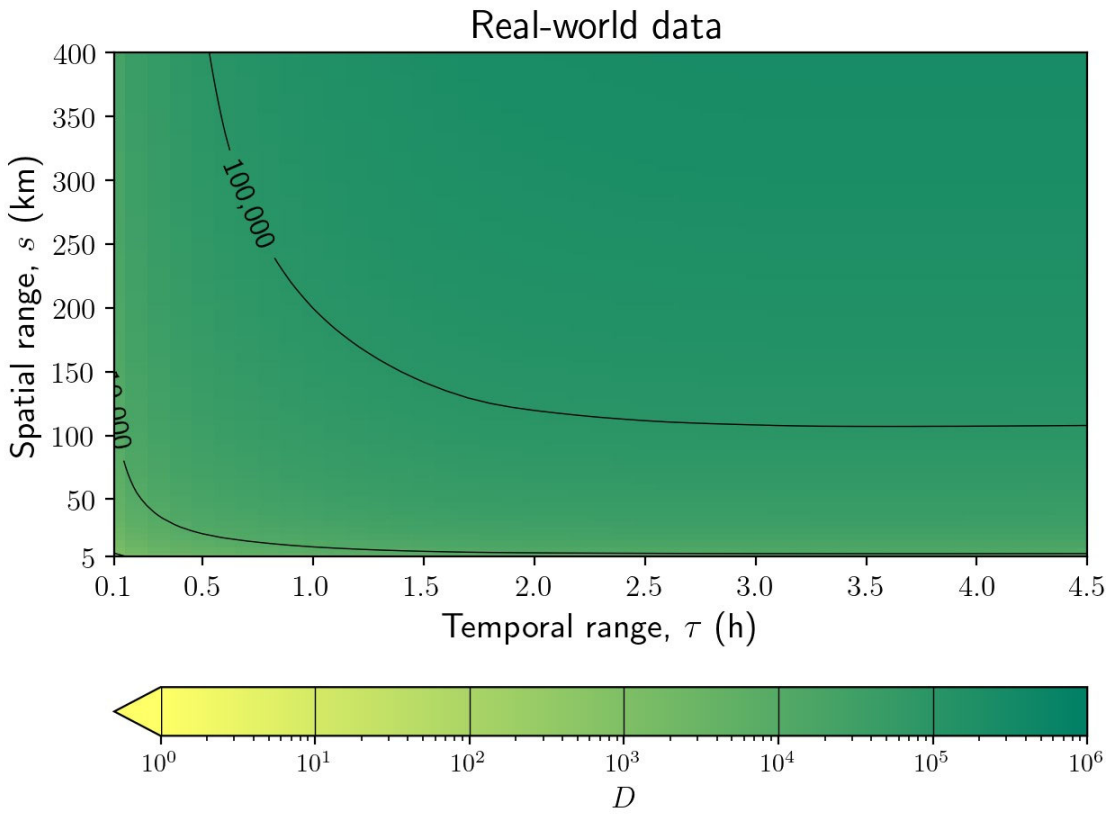

Figure 4.18. The clustering measure $D = K(s,\tau) – K(s)K(\tau)$ values [spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$] as applied to lightning strike location data. Their values are calculated for various combinations of spatial range s (km) and temporal range τ (hours). Note the different colour scale upper limit for the clustering measure D from K-function analysis framework result figures in **Section 4.2**.

Compared to the clustering measure D values of the various synthetic simulation study datasets in **Section 2.2**, the results in **Figure 4.18** exhibit behaviour clearly indicative of a clustered dataset over the selected spatial and temporal ranges. For the selected spatial and temporal range combinations, the global clustering measure shows a pattern of increasing clustering measure D values. It reaches peak value for τ = 4.4 h and s = 320 km, indicating the range of clustering for the lightning strike locations.

As per the spatio-temporal clustering methodology application framework (**Section 2.5**), the spatial and temporal range values with maximum clustering measure D values $\{s_{MAX}, \tau_{MAX}\}$ as returned by the K-function analysis framework are used for selecting a set of potential kernel bandwidth values for the kernel density estimation methodology (**Section 4.4.2**).

4.4.2. Kernel density estimation

The kernel density estimation (**Section 2.3.1.2**) is a part of the spatio-temporal clustering methodology framework (**Section 2.5**) as it allows us to identify space-time volumes that are of interest for identification of individual clusters by any local clustering measure. Global clustering measures (i.e. K-function analysis from **Section 4.4.1**) provide the necessary estimation of the spatial and temporal scales over which there is a presence of spatio-temporal clustering.

For the real-world data, the kernel density estimation methodology can be applied through the same approach as **Section 4.2** within the $\{X = 1000$ km, $Y = 1000$ km, $T = 24$ h$\}$ space-time volume. Using the K-function analysis results, a reasonable set of bandwidth values was selected as {2, 5, 10, 25, 50, 100} km for the spatial kernels and {0.05, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0} h for the temporal kernel. The leave-one-out density estimation for the optimal bandwidth values returned the set $\{h_x = 5$ km, $h_y = 5$ km, $h_t = 0.1$ h$\}$ for real-world dataset.

For calculating the probability density values, the voxel dimensions were selected as {10 km, 10 km, 0.1 h} covering the spatial intervals from 150–600 km in the easting x direction, 340–800 km in the northing y direction and the temporal interval from 7–24 h. The choice of these values is based on the coverage of the full lightning strike encompassing volume and the computational complexity for calculating and visualising the results. The volume rendering results for the full real-world lightning strike dataset with all probability density values and the highest 0.1% of probability density values are shown in **Figure 4.19**.

In **Figure 4.19**, the volume rendering results of all probability density values (**Figure 4.19a**) indicate that there is a clear space-time volume of higher probability density values. From the temporal and spatial location of this volume, it corresponds to the Storm 2A from the second case study in **Section 3.3.2.3**. When displaying only the highest 0.1% of probability density values (**Figure 4.19b**), the lightning strikes within Storm 2A generate the majority of the voxels left after the introduction of the threshold. When comparing these three-dimensional figures to the three-dimensional view of the three storms of the second case study (**Figure 3.10**), the additional volume (occurring between $t = 19.5$ h to $t = 20.5$ h), still identified after introduction of the value threshold, corresponds to Storm 2B.

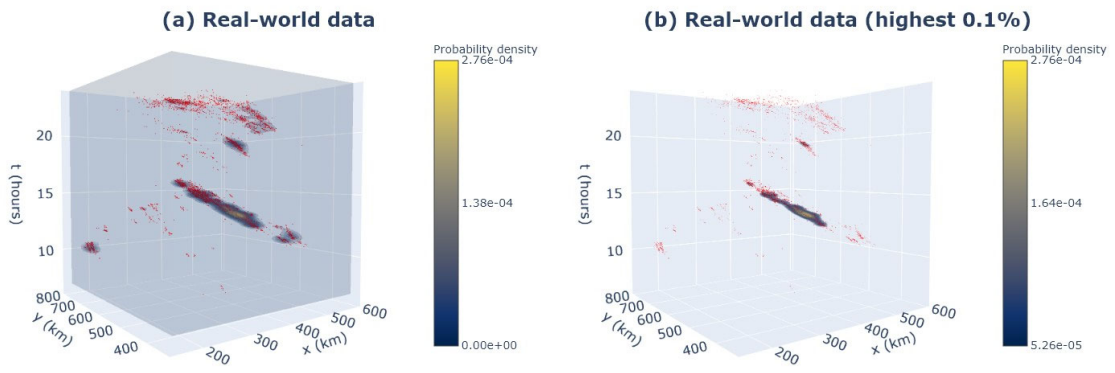

Figure 4.19. Volume rendering of **(a)** all and **(b)** only highest 0.1% of the probability density values as returned by the kernel density estimation methodology for the real-world lightning strike dataset (**Section 3.2.1**). Red points indicate the corresponding lightning strikes as point events. The probability density value range is given by the colour scale. Note the variations in the colour scale limits between both subplots, while the type of colour scale remains the same.

The space-time voxels that were identified as having the highest 0.1% of probability density values as returned by the application of the kernel density estimation were separated. The coordinates of these voxels are then used as the centroid locations for the elliptic space-time scan statistic (**Section 4.4.3**) as per the spatio-temporal clustering methodology application framework (**Section 2.5**).

4.4.3. Space-time scan statistic

The space-time scan statistic results for the point events simulation study (**Section 4.2.3**) established how the elliptic space-time scan statistic performs on various reference models and within the common methodology application framework (**Section 2.5**). The input parameters necessary to implement the space-time scan statistic on the real-world lightning strike dataset were assessed based on the K-function analysis results for spatial and temporal ranges of space-time interaction (**Section 4.4.1**) and the kernel density estimation results (**Section 4.4.2**) with a set probability density value threshold to identify particular space-time volumes. Based on these results, the necessary input parameters for the elliptic space-time scan statistic and the sets of their values selected are given in **Table 4.8**.

Table 4.8. Input parameters and their values for space-time scan statistic on the real-world lightning strike dataset.

Input parameter	Values
Centroid locations	Voxels from **Figure 4.19b**
Length of semi-major axis	{25, 50, …, 150} km
Shape of elliptic base	{1, 1.5, …, 4}
Angle	$\{0, (1/6)\,\pi, \ldots, (5/6)\,\pi\}$
Height	{0.5, 1, …, 4.5} h

From the given input parameter values in **Table 4.8**, permutations of the values are generated and kept constant for implementation of the elliptic space-time scan statistic. Various angles of the spatio-temporal domain are again used for visualising the clustering results of the space-time scan statistic. All point events within the identified scanning volume are considered part of the cluster. The elliptic space-time scan statistic results for the real-world lightning strike dataset are shown in **Figure 4.20** and the input parameters and their values for defining the identified elliptic cylinder encompassing the lightning strike cluster are listed in **Table 4.9**.

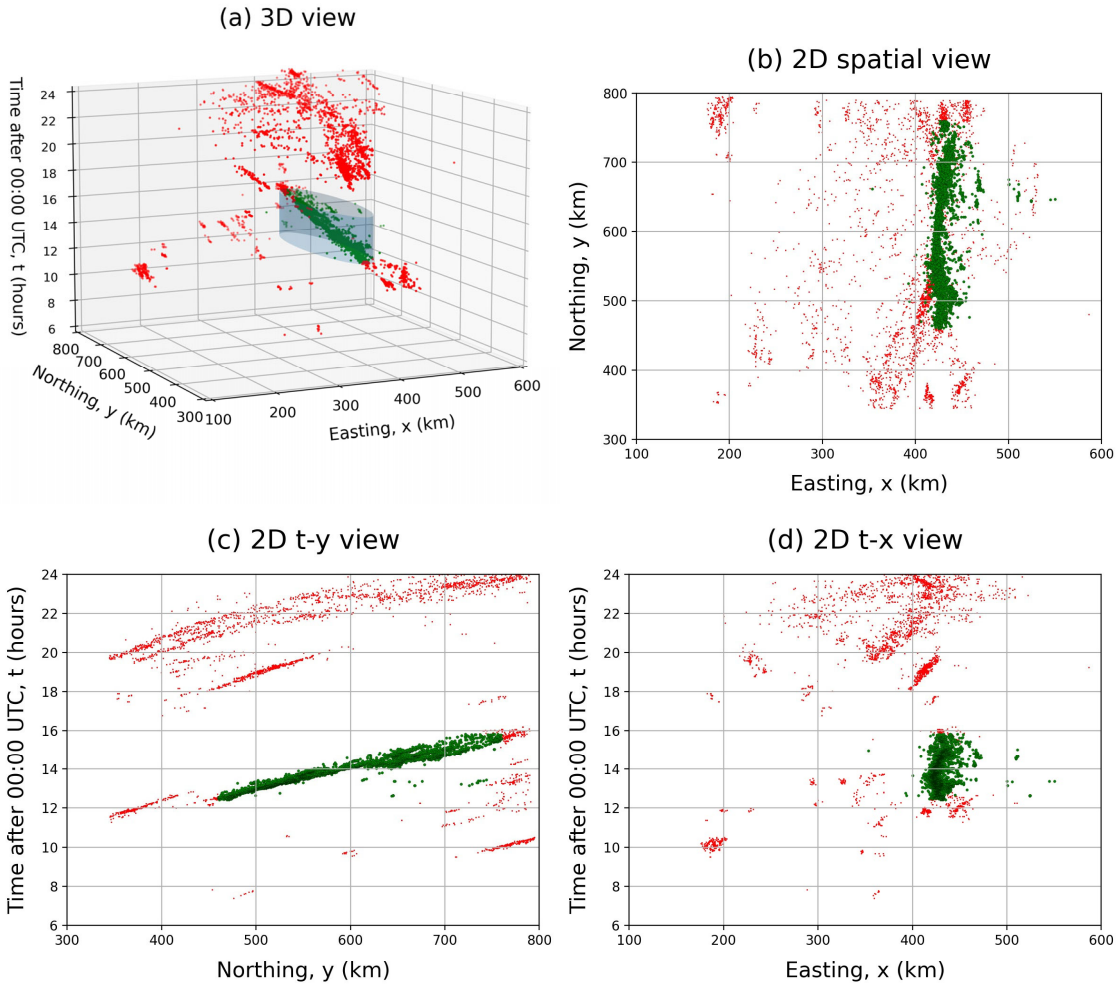

Figure 4.20. Various angles [**(a)** 3D view of the spatio-temporal domain, **(b)** 2D spatial view, **(c)** 2D *t–y* view, **(d)** 2D *t–x* view] of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to the real-world lightning strike data. In all subfigures, the events contained within the scan statistic are coloured green, while those outside in red. In **(a)**, the elliptic scan statistic is drawn in transparent light blue.

Table 4.9. Identified parameter values for elliptic space-time scan statistic on the real-world lightning strike dataset.

Input parameter	Identified values
Centroid location	{430, 610, 14.1}
Length of the semi-major axis	150 km
Shape of elliptic base	4
Angle	$(1/2)\,\pi$
Height	4 h

In **Figure 4.20**, the elliptic scan statistic has identified a cluster of lightning strike locations coinciding with the approximate initiation, duration, and termination of the Storm 2A of the second case study (**Section 3.3.2.3**). The elliptic cylinder as the scanning volume has managed to capture the northwards propagation direction (**Figures 3.10** and **3.11**) of the Storm 2A as the angle parameter is found to be (1/2) π relative to the first spatial dimension (i.e. easting) and the shape of the elliptic base is 4 (**Table 4.9**). The elliptic shape clearly encompasses the spatial spread of lightning strikes along the northing direction (**Figure 3.11**).

From the synoptic situation analysis for the three-storm system in the second case study [**Lewis and Silkstone, 2017**], the High Moorsley Doppler radar data snapshot at 14:08 UTC on 1 July 2015 (**Figure 3.11b**) sampled this storm through its most active phase, indicating that the identified centroid at $t = 14.1$ h is consistent with expectations based on the available auxiliary information. The elliptic scan statistic therefore identified thunderstorm with the highest density of lightning strikes as this space-time volume was already indicated as containing the highest density of lightning strikes from the kernel density estimation results (**Section 4.4.2** and **Figure 4.19b**).

4.4.4. Evaluation of limitations of the selected methodologies

In **Sections 4.4.1**–**4.4.3**, the spatio-temporal clustering methodology application framework (**Section 2.5**) was applied to the real-world lightning strike dataset (**Section 3.2.1**). This dataset was already used as the second case study of increased lightning activity in **Section 3.3.2.3** with characterisation of individual storms in **Section 3.3.4**. This allows one to analyse the clustering results with an existing understanding of the individual storm dimensions, assigned number of strikes within each and physical variable value ranges that would have an expected effect on the clustering results.

The selected application framework returned global and local clustering measure results consistent with the expectations based on the characterised dataset. The K-function analysis results (**Figure 4.18**) show spatial and temporal range values of $\{s_{MAX} = 320$ km, $\tau_{MAX} = 4.4$ h$\}$ for the maximum clustering measure D value. These values are consistent with the spatial extent of the individual storms {2A, 2B, 2C} in **Figure 3.12** and **Appendix Figure I1**. As the three individual storms in the second case study are temporally well separated (**Figure 3.10**), Storm 2A with $n = 4{,}019$ out of total $N = 12{,}521$ lightning strikes and with significantly smaller

average and median inter-event time values (**Figure 3.14**) affects the global clustering results more significantly than Storms 2B and 2C.

Similar to the K-function analysis results, the kernel density estimation indicates space-time volume within the Storm 2A as the volume with the highest probability density values (**Figure 4.19**). When considering the highest 0.1% value threshold (**Figure 4.19b**), the high point-event density volumes within Storm 2A are visually much denser than for Storms 2B and 2C, and any unidentified smaller clusters of lightning strikes. As in the results for the point event simulation study (**Section 4.2**), the use of K-function estimated spatial and temporal ranges of maximum space-time interaction $\{s_{MAX}, \tau_{MAX}\}$ as optimal bandwidth would not have identified smaller high-density volumes within Storm 2A. The leave-one-out estimator for the optimal bandwidth values returned values smaller than $\{s_{MAX}, \tau_{MAX}\}$.

After evaluation, the elliptic space-time scan statistic results for the point event simulation study (**Section 4.2.4**), the halving of the $\{s_{MAX}, \tau_{MAX}\}$ values and use of them as the maximum spatial and temporal extents of the elliptic cylinder returned an individual cluster with most of Storm 2A events contained within it. The maxima of both ranges were still identified as the input parameter values returning the most point-event dense space-time volume. The three storms within the real-world lightning strike dataset were also temporally well-separated (**Table 3.2**) relative to Models 4 and its variations, 4.1, 4.2, 4.3. As the scanning volume expanded, there were no other storms from which to assign the events to a cluster. The performance and ability of the elliptic scan statistic to separate clusters within a point event dataset produced by multiple moving point sources is dependent on the separation of the sources either spatially or temporally, clearly indicating a limitation of the current geometry and input parameters of the scanning volume (addressed in **Chapter 5**).

4.5. Results and analysis for real-world trajectory datasets

For the trajectory data type, one particular category of clustering methodologies was presented in **Section 2.2.2** and selected for implementation: density-based methodologies. Density-based methodologies establish high trajectory-density neighbourhoods around trajectories to separate inter-class trajectories from noise. The Trajectory-OPTICS algorithm was implemented from this methodology category due to its extension for the trajectory data type by using the spatio-temporal distance as a distance function.

Section 3.4.2 presented two synthetic benchmark datasets representing the homogeneous and clustered trajectory datasets. The results of applying the Trajectory-OPTICS methodology to the synthetic datasets in order to establish performance benchmarks are presented in **Section 4.3**. For the real-world dataset of hurricane tracks (presented in **Section 3.2.2**), the results of applying the Trajectory-OPTICS methodology are analysed in this section.

To apply the Trajectory-OPTICS clustering methodology with the spatio-temporal distance as defined in **Appendix Equation E.1**, the real-world hurricane tracks need to be aligned and of the same temporal length. These two requirements were intrinsically included in the two synthetic trajectory datasets as they had time-steps of equal length, and all trajectories were initiated at time $t = 0$ and allowed to propagate until reaching $t = T = 20$.

For the real-world hurricane track data, the time-steps are of equal temporal length of 6 hours but are not aligned and do not have the same total duration. Aligning the initial positions of all trajectories at $t = 0$ would allow the use of the clustering results to better understand how their physical properties (e.g. propagation speed, direction, distribution of points around the point of origin) develop over the same timescales (i.e. over their total duration). Once spatio-temporal models with parameters representing physical properties are introduced, their clustering results will be comparable to the patterns observed in real-world data and interpretable in the context of what properties, in particular, govern the creation of such clustering patterns.

When the hurricane trajectories are aligned at time $t = 0$, the Trajectory-OPTICS methodology with the selected spatio-temporal distance function further necessitates that they have to be of the same total duration. The choice for the particular duration value is a trade-off between the completeness of the trajectory tracks and the number of trajectories left in the dataset. Selecting a larger duration value allows for retaining more information about the trajectories, but there are fewer trajectories that would have persisted for that larger duration value. The performance of the clustering methodologies on real-world data is assessed generally, without an imposed focus on the particularities (e.g. selecting trajectories corresponding to time intervals over which they made landfall).

All trajectories are aligned at their points of origin at $t = 0$ and only their first $n_t = 30$ time steps are retained. As the time step value for this dataset is 6 hours, all trajectories are of equal

duration T = 7.5 d. The resultant real-world trajectory dataset consisting of N = 136 total trajectories is illustrated in **Figure 4.21**.

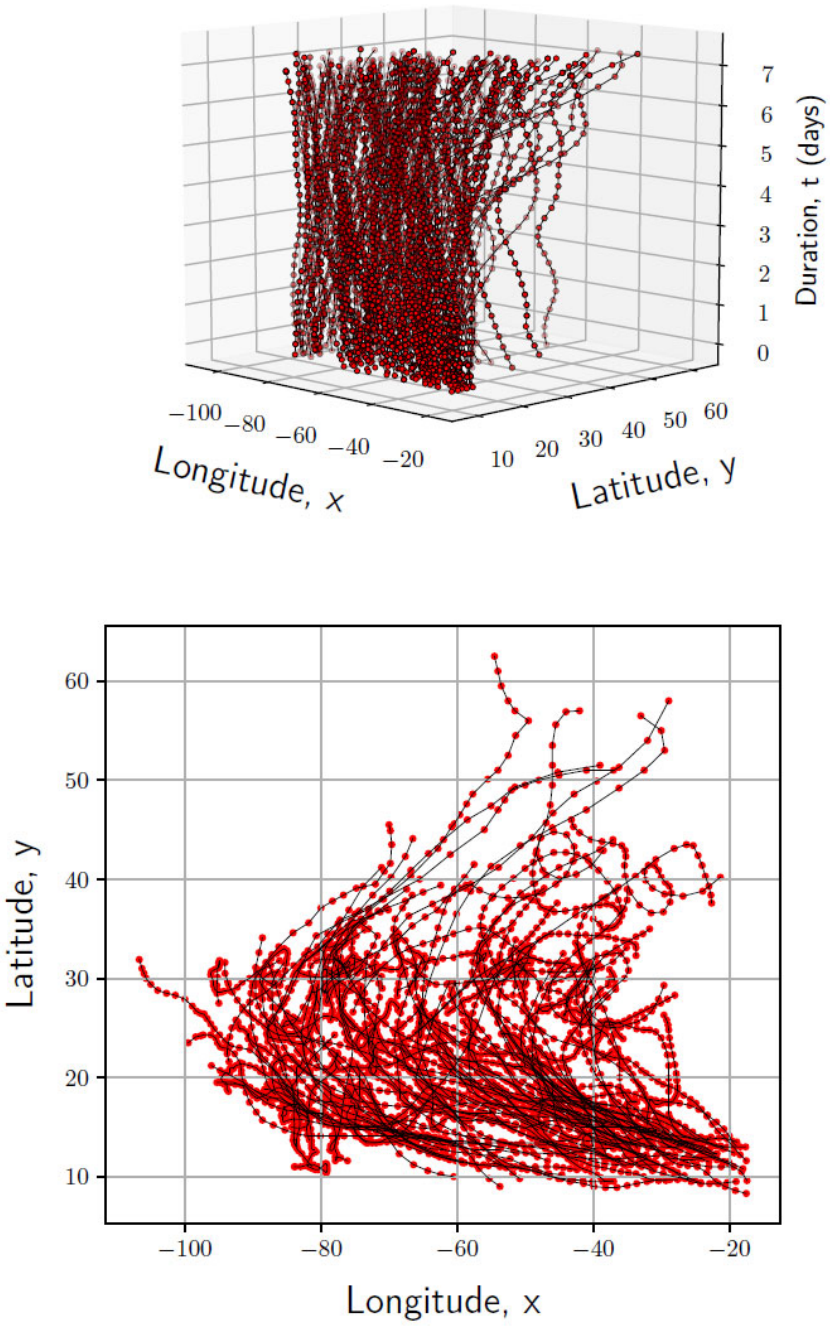

Figure 4.21. Real-world hurricane trajectory dataset after alignment of all trajectories at t = 0 and application of a duration threshold of n_t = 30, which corresponds to T = 7.5 d. Upper subfigure shows the propagation of the trajectories in the space-time rectangular cuboid, while lower subfigure illustrates the propagation of trajectories once they are collapsed onto the two spatial dimensions. There are a total of N = 136 trajectories.

After the alignment and application of a duration threshold of $n_t = 30$ time steps, the Trajectory-OPTICS methodology is applied to the modified dataset with the same application steps as established for the synthetic benchmarks (**Section 4.3**). The results after applying all steps of the methodology and the resulting identified clusters are presented in **Figure 4.22**.

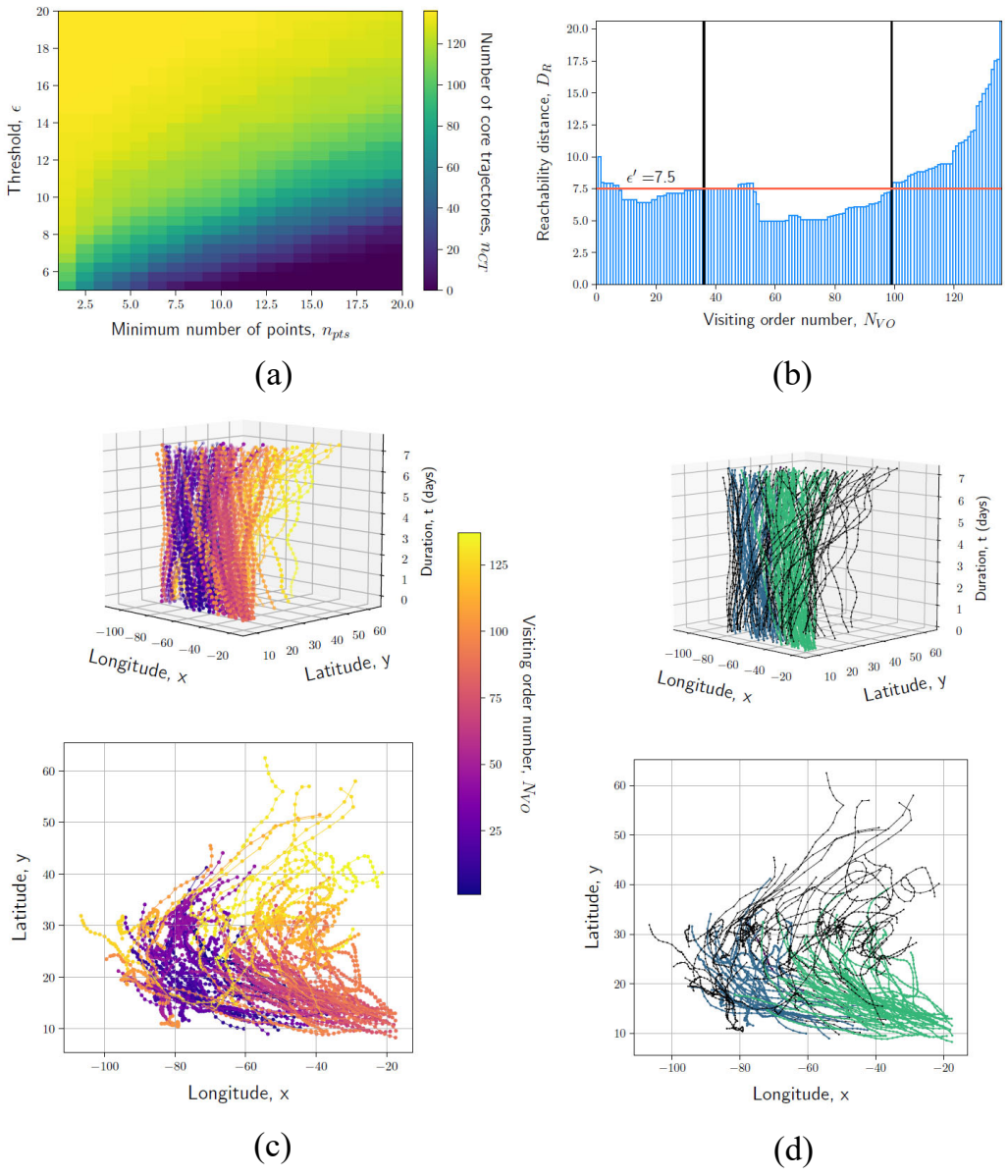

Figure 4.22. Clustering results of the $N = 136$ real-world hurricane trajectories with alignment at $t = 0$ and duration cut-off of $T = 7.5$ d. **(a)** The number of core trajectories n_{CT} for various neighbourhood distance threshold ε and minimum number of trajectories within given threshold n_{pts} combinations. A combination of input parameters $\varepsilon = 10$, $n_{pts} = 10$ was selected. **(b)** Reachability distance D_R plot with the reachability distance threshold E' indicated by a red line and its value given above it. Black vertical lines indicate first trajectories for which $D_R \geq \varepsilon'$. **(c)** Visiting order of the trajectories during the execution of the Trajectory-OPTICS methodology. Visiting order number N_{VO} is indicated by the colour scale. **(d)** Clustering results with each cluster assigned a colour, while all noise trajectories ($D_R \geq \varepsilon'$) are coloured black. In **(c)** and **(d)**, upper subfigures show the trajectory propagation in the space-time rectangular cuboid, while lower subfigures illustrate the propagation onto the two spatial dimensions.

Each subplot in **Figure 4.22** can be directly compared to the benchmarks for the results of the Trajectory-OPTICS methodology when applied to homogeneous or clustered trajectories established in **Section 4.3**. For selecting the two input parameters (**Figure 4.22a**), a set of values was chosen to have a similar proportion of all trajectories be assigned as 'core' as for the analysed synthetic trajectory datasets. A set of parameter values $\{\varepsilon = 10, n_{pts} = 10\}$ was selected and used as input for the Trajectory- OPTICS methodology to generate the reachability distance plot (**Figure 4.22b**).

The reachability distance plot shows two intervals of D_R values. While not as pronounced as for the clustered trajectories (**Figure 4.16b**), these two intervals represent two trajectory clusters as a reachability distance threshold $\varepsilon' = 7.5$ is set from visual inspection. From relating the visiting order of trajectories (**Figure 4.22c**) with the reachability distance plot (**Figure 4.22b**), the trajectories that form the two clusters have similar physical properties between themselves (e.g. points of origin, propagation speed and direction). Noise trajectories ($D_R \geq \varepsilon'$ and $N_{VO} \geq 100$) have varying points of origin, propagate in varying directions and speeds. Trajectories, which correspond to the two trajectory clusters identified as having $D_R < \varepsilon'$, confirm the visual observations from the visiting order plot. The two trajectory clusters have similar propagation directions and speeds, but varying points of origin (a difference of approximately 25° longitude).

The Trajectory-OPTICS methodology as described in **Section 2.2.2.1** was used to establish performance benchmarks on synthetic homogeneous and clustered trajectory datasets (**Section 3.4.2**). When applied to real-world hurricane track data (**Section 3.2.2**), the methodology returned two distinct clusters with varying points of origin. While having a visual distinction and a potential physical property of the hazardous process that would explain it, these clusters were found using a particular set of input parameters, applying alignment and duration cut-off to the trajectory dataset. These three factors need to be evaluated and their effect understood.

First, **input parameters** – the choice of input parameters govern which and how many of the trajectories are assigned as being 'core' trajectories. More core trajectories imply more points of reference for measuring the reachability distance. The set of parameter values used in applying Trajectory-OPTICS in **Section 4.3** was $\{\varepsilon = 1, n_{pts} = 10\}$. To evaluate how the changes in the two input parameters affect the reachability distance plots, four sets of input parameter values are selected for the synthetic clustered trajectory dataset (**Section 3.4.2**): $\{\varepsilon = 0.75, n_{pts}$

= 10}, {$\varepsilon = 1.25$, $n_{pts} = 10$}, {$\varepsilon = 1$, $n_{pts} = 5$} and {$\varepsilon = 1$, $n_{pts} = 20$}. The resulting changes in the reachability distance plots are shown in **Figure 3.23**.

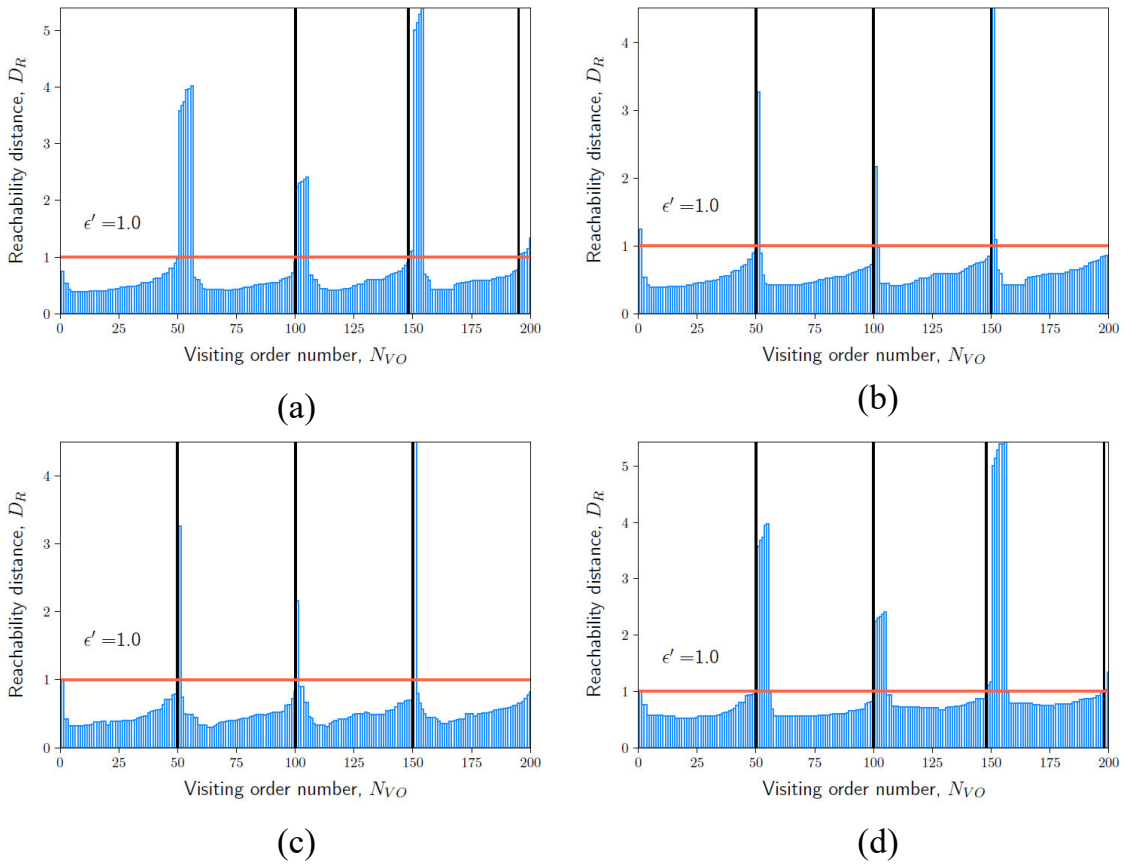

Figure 4.23. Reachability distance D_R plots for various sets of input parameters ε and n_{pts}: **(a)** {$\varepsilon = 0.75$, $n_{pts} = 10$}, **(b)** {$\varepsilon = 1.25$, $n_{pts} = 10$}, **(c)** {$\varepsilon = 1$, $n_{pts} = 5$} and **(d)** {$\varepsilon = 1$, $n_{pts} = 20$}. The reachability distance threshold $\varepsilon' = 1.0$ is kept constant and indicated by a red line. Black vertical lines indicate first trajectories for which $D_R \geq \varepsilon'$.

When compared to the set of input parameters chosen for implementation {$\varepsilon = 1$, n_{pts}=10} (**Figure 4.16b**), changing the input parameters controls the noise tolerance in the clustering results. When lowering ε (**Figure 4.23a**), fewer trajectories are assigned as core (evident in **Figure 4.15b**), the peaks in D_R values are not as pronounced and more trajectories are left unclustered. The opposite is observed when increasing ε (**Figure 4.23b**), where the peaks in D_R are more pronounced and fewer trajectories are unclustered relative to **Figure 4.16b**. Similarly evident from **Figure 4.15b** is the increase in the number of core trajectories n_{CT} with increasing n_{pts} and vice versa. For reachability distance this influences the noise tolerance as well as smaller n_{pts} values (**Figure 4.23c**) create more pronounced peaks in D_R with fewer trajectories left unclustered and vice versa (**Figure 4.23d**). Therefore, as evident in **Figure**

4.23, the choice of input parameters controls the noise tolerance in the clustering results, but all four clusters of synthetic trajectories can be identified.

Second, **alignment of the trajectories**: evaluation of the alignment of all trajectories at $t = 0$ is based on the objective of the clustering task. Without alignment, trajectories are inherently clustered in the temporal domain based on the hurricane season. To cluster trajectories based on their physical properties, a first-observation alignment [**Chudova *et al.*, 2003**; **Blender *et al.*, 1997**] is implemented to align the observed points of origin of the trajectories. Alternative approaches to the alignment of trajectories have not been considered.

Third, **the choice of the duration cut-off**: the choice of duration cut-off in terms of the number of time steps n_t is a trade-off between the number of trajectories in the dataset and the completeness of information about the propagation of trajectories (i.e. large threshold will retain most of the trajectories, but only a few hurricanes will have persisted). The selected duration threshold value was $n_t = 30$ or $T = 7.5$ d. To evaluate how the changes in the duration threshold affect the reachability distance plots, three additional values are selected for the real-world hurricane dataset (**Section 3.2.2**) $\{n_t = 20, n_t = 25, n_t = 35\}$. The resulting reachability distance plots and clustering results are shown in **Figure 4.24**.

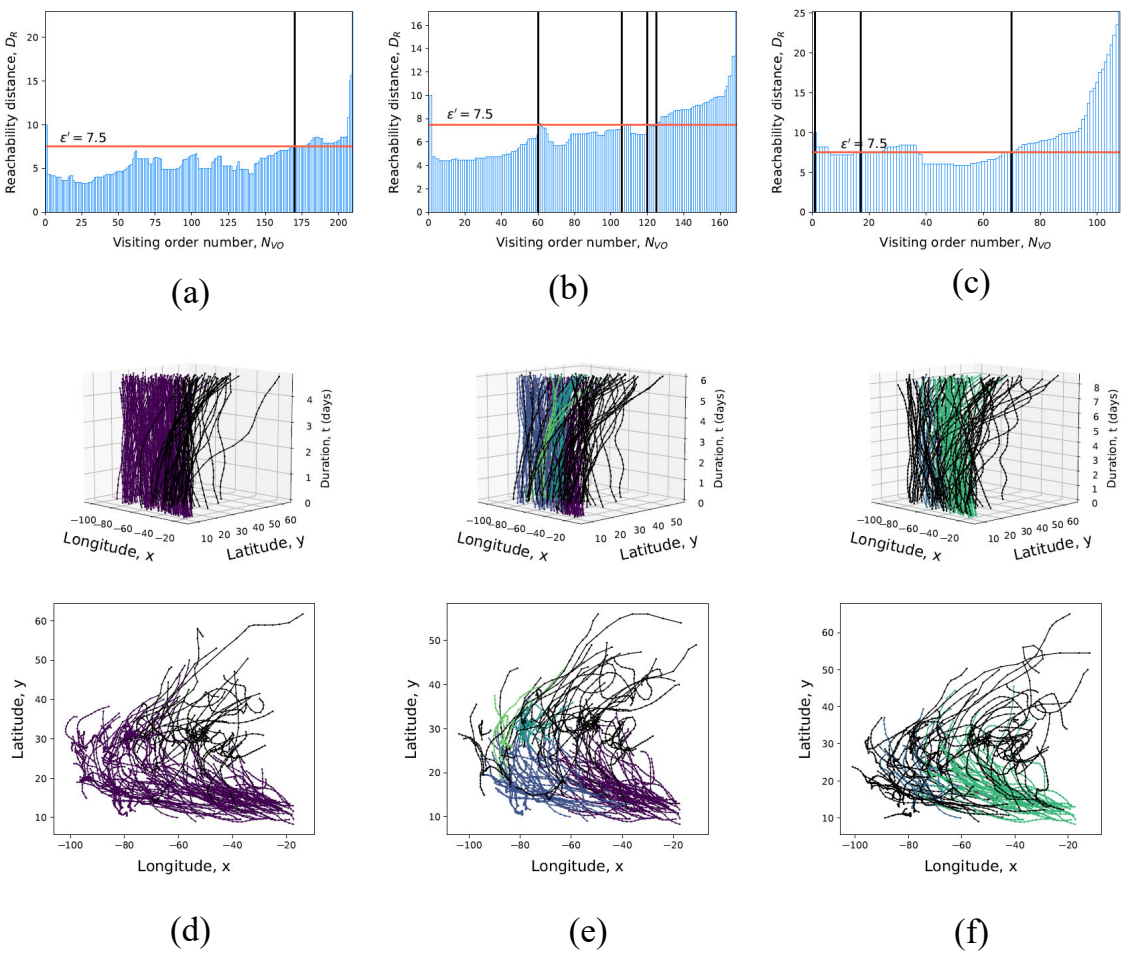

Figure 4.24. Reachability distance D_R plots [**(a)**, **(b)**, **(c)**] and clustering results [**(d)**, **(e)**, **(f)**] for various trajectory duration threshold as imposed by the number of selected time steps n_t: [**(a)**, **(d)**] $n_t = 20$, [**(b)**, **(e)**] $n_t = 25$, [**(c)**, **(f)**] $n_t = 35$. For all three threshold values, the combination of input parameters $\varepsilon = 10$, $n_{pts} = 10$ was selected. A threshold on the reachability distance value was set at $\varepsilon' = 7.5$ from visual inspection and is indicated by a red line. For clustering results [**(d)**, **(e)**, **(f)**], each identified cluster is assigned a unique colour, while all noise trajectories ($D_R \geq \varepsilon'$) are coloured in black. Upper subfigures show the propagation of the trajectories in the space-time rectangular cuboid, while lower subfigures illustrate the propagation of the trajectories once they are collapsed onto the two spatial dimensions.

When compared to the reachability distance plot and clustering results for $n_t = 30$ (**Figure 4.22**), changes in duration thresholds directly affect the number of compositions of the clusters identified by the Trajectory–OPTICS methodology. For $n_t = 20$ (**Figures 4.24a** and **4.24d**), for the same input parameters and reachability threshold value, separation into two distinct clusters is no longer possible. There are more than 200 trajectories in the dataset and with a small reachability threshold value, several small clusters could be identified. If $n_t = 25$ (**Figures 4.24b** and **4.24e**), the previous reachability distance threshold value returns four clusters of varying sizes with two of them visually similar to the ones identified in the clustering results for $n_t =$

30 (**Figure 4.22d**). Increasing the duration threshold to n_t = 35 with the same reachability distance threshold value returns results similar to **Figure 4.22d** as two clusters with varying points of trajectory origins are identified. Therefore, the choice of the trajectory duration threshold in terms of the number of time steps n_t affects the clustering results directly. Changing the threshold value controls the size of the final dataset, which consequently requires selection of input parameters and reachability distance threshold values depending on the particular data subset. Without adjustment of parameters, the clustering results are sensitive to the selected duration threshold.

A density-based spatio-temporal clustering methodology Trajectory-OPTICS (**Section 2.2.2.1**) was applied to homogeneous and clustered synthetic trajectory datasets (**Section 3.4.2**) and their results were used as benchmarks to assess how the methodology would perform on a real-world hurricane dataset (**Section 3.2.2**). The methodology uses a set of core trajectories (i.e., trajectories with at least n_{pts} other trajectories within a spatio-temporal distance ε of them) as reference trajectories for measuring the separation (i.e., reachability distance) to any given trajectory. By visiting each trajectory only once, the methodology identifies trajectory intervals that have low reachability distance values. If these values are below a threshold *ε'*, they are classified into a cluster and any trajectories with reachability distance above the threshold are noise trajectories and separate subsequent clusters.

Synthetic dataset benchmarks (**Figure 4.16**) for the reachability distance plots exhibited two contrasting behaviours. For clustered trajectories, four intervals of low reachability distance values were confirmed to represent the four generated clusters (**Figure 4.17b**). For homogeneous trajectories, there were no clear persistent intervals of low reachability distance values and visual selection of a suitable ε' values was not trivial.

For the real-world hurricane trajectory dataset, alignment and duration thresholds were introduced. Alignment of all trajectory points of origins at time t = 0 was justified as a necessity for completion of an objective of this thesis as it allows to extract physical properties of the hazardous process. The final dataset consisted of 136 trajectories that persisted for a total duration of T = 7.5 d. Trajectory-OPTICS methodology identified two distinct clusters with similar propagation speeds and directions, but varying areas of origin.

Sensitivity of the clustering results to the three parameters of the methodology $\{\varepsilon, n_{pts}, \varepsilon'\}$ was evaluated and the selected values justified. The further discussion expanded on the alignment and duration thresholds for the real-world trajectory dataset. The clustering results in terms of the number of identified clusters and the number of events within each cluster were sensitive to the selected duration threshold and the set of input parameters had to be reassessed for each selection. Ways of addressing this are identified as further research direction in **Section 6.3.3**.

4.6. Chapter conclusions

In this research chapter, the spatio-temporal clustering application and analysis framework (**Section 2.5**) was applied to the real-world datasets and the simulation study (Model 1 and Model 2 as synthetic benchmarks and Model 3 and Model 4 as synthetic models representative of single and three thunderstorm systems) as presented in **Sections 3.2** and **3.4** respectively.

The research on a common spatio-temporal application and analysis framework by using global and local clustering measures for natural hazard data in [**Tonini *et al.*, 2022**; **Wang *et al.*, 2021**; **Tonini and Cama, 2019**; **Tonini *et al.*, 2017**] was discussed and extended in **Chapter 2** and applied to synthetic and lightning strike datasets in this chapter. No previous global and local clustering measures framework has been applied to lightning strike point event data in the spatio-temporal domain. Only the spatial domain has been considered in [**Hernandez-Magallanes and Genton, 2019**; **Berdahl, 2016**]. Existing lightning-related work with global clustering measures in the spatio-temporal domain includes lightning-caused wildfires [**Aftergood and Flannigan, 2022**; **Zhang *et al.*, 2013**; **Wang and Anderson, 2010**].

Within the methodology application framework, the K-function results for both simulation study and real-world data indicate the ambiguity in working with scan statistics. Instead of limiting the scanning volume extent based on the study volume [**Tonini *et al.*, 2022**; **Parente *et al.*, 2016**; **McNally and Colver, 2008**; **Kulldorff *et al.*, 2006**], this framework proposed to use the K-function estimated spatial and temporal ranges of maximum space-time interaction. The scan statistic geometry and set of input parameters did not identify individual point events clusters of individual storms for Model 4 (and variations 4.1, 4.2, 4.3), identifying simply the most point-event dense space-time volume. Further evaluation and improvements to the K-function analysis and the space-time scan statistic geometry are presented in **Chapter 5**.

5. Improvements on the spatio-temporal clustering methodologies for application to point events produced by a moving source

Chapter summary

This research chapter presents and implements improvements to the spatio-temporal clustering methodology framework (**Section 2.5**) as applied and evaluated in **Chapter 4**. The results from **Chapter 4** are used to identify reasons behind performance shortcomings of the implemented methodologies. These improvements aim to modify the existing spatio-temporal clustering methodologies as part of the framework to improve their ability and/or efficiency in identifying clusters of point events produced by a moving source. This comprises of novel research on extensions of the existing and reviewed spatio-temporal clustering methodologies to account for the application to point event data produced by a moving source. The main improvement is made on the space-time scan statistic as the local clustering measure. In **Chapter 4**, the methodology could not account for the source movement in the space-time volume and when applied to moving source datasets with low temporal separation between individual clusters. An inclined variation of the scanning volume is proposed with additional input parameters providing physical interpretation to the clustering results. These and other changes in the implementation of the methodologies are highlighted and the expected improvements in their performance from these changes are discussed. For point events, including the simulation study datasets from **Section 3.4.1** allows for continued assessment and evaluation of the performance of the methodologies depending on the physical variables of the lightning strikes and the thunderstorm system producing them. The improved methodologies and the spatio-temporal clustering framework is again applied to the real-world lightning strikes dataset (**Section 3.2.1**) to evaluate the performance improvements.

This chapter aims to answer the following questions:

Q5.1. How can the performance shortcomings of the already implemented methodologies be addressed in the context of points events produced by a moving source?

Q5.2. Are improvements on these methodologies only in terms of the clustering accuracy or also in computational complexity, adjustments to the general methodology framework?

Q5.3. Can these improvements on the methodologies bring more detailed characterisation of the physical processes and variables?

5.1. Introduction

In this chapter, the improvements to the spatio-temporal clustering methodologies for the point event data type presented in **Section 2.3.1** and their common application framework (**Section 2.5**) are presented and applied to the synthetic benchmark datasets as part of the simulation study presented in **Section 3.4** and the real-world datasets presented in **Section 3.2**. From the clustering methodologies relevant to the point event data type, improvements to the global clustering measure (K-function analysis) and the local clustering measure (space-time scan statistic) are presented and discussed in **Section 5.2** and **Section 5.3,** respectively.

After evaluating the methodology performance on the simulation study and real-world lightning strike dataset in **Chapter 4**, these improvements are implemented. The clustering results for the synthetic datasets of the simulation study are presented in **Section 5.4**, while the results for the real-world lightning strike dataset follow in **Section 5.5**. The clustering results are presented visually and analyses on what are their implications for the individual methodologies and the general clustering methodology analysis framework (**Section 2.5**) are discussed within each section. The clustering results are compared with previous results in **Chapter 4** and their implication for the individual clustering methodologies and the overall methodology framework are evaluated in **Section 5.6**.

5.2. Improvements to the K-function analysis

From the results of the clustering measure D (**Section 4.2.1** and **Section 4.4.1**), the dependence of this methodology on cumulative results is evident. As the spatial and temporal range values $\{s,\tau\}$ are extended, the K-functions count all point events contained within these range values. In terms of variations of the global clustering measure, in [**Diggle *et al.*, 1995**] various additional clustering measures and extensions based on the clustering measure D (**Equation 2.1**) were presented. As example, the proportional increase in the number of further cases which would be expected for the given spatial (s) and temporal (τ) ranges due to space-time interaction, denoted as clustering measure $\widehat{D}$, can be expressed as

$$\widehat{D}(s,\tau) = \frac{D(s,\tau)}{K_S(s)K_T(\tau)} = \frac{K_{ST}(s,\tau)}{K_S(s)K_T(\tau)} - 1 \qquad (5.1)$$

In **Equation 5.1**, positive values have identical interpretation to clustering measure *D*, while in absolute values, clustering measure $\widehat{D}$ indicates the order of magnitude by which the spatio-temporal K-function $K_{ST}(s,\tau)$ is larger than the product of the individual spatial and temporal K-functions $K_S(s)K_T(\tau)$. In this thesis, the focus was on identifying improvements on the K-function expressions directly by considering various aspects of the dataset.

When evaluation the K-function analysis results for the simulation study and the real-world data in **Sections 4.2.4** and **4.4.4**, respectively, two main observations were made about their performance and value within the methodology framework. First, using the spatial and temporal ranges $\{s_{MAX}, \tau_{MAX}\}$ for which the clustering measure *D* reached maximum was not efficient for estimating the optimal bandwidths for the kernel density estimation (**Table 4.1** and **Section 4.2.2**). Second, the use of these ranges was not immediately applicable to the spatial and temporal extents of the space-time scan statistic scanning volume dimensions (**Section 4.2.3**). The scan statistic used the maximum extents provided by the set of input parameters and their values without imposed limitations from the study area and duration [**Tonini *et al.*, 2022**; **Parente *et al.*, 2016**; **McNally and Colver, 2008**; **Kulldorff *et al.*, 2006**].

Differentiating the K-function results for small variations in spatial and temporal range values is not obvious. While satisfactory for a global clustering measure aimed at initial assessment of space-time interaction, to compare the clustering measure *D* results for various spatial and temporal scales, initial clustering measure *D* results can be standardised by the volume of the space-time search volume defined by the spatial and temporal range values. The expressions for the individual K-functions (**Appendix Equations B.4–B.6**) can be redefined as

$$K'_{ST}(s,\tau) = \frac{1}{2\tau\pi s^2} K_{ST}(s,\tau) \tag{5.2}$$

$$K'_{S}(s) = \frac{1}{\pi s^2} K_{S}(s) \tag{5.3}$$

$$K'_{T}(\tau) = \frac{1}{2\tau} K_{T}(\tau) \tag{5.4}$$

$$D'(s,\tau) = K'_{ST}(s,\tau) - K'_{S}(s)K'_{T}(\tau) \tag{5.5}$$

This standardisation of the individual K-functions by the space-time volume that is used to search for additional point events around an arbitrary one reduces the significance of detecting additional smaller clusters of point events at greater range values. An intuitive understanding of this significance reduction is provided in discussion about **Figure 5.2b** in **Section 5.4**, where

Model 3 K-function results illustrate how the additional point events from a different 'parent' cluster no longer contribute to space-time interaction.

5.3. Improvements to the space-time scan statistic

Section 4.2.3 applied the elliptic space-time scan statistic to synthetic single and three storm models. For single storm Models 3 and its variations, the scan-statistic results were dependant on the selected maximum spatial and temporal extents of the scanning volume. All models returned the maximum possible extent values as the identified dimensions of the elliptic cylinder. Without imposing limitations, the scan statistic would identify the storm as one individual cluster. With limitations from K-function analysis (**Section 4.2.1**), the scan statistic selects a space-time volume within the storm with highest point event density.

For the three storm Model 4 and its variations, the elliptic space-time statistic did not return individual clusters consisting of point events within one single storm source. The identified space-time volumes as individual clusters consist of point events from multiple storms as the geometry of the elliptic cylinder as the scanning volume cannot account for the moving source producing each storm. The evaluation of the simulation study results followed in **Section 4.2.4**.

For the real-world lightning strike dataset (**Section 3.2.1**), the clustering results of the space-time scan statistic with the elliptic cylinder as the scanning volume identified a cluster of lightning strike locations corresponding to the Storm 2A from the second case study of increased lightning activity **(Section 3.3.2.3**). From the characterisation results of the individual storms 2A–2C (1 July 2015), the scan statistic identified the space-time volume with the highest expected density of lightning strike events and correctly specified its propagation direction (**Section 4.4.3**). During the evaluation (**Section 4.4.4**), it was noted that for this particular real-world data, the storms were well-separated temporally, not challenging the methodology as was the case for Model 4 and its variations, 4.1, 4.2 and 4.3.

Selection of its input parameters for the elliptic scan statistic was supported by the spatial and temporal ranges of space-time interaction as identified by the global clustering measure *D* derived from using K-functions (**Section 4.4.1**) and the particular space-time volume regions (i.e. voxels) were indicated by the kernel density estimation results with a threshold of only showing the voxels with highest 0.1% of the probability density values (**Section 4.4.2**). These

results were evaluated (**Section 4.4.4**) and two properties were identified as targets for improvement. For this future direction, two approaches can be specified.

First, **identify and assess various scanning volume geometries:** at the current implementation of the scan statistic, the scanning volume is an elliptic cylinder. The approach to assessing various scanning volumes can be further divided into two approaches:

1. **Modifications to the elliptic cylinder:** at the time of point event generation, the thunderstorms producing the lightning strikes are considered spatially fixed sources in the context of this thesis (**Section 3.3.5).** While extending the length of the semi-major axis captures most of the lightning strike locations generated by the second thunderstorm and the elliptic base identifies the propagation direction, an elliptic cylinder is not able to capture the propagation speed and spread of individual lightning strike events within the thunderstorm. As one objective of this chapter is to extract as many physical properties of the hazardous process as possible, various geometries that aid in the completion of this objective will be considered. Visualisations of some of the shape geometries that will be implemented are shown in **Figure 5.1**.

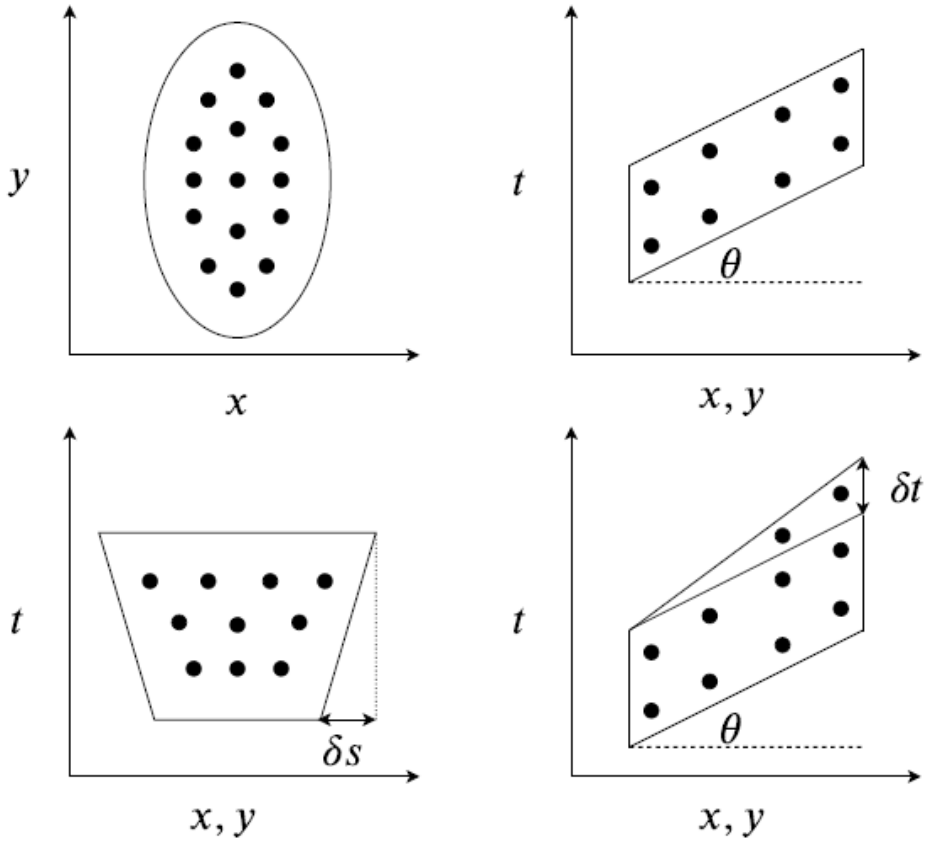

Figure 5.1. Various geometries of the space-time statistic for the lightning strike location (represented by arbitrary point events as black dots) dataset. Upper left subplot displays the current elliptic scan statistic in the 2D spatial domain. Upper right subplot displays the side-view in the space-time domain of the elliptic cylinder with an angle of inclination θ, capturing the direction and speed of propagation directly. Lower left subplot displays an isosceles trapezoid base scan statistic in the space-time view of the propagation direction, capturing the changes in spatial extent of the thunderstorm. Lower right subplot displays the side-view in the space-time domain of the isosceles trapezoid base scan statistic with changing height value, capturing the persistence of lightning strikes over various locations.

The elliptic scan statistic (upper left subfigure in **Figure 5.1**) captures the direction of propagation as it consists of the angle parameter relative to the first spatial dimension x. To account for a moving source producing the events, an inclination parameter θ is added to the scan statistic and the movement speed (if assumed constant as it was for the modelling framework in **Section 3.3.5**) can be captured directly. As additional extensions to the geometry of the scanning volume, to assess the spatial extent of the thunderstorm and its changes over time, an isosceles trapezoidal prism (i.e. two sides with right angles to the base) can be used as scanning volume (lower left subfigure in **Figure 5.1**). Parameter δs, as half the difference between the parallel sides of the trapezoid, estimates the changes in the spatial extent of the thunderstorm over time. The last dimension of the scanning volume

that can be adjusted is the height of it in the temporal domain (lower right subfigure in **Figure 5.1**). This height change by δt would assess the persistence of the thunderstorm over the particular area. The latter two extensions are not implemented in this thesis.

2. **Arbitrary shapes:** some of the scientific literature on arbitrarily-shaped scan statistics has already been reviewed at the critical literature review phase in **Section 2.2** (e.g. [**Assuncao *et al.*, 2006**; **Tango and Takahashi, 2005**; **Duczmal and Assuncao, 2004**; **Patil and Taillie, 2004**]), but have not been implemented as they take advantage of the auxiliary information available about the geographic regions and their connectivity, which is not within the objectives of this thesis. Further critical review of available literature in this research direction is left as a potential research direction.

Second, **modify the methodology for multiple cluster detection:** two different approaches on detecting multiple clusters with the scan statistic are identified and the results summarised as follows:

1. **Sequential method** [**Zhang *et al.*, 2010**]: the identified most likely cluster has an effect on calculating the likelihood ratio for the less likely clusters. The objective of this method is to eliminate this effect in order to test for the presence of additional clusters conditional on the presence of the previously detected ones. Once the initial cluster is identified, the point events contained within it and the encompassing study volume are removed from the dataset. The scan statistic is applied as before, but on a modified dataset with reduced number of events and study volume.
2. **Multiple cluster hypothesis method** [**Li *et al.*, 2011**]: an alternative hypothesis is defined, allowing multiple volumes of clustered events, which accounts for co-existing clusters in the detecting and evaluating process. Various combinations of the multiple scanning volumes have to be implemented, requiring higher level of computational complexity.

5.4. Results of improved clustering measures on synthetic datasets produced by the lightning strike models

As the first methodology within the spatio-temporal clustering framework (**Section 2.5**), the K-function analysis framework with the standardisation applied to individual K-function is applied to the datasets from the simulation study (**Section 3.4.1**). The individual K-functions

and the resulting clustering measure D are standardised by the search window volume as defined by the spatial range s and temporal range τ (**Equations 5.2–5.5**). The results for the standardised clustering measure D' and its components with the applied standardisation by search window volume over the same spatial and temporal range values for the homogeneous (Model 1) and clustered (Model 2) point event datasets are shown in **Figure 5.2**. These are the same synthetic dataset initialisations of the $N = 1{,}000$ point events placed in a $\{X = 20, Y = 20, T = 20\}$ space-time volume according to the point event generation algorithms as used to generate results for clustering measure D in **Figures 4.1** and **4.2**. The spatial and temporal ranges are set as $s = \{0.1, 0.2, ..., 5\}$ and $\tau = \{0.1, 0.2, ..., 5\}$.

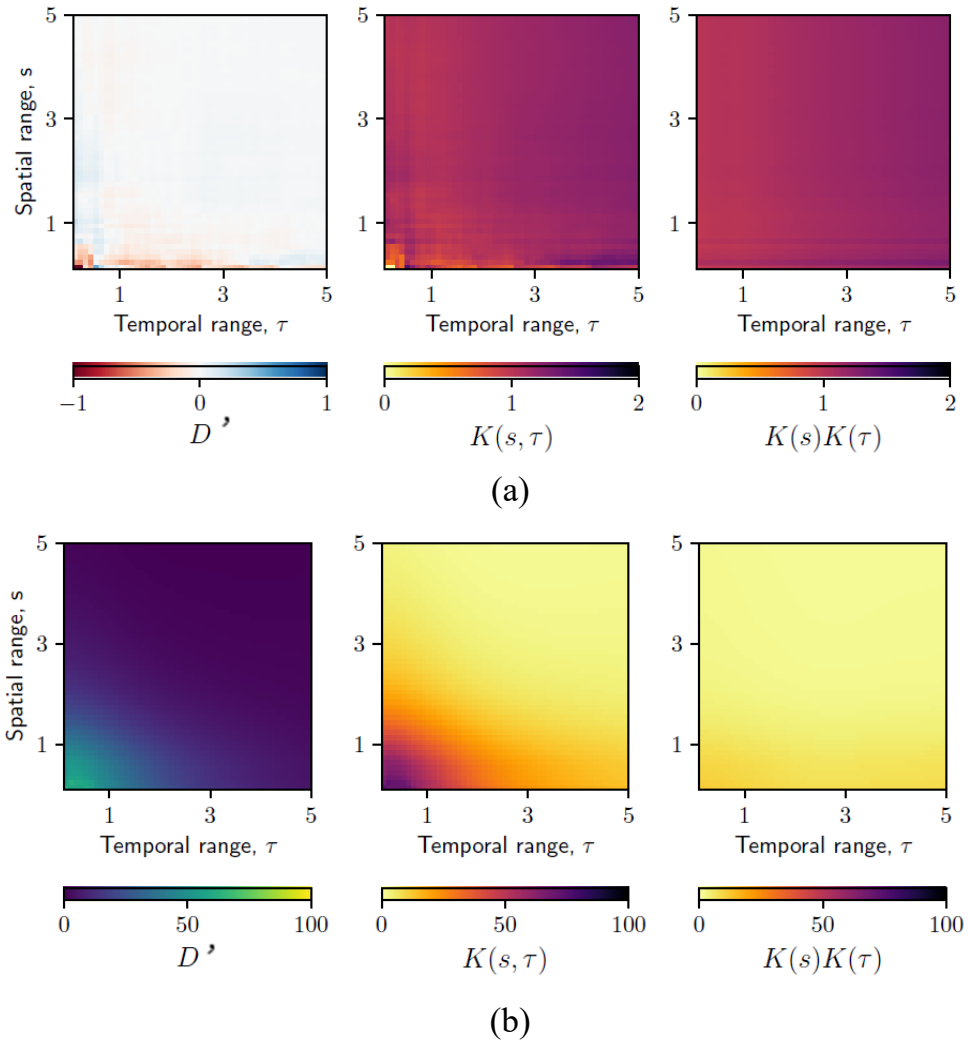

Figure 5.2. Three subplots of the standardised clustering measure $D' = K(s,\tau)'–K'(s)K'(\tau)$, spatio-temporal K-function $K'(s,\tau)$ and the product of spatial and temporal K-functions $K'(s)K'(\tau)$ as applied to **(a)** homogeneous (**Model 1**) and **(b)** clustered (**Model 2**) point event data. Their values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the change in colour scale ranges from **Figures 4.1** and **4.2**.

For the homogeneous point event dataset (**Figure 5.2a**), the clustering measure *D'* does not change its behaviour relative to clustering measure *D* (**Figure 4.1**). For the clustered point event dataset (**Figure 5.2b**), values of clustering measure *D* without standardisation (**Figure 4.2**) increased for increasing spatial and temporal range values until their peak at (s = 2, τ = 1.6). The distribution of the standardised clustering measure *D'* as calculated from the standardised K-functions for the clustered point event dataset shows a pattern of maximum values for small spatial and temporal range values with decreasing trend that reaches almost zero after (s = 2, τ = 1.6).

The main visual difference between the clustering measure *D* and its standardised form is the absence of clear regions of increasing values after reaching the peak at (s = 2, τ = 1.6). The given combination of spatial and temporal ranges corresponds to the expected dimensions of each individual 'children' point event cluster. From **Table 3.5** and discussion of **Figure 4.2**, the reasoning behind encountering local increases in clustering measure *D* values was due to the selected Gaussian distribution standard deviation value of σ = 0.5. The selection of the standard deviation parameter then determines how spread around each 'parent' event are their corresponding 'children' events that comprise the final point event dataset Model 2. As the spatial and temporal ranges are increased around an arbitrary point event, the methodology encounters events from other clusters ('children' events around a different 'parent' event). The standardised clustering measure *D'* then identified only the first (global) peak in its values as further spatial and temporal ranges have *D* values decreased due to the standardisation.

For Model 3 and its variations as listed in **Table 3.6**, the K-function analysis results for the standardised K-functions and the standardised clustering measure *D'* are shown in **Figure 5.3**. It is noted that the same Models 3 and its variations and Models 4 and its variations, the initialisations are used for applying the standardised K-functions. Each model consists of the same N = 3,000 and N = 10,000, respectively, for Models 3 and 4. The spatial and temporal ranges are kept the same as for the analysis of clustering measure *D* for Models 3, 4 and their variations at s = {2, 4, ..., 120} km and τ = {0.05, 0.1, ..., 2} h.

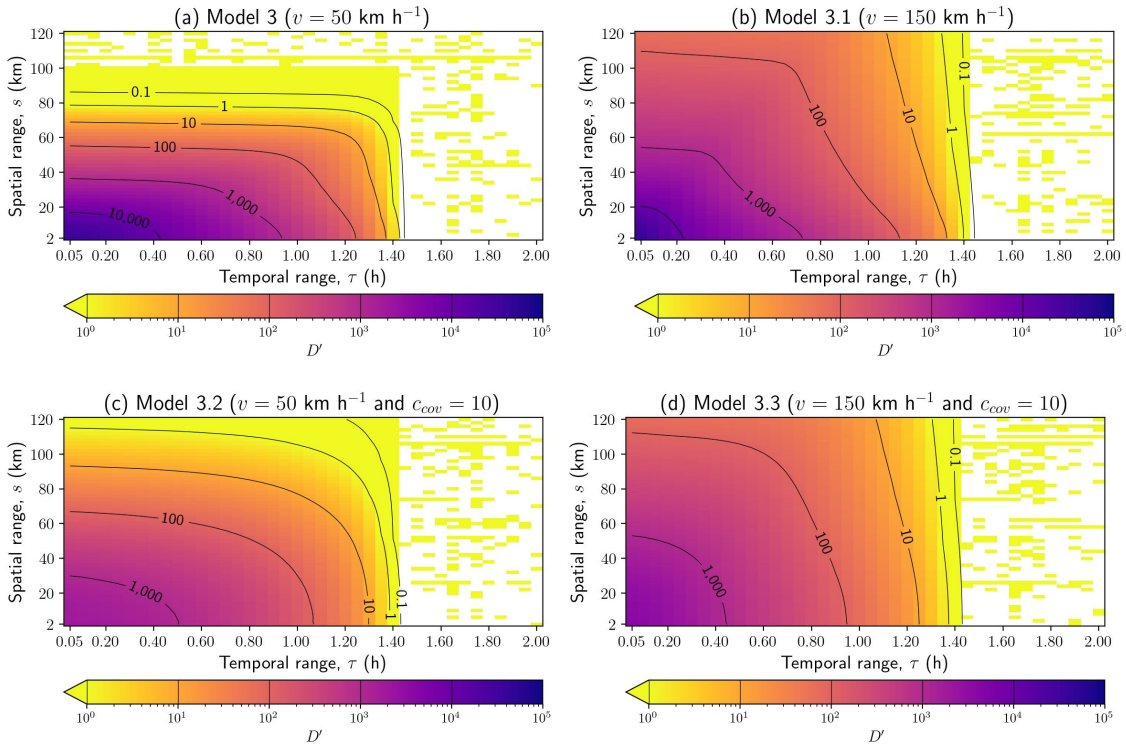

Figure 5.3. The standardised clustering measure $D' = K'(s,\tau)–K'(s)K'(\tau)$ as applied to the **(a)** reference single thunderstorm model (**Model 3**) and its variations: **(b) Model 3.1**, **(c) Model 3.2** and **(d) Model 3.3**. The particular differences in the model variations are indicated in the titles of the subfigures. The D' values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the different colour scales for the clustering measure D' from previous figures. The colour scale is a log-scale and continues as 'yellow' for any D' values below 1 and is 'white' for negative or non-existent values. The various D' value thresholds are indicated by black contours.

The clustering measure D' values for Model 3 and its variations in **Figure 5.3** show the behaviour expected after standardising the clustering measure D, when compared to the results from **Figure 4.3**. The reference Model 3 (**Figure 5.3a**) shows similar spatial and temporal extents to **Figure 4.3a** (s = 100 km, τ = 1.4 h) as the standardisation does not change the relative position of the point events or how the spatial and temporal ranges are measured. The dimensions of the synthetic storm are also immediately detectable by the standardised clustering measure. The range of the D' values as marked by the black contours extends over five orders of magnitude from the contour D' = 1. In **Figure 4.3a**, the clustering measure D increases, reaching maximum values around (s = 40 km, τ = 0.55 h) and decreasing afterwards. The clustering measure D' values show a constant decrease for increasing spatial and temporal range values as the extended scanning volume affects the clustering measure.

The variations to the reference Model 3 show contrasting behaviours relative to the reference as expected. Increasing the movement speed of the lightning strike cluster from v = 50 km h^{-1} to v = 150 km h^{-1} for Model 3.1 (**Figure 5.3b**) extends the spatial range with relatively significant *D'* values. Previously *D'* value contour *D'* = 100 was observed to be around s = 55 km for low τ values, but the model variation has extended this value contour to s = 110 km. When leaving the movement speed as the reference Model 3, but changing the covariance matrix factor for the distance to the movement track variable, Model 3.2 (**Figure 5.3c**) also exhibits noted varied behaviour.

Similarly to Model 2, the spatial range values with relatively significant *D'* values is extended, while not as much as Model 2 (*D'* value contour *D'* = 100 extents to s = 65 km for low τ values). There are also no *D'* values higher than the contour *D'* = 10,000 as were observed for Model 3 and Model 3.1. The combined increase in the movement speed and the distance to the movement track in Model 3.3 exhibits the behaviour expected following the analysis of the previous two models (**Figure 5.3d**). The *D'* value contour *D'* = 100 is observed to s = 115 km for low τ values. Similarly to Model 3.2, no *D'* values higher than the contour *D'* = 10,000. When comparing directly to Model 3.1, the *D'* value contours show a decrease in values with increasing temporal range τ value for small s values (e.g. for s = 2 km, contour *D'* = 100 is at τ = 0.95 h compared to τ = 1.12 h for Model 3.1).

As mentioned in the analysis of **Figure 5.3a** and observable in all other subfigures, the standardised clustering measure *D'* values reach their maximal values at the point of origin – the selected spatial and temporal step size values (i.e. 2 km and 0.05 h for Model 3) and decreased for larger range values. Within the spatio-temporal clustering application and analysis framework (**Section 2.5**), the standardisation then raises the issue of selecting the set of optimal bandwidth values for the kernel density estimation and the spatial and temporal extents of the space-time scan statistic. A significance level for the clustering measure *D*' values is established to approach this issue. The clustering measure *D* in **Figure 4.3**, once calculated for spatial and temporal ranges close to the extent of the dataset, decrease abruptly. The standardised clustering measure clustering measure *D'* follows a more gradual decrease due to the standardisation. Establishing a certain measure value, indicative of significant space-time interaction, would give a simple visual assessment of the ranges of clustering.

As a significance test, a Monte Carlo test is used to establish the spatial and temporal ranges for which the clustering measure *D'* indicates significant space-time interaction. In the Monte Carlo test, a total of N_{M1} = 99 Model 1 runs are initiated and the standardised K-function analysis methodology applied to all initiations and over the same spatial and temporal ranges of s = {2, 4, ..., 120} km and τ = {0.05, 0.1, ..., 2} h. If the clustering measure *D'* value is higher than all N_{M1} = 99 Model 1 runs for a given spatial and temporal threshold combination (s, τ), the result is said to be 'significant' ($p < 0.01$). Before the Monte Carlo test to assess the significance of the clustering measure *D'* values, an initiation of the homogeneous Poisson point process (Model 1) with N = 3,000 point events is shown in **Figure 5.4**.

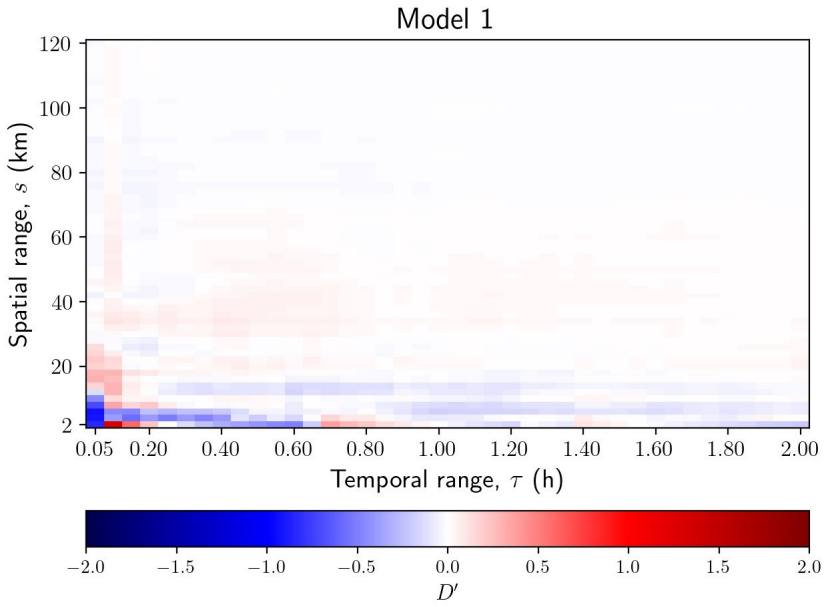

Figure 5.4. The standardised clustering measure $D' = K'(s,\tau) - K'(s)K'(\tau)$ as applied to a single initiation of the homogeneous Poisson process (**Model 1**) with N = 3,000 point events over the $\{X = 1000, Y = 1000, T = 24\}$ space-time volume. The *D'* values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the different colour scale for the clustering measure *D'* from previous figures. Note that the colour scale for **Model 1** results is on a linear scale.

When comparing the clustering measure *D'* values in **Figure 5.4** to **Figure 5.3**, there is a lack of structure and lack of significant absolute difference in the measure values. A linear scale of $D' = [-2, 2]$ fully covers all *D'* values with the extrema observed only for the lower spatial and temporal threshold values as these value patterns for small range values are dependent on the particular initiation of Model 1. When comparing these results to **Figure 5.2**, the increase in the total number of point events from N = 1,000 did not affect the clustering measure as expected from Model 1. The results of the Monte Carlo test for the reference Model 3 and its variations are shown in **Figure 5.5**.

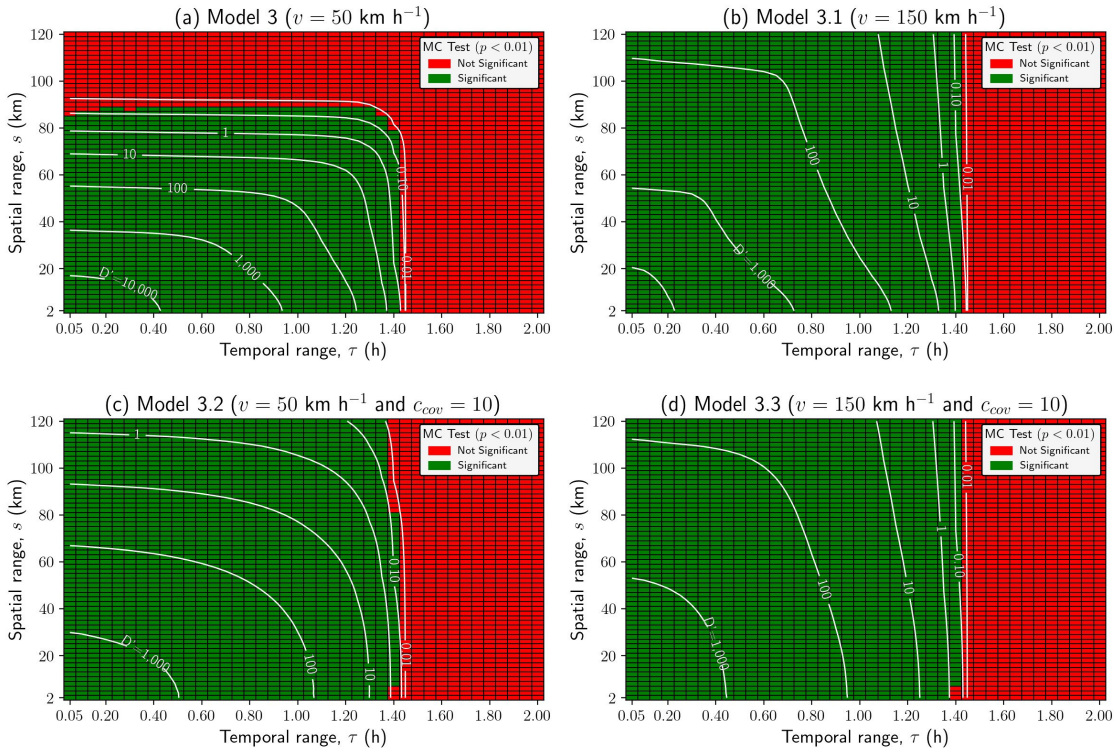

Figure 5.5. The results of a Monte Carlo (MC) test for significance of the standardised clustering measure *D'* results. A total of 99 initiations of **Model 1** are run for this test and the clustering measure *D'* value for a given pair of spatial range *s* and temporal range *τ* values is significant if it is larger than all 99 clustering measure *D'* values returned by **Model 1**. The results are shown for the **(a)** reference single thunderstorm model (**Model 3**) and its variations: **(b) Model 3.1**, **(c) Model 3.2** and **(d) Model 3.3**. Significant results (i.e. cells representative of the combinations) for given combinations of {*s*, *τ*} are marked as green and not significant as red. The various *D'* value thresholds are indicated by white contours overlaying the subfigures.

In **Figure 5.5**, the Monte Carlo test is used to establish the clustering measure *D'* values that can be deemed 'significant' (relative to N_{M1} = 99 Model 1 runs) for the reference single thunder storm Model 3 and its variations. For the reference Model 3 (**Figure 5.5a**), the significance tests indicate the significance cut-off to occur between the contour *D'* = 0.10 and the contour *D'* = 0.01, marking a decrease (from *s* = 100 km to 85 km) in the initially estimated spatial extent of space-time interaction for Model 3.

While not performing the Monte Carlo test used to establish the clustering measure *D'* values that can be deemed 'significant' for Model 4 (and variations) as in **Table 3.7**, the standardised clustering measure *D'* values are shown in **Figure 5.6**. As for **Figure 4.4**, two extensions are added with larger spacing values. Spatial range is extended to *s* = {110, 115, ..., 300} km for *τ* = {0.1, 0.2, ..., 1.0} h. The temporal range is extended to *τ* = {1.85, 1.95, ..., 2.25} h for *s* = {5, 10, ..., 110} km. The region of *s* = {110, 300} km and *τ* = {1.0, 2.25} h is left without results.

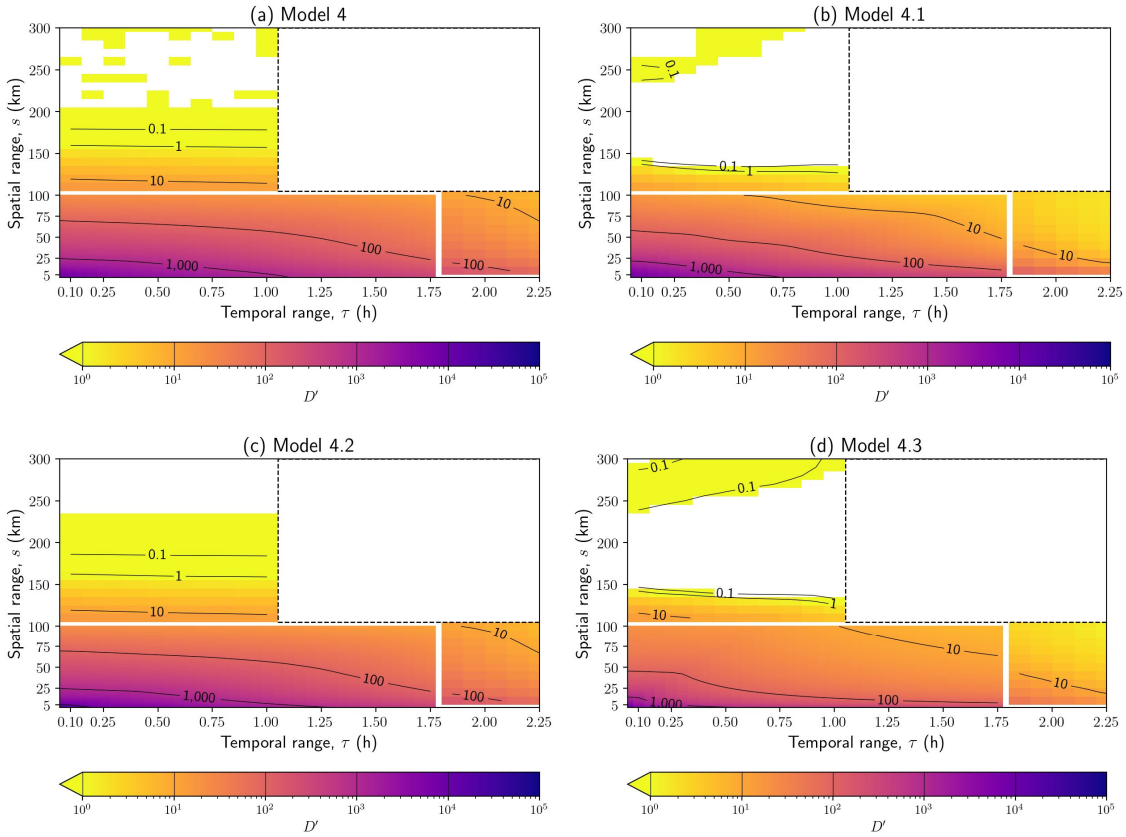

Figure 5.6. The standardised clustering measure $D' = K'(s,\tau)–K'(s)K'(\tau)$ as applied to the **(a)** reference three thunderstorm model (**Model 4**) and its variations: **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. The particular differences in the model variations are indicated in the titles of the subfigures. The D' values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales. Note the different colour scales for the clustering measure D' from previous figures (but same as **Figure 5.3**). The colour scale is a log-scale and continues as 'yellow' for any D' values below 1 and is 'white' for negative or non-existent values. The various D' value thresholds are indicated by black contours. The region of s = (105, 300) km and τ = (1.05, 2.25) h is left without results (marked by dashed lines).

The clustering measure D' values for Model 3 and its variations in **Figure 5.6** show the behaviour expected after analysis of **Figures 4.4** and **5.3**. Variations in the model parameters result in the same observable changes in the clustering measure value extents. The standardised clustering measure D' values reach their maximal values at the point of origin – the selected spatial and temporal step size values (i.e. 2 km and 0.05 h for Model 3) and decreased for larger range values. For Models 4.1 and 4.3 (**Figures 5.6b** and **5.6d**), the region (s = {145, 240} km for τ = {0.1, 1} h) of negative D' values and the positive values for s > 240 km were discussed and evaluated in the context of **Figures 4.4** and **4.5**. The main benefit of the standardisation for Model 4 (and its variations) is similar to Model 2 (**Figure 5.2**), as the positive clustering measure values for s > 240 km are now standardised by the search volume and do not have values indicating that the space-time interaction is as significant as for smaller range values.

For improvements on the non-standardised clustering measure *D* and its analysis framework, this standardisation of the clustering measure does not improve any computational complexity aspects. The standardisation allows to assess the space-time interaction closer to the spatial and temporal extents of the dataset as the values decrease gradually compared to the clustering measure *D*. The Monte Carlo test (**Figure 5.5**) indicated a value threshold between the contours $D' = \{0.01, 0.10\}$ for Model 3 and its variations. Similarly to Model 2 (**Figure 5.2**), the standardisation minimises the effect of encountering space-time interaction over larger range values, indicating the spatial and temporal extents of the most significant point events cluster.

For the spatio-temporal application and analysis framework (**Section 2.5**), the clustering measure *D* maximum value ranges informed the set of potential bandwidth values for the kernel density estimation (**Section 4.2.2**) and the spatial and temporal extents for the space-time scan statistic scanning volume (**Section 4.2.3**). While in **Sections 4.2** and **4.4**, the results indicated that the kernel density estimation, when using the leave-one-out estimator selects optimal bandwidth values smaller than the clustering measure *D* maximum value ranges, while the space-time scan statistic identified the maximum allowed spatial and temporal extend values. Further research could indicate whether there is a certain *D'* value (as a contour) that can correspond to the maximum measure *D* value ranges. As the K-function value with increasing spatial and temporal ranges would increase as the scanning volume [**Perry *et al.*, 2006**; **Diggle *et al.*, 1995**] for a homogeneous Poisson process, the standardisation would indicate when individual K-functions are returning results of increased space-time interaction relative to a homogeneous dataset.

In this section of presenting the improvements to the individual methodologies, the kernel density estimation results from **Section 4.2.3** are used here to select the space-time voxels with the probability density value threshold of 0.1% of the highest probability density values. As the elliptic scan statistic results for Model 4 and its variations (**Figure 4.14**) indicated the inability of the methodology to identify point events clusters produced by a moving source due to the relatively small temporal separation, further results will only be presented and analysed for Models 4 and its variations in this section. The already established input parameters (**Table 4.6**) and the inclination angle (**Figure 5.1**) for the inclined elliptic space-time scan statistic and the sets of their values selected are given in **Table 5.1**.

Table 5.1. Input parameters and their values for the inclined elliptic space-time scan statistic for single storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**).

Input parameter	**Input values**			
	Model 4	**Model 4.1**	**Model 4.2**	**Model 4.3**
Centroid locations	From **Fig. 4.11a**	From **Fig. 4.11b**	From **Fig. 4.11c**	From **Fig. 4.11d**
Length of semi-major axis	{5, 10, …, 30}	{5, 10, …, 25}	{5, 10, …, 30}	{5, 10, …, 50}
Shape of elliptic base	$\{1, 1.5, \ldots, 3, 3.5, 4\}$			
Angle	$\{0, (1/4)\pi, \ldots, (3/4)\pi\}$			
Inclination height	{0.50, 0.75, …, 1.25}	{0.25, 0.75, …, 1}	{0.50, 0.75, …, 1.25}	{0.25, 0.75, …, 1}
Height – Inclination height	{0.05, 0.1, …, 0.50}	{0.05, 0.1, …, 0.40}	{0.05, 0.1, …, 0.50}	{0.05, 0.1, …, 0.50}

The input parameters already established and used in **Chapter 4** are not further discussed and their ranges are kept constant (only the semi-major axis length is halved as discussed in **Section 4.2.4**). The 'Inclination height' parameter in **Table 5.1** is related to the inclination angle parameter θ in **Figure 5.1**. For easier implementation, this parameter is expressed as the temporal distance (in hours) between the two points of the major axis of the elliptic cylinder. Due to the inclination, two parallel planes pass through the elliptic cylinder scanning volume, which assigns all points events (as enclosed by the planes and the scanning volume) to the identified local point event cluster.

The choice of this 'Inclination height' parameter is dependent on the expected movement speed values of the moving source. This was taken into when creating the set of input values for this parameter. These implementation details are visualised with the space-time scan statistic results and further discussion follows. The input parameters for the inclined elliptic space-time scan statistic and the sets of their values selected are given in **Table 5.1** and the results for Model 4 are shown in **Figure 5.7** with results for all Model 4 variations in **Appendix Q**. The discussion on the exact details of limiting the space-time volume with the inclination height follow the results and is applicable all model variations.

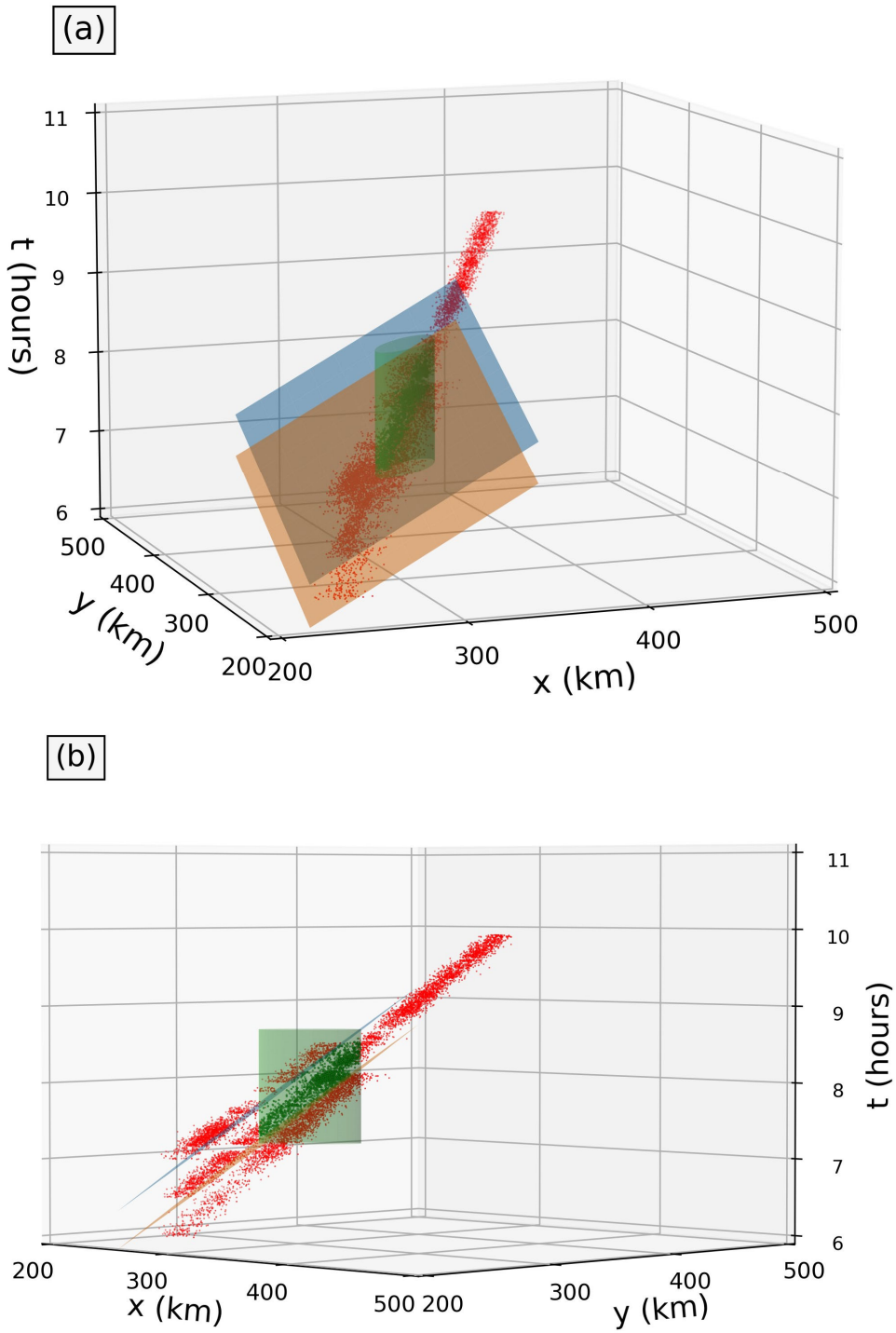

Figure 5.7. 3D views [**(a)** from the same angle as **Figure 4.14** and **(b)** looking perpendicular to the movement direction**]** of the spatio-temporal domain showing the point events contained within the inclined elliptic space-time scan statistic as applied to **Model 4**. In all subfigures, the events contained within the elliptic space-time volume with an inclination height as identified by the scan statistic are coloured green, while those outside in red. The elliptic scanning volume is drawn in transparent green. The upper and lower planes, controlling the inclination height of the space-time scanning volume are coloured in light blue and orange, respectively.

In **Figure 5.7** and **Appendix Figure R1**, when viewed from the direction perpendicular to the movement direction of the synthetic storm source (**Figure 5.7b**), the 'Inclination height' can be marked as the time interval between the time coordinate of the points where the lower orange planes intersects the projected major axis point of the elliptic space-time cylinder. This changed the definition of the 'Height' input parameter from **Section 4.2.3** to a new parameter 'Height – Inclination height', defining the temporal distance between the orange and the blue planes intersecting the elliptic cylinder with given 'Height' values from the previous definition. A further visual representation and explanation on this can be seen in **Appendix Figure R2**.

When compared to **Figure 4.14a**, the improvement to the accuracy of the space-time scan statistic in identifying a cluster of point events within the same storm source is evident. The point events of the storm source S2 (**Table 3.7**) form the majority of the identified individual cluster with the 'Height – Inclination height' being too large and assigning some events from S3 to this cluster. The methodology required a detailed and careful selection of the input parameter ranges as shown in **Table 5.1**. Further evaluation of this aspect follows in **Section 5.6**. The input parameters for the inclined elliptic space-time scan statistic and the sets of their values selected for Model 4 and its variations are given in **Table 5.2**.

Table 5.2. Identified input parameter values by the inclined elliptic space-time scan statistic on three storm reference **Model 4** and its variations (**Models 4.1, 4.2, 4.3**).

Input parameter	**Identified values**			
	Model 4	**Model 4.1**	**Model 4.2**	**Model 4.3**
Centroid location	{305, 305, 7.45}	{250, 255, 7.10}	{310, 310, 7.10}	{355, 355, 8.6}
Length of the semi-major axis	30 km	20 km	30 km	30 km
Shape of elliptic base	3	2	4	4
Angle	$(1/4)\,\pi$	$(1/4)\,\pi$	$(1/4)\,\pi$	$1/4)\,\pi$
Inclination height	1 h	0.75 h	1.25 h	1 h
Height – Inclination height	0.5 h	0.4 h	0.5 h	0.5 h

5.5. Results of improved clustering measures on real-world datasets

When applying the spatio-temporal clustering analysis framework (**Section 2.5**) to the real-world lightning strike dataset (**Section 3.2.1**) in **Section 4.4**, the results of the individual methodologies were as expected (from characterisation results in **Section 3.3**) and informative. The framework correctly identified the ranges of space-time interaction and the elliptic scan statistic identified Storm 2A (**Section 3.3.2.3**) as the space-time volume with the highest density of point events, given the scanning volume input parameters.

The selected local clustering measure, elliptic space-time scan statistic was unable to identify individual clusters of point events within the same storm source when applied to the three storm reference Model 4 and its variations. These limitations were addressed by introducing inclination in the scanning volume and results were presented for Model 4 in **Section 5.4**. Due to the storms within the real-world lightning strike dataset being temporally well-separated, the elliptic space-time statistic performed well. The analysis in this section focuses on comparing the standardised K-function results with **Section 5.4** and identifying the changes in the input parameters necessary to adjust for the new scan statistic input parameters (**Table 5.1**) when applied to the real-world lightning strike dataset.

The results of the K-function analysis with the standardization introduced in **Section 5.2** as applied to the real-world lightning strike dataset are shown in **Figure 5.8**. The lightning strike spatial coordinates are projected onto the OSGB 1936 (EPSG:27700) northing/easting coordinates (**Section 3.3**). The scale of the dataset is unchanged from the dataset to which the clustering measure D was applied in **Figure 4.18**.The spatial and temporal ranges remain set as $s = \{5, 10, ..., 400\}$ km and $\tau = \{0.1, 0.2, ..., 4.5\}$ h.

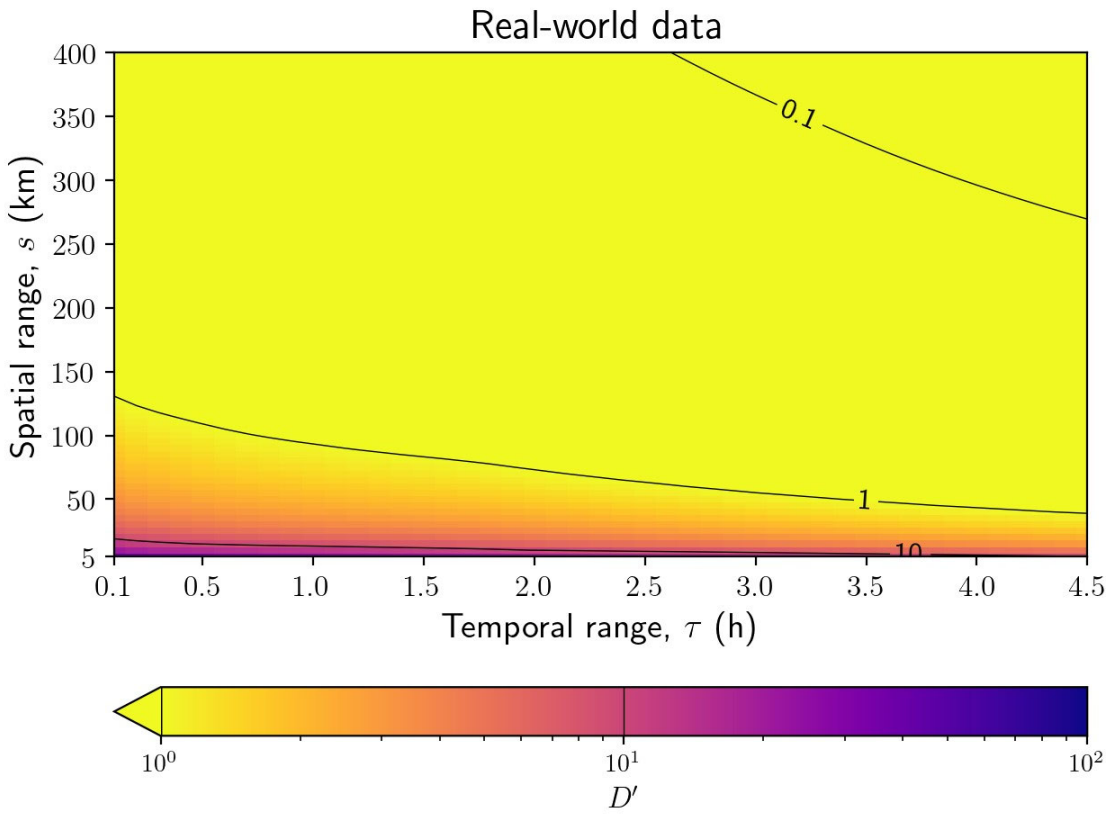

Figure 5.8. The standardised clustering measure $D' = K'(s,\tau)–K(s)K(\tau)$ [spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$] as applied to lightning strike location data. Their values are calculated for various combinations of spatial range s (km) and temporal range τ (hours). Note the different colour scale upper limit for the clustering measure D from K-function analysis framework result figures in **Section 4.4**.

In **Figure 5.8**, the standardised clustering measure D' reaches it maximum value for the spatial and temporal ranges origin values (i.e. the selected step sizes for both ranges) at $\{s = 5$ km, $\tau = 0.1$ h$\}$. The maximum value within the range $D' = (10,100)$ is four orders of magnitude smaller than the clustering measure D maximum in **Figure 4.18**. Due to the standardisation, the clustering measure D' values are between value contours $D' = (0.1,1)$ for most of the spatial and temporal range region used for this dataset. When comparing these results to non-standardised clustering measure D values in **Figure 4.18** and the standardised results for Model 3 and its variations in **Figure 5.6**, the previous identified maximum clustering measure D value ranges $\{s_{MAX} = 320$ km, $\tau_{MAX} = 4.4$ km$\}$ are by the $D' = 0.1$ contour. As mentioned in **Section 5.4**, further research could indicate whether a certain D' value (as a contour)can correspond to the maximum measure D value ranges.

In this section, as part of the spatio-temporal application and analysis framework (**Section 2.5**) kernel density estimation results from **Section 4.4.3** and **Figure 4.19** are used here to select the

space-time voxels with the probability density value threshold of 0.1% of the highest probability density values as the centroid locations. The already established input parameters (**Table 4.8**) and the inclination angle (**Figure 5.1**) for the inclined elliptic space-time scan statistic and the sets of their values selected are given in **Table 5.3**.

Table 5.3. Input parameters and their values for the inclined elliptic space-time scan statistic as applied to the real-world lightning strike dataset.

Input parameter	Input values
Centroid location	From **Fig. 4.19b**
Length of the semi-major axis	{100, 105, …, 150}
Shape of elliptic base	{1, 1.5, …, 3, 3.5, 4}
Angle	$\{0, (1/4)\pi, \ldots, (3/4)\pi\}$
Inclination height	{2, 2.25, …, 4}
Height – Inclination height	{0.05, 0.1, …, 0.50}

As noted in **Section 5.4**, the choice of the input value range for the 'Inclination height' dependent on the expected movement speed values of the moving source. Given that the 'Length of the semi-major axis' parameter range extends up to 150 km and the previous 'Height' parameter was found to be 4.0 h (**Figure 4.9**), the expected movement speed would be approximately 75 km h^{-1}. This is not far off the characterised movement speed v of Storm 2A (**Table 3.3** in **Section 3.3.4.2**), considering the step size in input parameters for the space-time scan statistic. Therefore, the 'Inclination height' and 'Height - Inclination height' input parameters have to add up to the previous 'Height' parameter. For the selected input parameter value ranges in **Table 5.3**, the inclined elliptic space-time scan statistic results for the real-world lightning strike dataset are shown in **Figure 5.9** and the input parameters and their values for defining the identified elliptic cylinder and the two planes encompassing the identified lightning strike cluster are listed in **Table 5.4**.

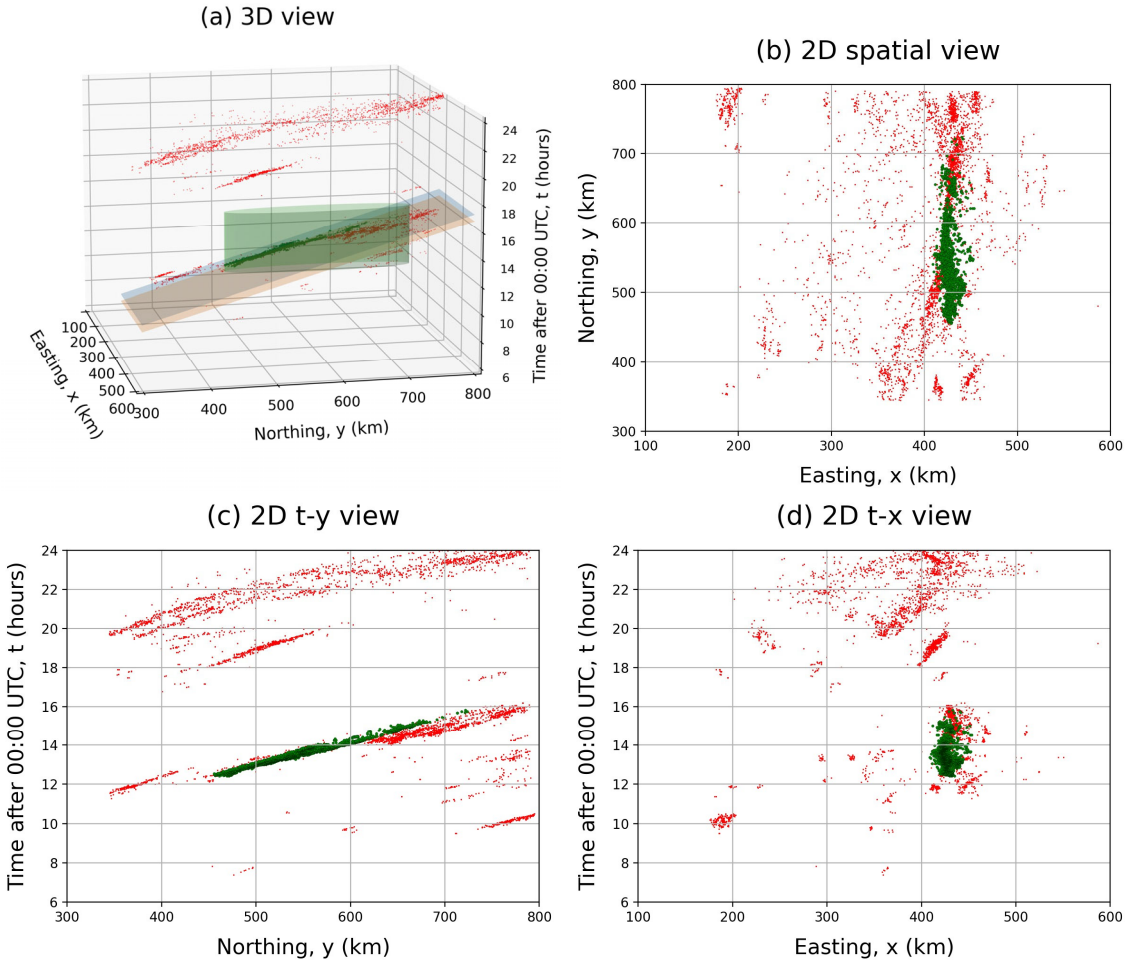

Figure 5.9. Various angles [**(a)** 3D view of the spatio-temporal domain, **(b)** 2D spatial view, **(c)** 2D *t*–*y* view, **(d)** 2D *t*–*x* view] of the spatio-temporal domain showing the point events contained within the inclined elliptic space-time scan statistic as applied to the real-world lightning strike data. In all subfigures, the events contained within the scan statistic are coloured green, while those outside in red. In **(a)**, the elliptic scan statistic is drawn in transparent green. The upper and lower planes, controlling the inclination height of the space-time scanning volume are coloured in light blue and orange, respectively.

Table 5.4. Identified parameter values for the inclined elliptic space-time scan statistic as applied to the real-world lightning strike dataset.

Input parameter	Identified values
Centroid location	{430, 600, 14.1}
Length of the semi-major axis	145 km
Shape of elliptic base	4
Angle	$(1/2)\,\pi$
Inclination height	3.5 h
Height – Inclination height	0.5 h

In **Figure 5.9**, the inclined elliptic scan statistic has identified a cluster of lightning strike locations coinciding with the approximate initiation, duration, and termination of the Storm 2A of the second case study (**Section 3.3.2.3**), just as seen in **Figure 4.20**. The elliptic cylinder as the scanning volume has managed to capture the northwards propagation direction of Storm 2A as the angle parameter is found to be 0.5π relative to the first spatial dimension (i.e. easting) and the shape of the elliptic base is 4.0 (**Table 5.4**). The value of applying the additional 'Inclined height' parameter is seen from the identified value for 'Height – Inclination height' at 0.5 h, indicating temporal extent of lightning activity over a particular spatial location. It is also observed that the identified values for the 'Inclination height' and 'Height – Inclination height' input parameters add up to 4 h (**Table 5.4**) as the previous parameter 'Height' in **Table 4.9**.

Compared to **Figure 4.20**, there are minor changes to the centroid location and the semi-major axis length. The inclination in the space-time scanning volume also allows assigning the initial point events at the beginning of the main cluster (**Figure 5.9c** compared to **Figure 4.20c**). Due to the 'Height – Inclination height' parameter, no additional point events are assigned from 'below' the main cluster near the end of its duration in **Figure 5.9c**.

5.6. Evaluation of improved clustering measures

In **Sections 5.2** and **5.3**, the spatio-temporal clustering methodology application framework (**Section 2.5**) for the point event data type was improved in terms of the limitations identified for individual clustering methodologies from the evaluation in **Sections 4.2.4** and **4.4.4**. The various K-functions (**Section 2.3.1.1**) were standardised by the search window volume for the various spatial and temporal range values combinations in **Section 5.2**, while the elliptic scan statistic was extended to include an inclination to account for a moving source producing the point events (**Section 5.3**). These extensions to the methodologies were applied to the synthetic simulation study (**Section 5.4**) and the real-world lightning strike (**Section 5.5**) datasets.

The selected application framework improvements returned results expected from the initial intuition for their use. The standardised K-function analysis results (**Figures 5.2–5.6**) show clear maximum values for the smallest spatial and temporal range values, while decreasing more gradually with increase ranges compared to **Figures 4.2**–**4.4**. For Models 2, 4.1 and 4.3 (**Figures 5.2b**, **5.6b** and **5.6d**), the standardisation of the K-function values does not allow

significantly larger range values to exhibit values as significant as for smaller ranges, indicating more reasonable cluster threshold range values. This standardisation does not affect the overall application framework as it is a more visual change. A Monte Carlo test was used (**Figure 5.5**) to establish a significance level for the standardised results, returning a significance cut-off between the contour $D' = 0.10$ and the contour $D' = 0.01$.

For use in kernel density estimation, the estimated spatial and temporal ranges of maximum space-time interaction $\{s_{MAX}, \tau_{MAX}\}$ by the clustering measure D serve as reasonable initial indication of the extents of space-time interaction. In this section, the kernel density estimation results from **Section 4.2.3** were used without change. For use in the space-time scan statistic methodology as centroid location, the space-time voxels with the probability density value threshold of 0.1% of the highest probability density values.

After evaluation of the elliptic space-time scan statistic results for the point event simulation study (**Section 4.2.4**), the results for Models 4, 4.1, 4.2 and 4.3 (**Figure 4.14**) indicated the inability of the methodology to identify point events clusters produced by a moving source due to the relatively small temporal separation. In **Figure 5.7**, the intuition for the use of inclined (relative to the 2D spatial domain) planes to modify the elliptic scan statistic was presented and applied to Models 4, 4.1, 4.2 and 4.3. The improved methodology was able to now correctly identify individual clusters containing point events from one single storm source even with small temporal separation between them. This improvement to the methodology was also observed even in the temporally well-separated real-world lightning strike dataset in **Figure 5.9**. The inclined scan statistic returned a more compact volume and did not include additional point events around the main structure of Storm 2A (**Section 3.3.2.3**) as observed when applying just the elliptic cylinder volume (**Figure 4.20**).

For the inclined elliptic scan-statistic, the value of applying the additional 'Inclined height' and 'Height - Inclination height' input parameters was evident in providing additional physical interpretation. The previous 'Height' parameter is split into these two additional parameters with 'Inclined height' linked to the movement speed of the moving source producing the point events and 'Height – Inclination height' indicating the storm's temporal duration at a particular spatial location (averaged over total storm duration). This is a significant improvement both in terms of clustering performance and physical interpretability of the clustering results.

5.7. Chapter conclusions

In this research chapter, improvements to the individual methodologies within the spatio-temporal clustering application and analysis framework (**Section 2.5**) were proposed and applied to the real-world lightning strike dataset and the simulation study (Model 1 and Model 2 as synthetic benchmarks and Model 3 and Model 4 as synthetic models representative of single and three thunderstorm systems) as presented in **Sections 3.2** and **3.4** respectively.

The research on a common spatio-temporal application and analysis framework by using global and local clustering measures for natural hazard data in [**Tonini *et al.*, 2022**; **Wang *et al.*, 2021**; **Tonini and Cama, 2019**; **Tonini *et al.*, 2017**] was discussed and extended in **Chapter 2** and applied to synthetic and lightning strike datasets in **Chapter 4**. No previous global and local clustering measures framework has been applied to lightning strike point event data in the spatio-temporal domain. The results indicated the inability of the elliptic space-time scan statistic methodology [**Kulldorff *et al.*, 2006**] to identify point events clusters produced by a moving source due to the relatively small temporal separation (Models 4, 4.1, 4.2 and 4.3).

The first proposed improvement was the standardisation of the K-functions with the scanning volume around an arbitrary point as defined by the spatial and temporal ranges. This does not allow significantly larger spatial and temporal range values to exhibit clustering measure values as significant as for smaller ranges, indicating more reasonable cluster dimensions. This standardisation does not affect the overall application framework as it is a more visual change. A potential further research direction on established a significant value threshold is discussed.

The second proposed improvement was the elliptic space-time scan statistic adjusted to include an inclination in its scanning volume. The elliptic cylinder scanning volume is intersected by two planes with one defined by the temporal inclination between the ellipse major-axis points (i.e. the second point projected at the inclination height) and the other by their temporal distance. This significantly improved the space-time scan statistic performance for Models 4, 4.1, 4.2 and 4.3 and the real-world lightning strike data. Additional physical interpretation was discussed as a result of this approach specific to point events produced by a moving source.

6. Summary, conclusions and further research directions

6.1. Introduction

This final chapter of this thesis provides a summary, conclusions and some future research directions. As introduced in **Chapter 1**, natural hazards are inherently spatio-temporal processes as they have various properties in the separate spatial and temporal domains and the combined spatio-temporal domain. The initiation and propagation of natural hazards pose several risk factors for both urban and natural environments, resulting in various economic, environmental and human costs. Therefore, better understanding the spatio-temporal patterns of the various types of natural hazards can aid in better risk quantification and understanding of the natural hazard process. Working in the spatial and temporal domains separately might hinder insight into the processes and quantification of the physical variables of the natural hazard that are observable only in the combined spatio-temporal domain. Understanding the spatio-temporal clustering patterns and their relationship and connection with the various physical variables of the natural hazard process is essential.

This thesis has aimed (i) to synthesise into a common framework and apply spatio-temporal clustering analysis methodologies and (ii) to develop spatio-temporal models to reproduce observed natural hazard clustering patterns. **Chapter 1** identified four main research objectives (**O.1**–**O.4**), with specific research directions to complete these aims. The results of this thesis are contextualised within these four objectives in **Section 6.2**, identifying synthesised results from **Chapters 2**–**5** that contribute to completing each of the individual research objectives and answering any research questions relevant to each individual objective and chapter.

Through the synthesis of the thesis research results and as already mentioned in some limitation discussions and conclusions for **Chapters 2**–**5**, three potential further research directions are proposed and presented in **Section 6.3**. These research directions were initiated during this thesis but were not completed either due to being out of the scope of this thesis or requiring additional expertise and research. The further research directions address concerns with some aspects of the research undertaken in this thesis. Overall concluding remarks of the thesis are then presented in **Section 6.4**.

6.2. Relationship of thesis to original research objectives

Here the results and conclusions of **Chapters 2**–**5** are synthesised to indicate their contribution to completing the four thesis research objectives from **Chapter 1** to meet the research aims. The four identified thesis research objectives (**O.1** to **O.4**) are as follows with a synthesis of all relevant chapters and research results mentioned after each individual objective:

(O.1) Critical literature review of spatio-temporal clustering methodologies.

This first objective was mostly addressed through the critical literature review in **Chapter 2**. A list of research questions for this chapter was as follows:

Q2.1. What spatio-temporal clustering methodologies are available for the particular data types: point events and trajectories?
Q2.2. What is the usage frequency of various spatio-temporal clustering methodologies?
Q2.3. Which methods are well-established, and which are still being developed (i.e. publication timeline)?
Q2.4. Which fields of study apply clustering methodologies in the spatio-temporal domain?
Q2.5. Are there particular methodologies that have been extensively used in only particular fields of study?
Q2.6. Can the reviewed methodologies be categorized in broader groups depending on their application purpose and used in a common framework?

The research questions posed for **Chapter 2** were answered in that chapter and appendices. Identifying the various clustering methodologies in **Section 2.2** and their intuition and implementation details in **Section 2.3** address **Q2.1** and **Q2.2** in a non-exhaustive way, as only the most common methodologies were identified and selected for implementation. For the point event data type, these were the K-functions analysis, kernel density estimation and the space-time scan statistic. For trajectories, Trajectory-OPTICS methodology was selected as part of the density-based methodology category identified by the critical literature review.

Section 2.2 addressed the research questions **Q2.3**–**Q2.5**, as a research timeline was discussed in **Section 2.2.2** with indications about the most common fields of study and application, based on the critically reviewed $N = 143$ scientific publications. The space-time scan statistic was

identified as a methodology that originally was developed for application in epidemiology research [**Kulldorff *et al.*, 2005**; **Kulldorff *et al.*, 1998**; **Kulldorff, 1997**], but has since been used for various natural hazards: landslides [**Tonini *et al.*, 2022**; **Tonini and Cama, 2019**], wildfires [**Parente *et al.*, 2016**; **Pereira *et al.*, 2015**; **Orozco *et al.*, 2012**; **Tuia *et al.*, 2008**] and floods [**Wang *et al.*, 2021**].

The categorisation of methodologies into global and local clustering measures is presented and discussed in **Section 2.4**, detailing their differences in clustering interpretation and purpose of application. A common application framework is proposed in **Section 2.5** based on the already established use of methodology combinations for various natural hazards: landslides [**Tonini *et al.*, 2022**; **Tonini and Cama, 2019**], wildfires [**Tonini *et al.*, 2017**], and floods [**Wang *et al.*, 2021**]. This addressed the research question **Q2.6**. This identified framework is then used throughout the application sections of this thesis in **Chapters 4** and **5**.

(O.2) Creation of spatio-temporal models representative of natural hazards.

The second objective of this thesis is completed by the characterisation of physical variables and modelling research done on two lightning strike case studies of increased lightning activity in **Chapter 3**. These models were used to generate synthetic point event datasets representative of lightning strikes, creating a simulation study of point event data for the application of the various clustering methodologies and their framework from **Chapter 2** in **Chapters 4** and **5**. The research questions proposed for the research in **Chapter 3** were as follows:

Q3.1. What are the established synthetic models for point-event natural hazards?
Q3.2. How are the physical properties of the natural hazards incorporated in these models?
Q3.3. For lightning strikes as point events produced by a moving source (i.e. thunderstorm system) – what physical properties of the hazard can be used to create a synthetic model?
Q3.4. What input parameter ranges can be classified as 'representative' of lightning?
Q3.5. What synthetic datasets representative of lightning strikes can be established in order to use them for application of the spatio-temporal clustering methodologies?

To understand the best approach to modelling lightning strikes as points, **Q3.1** and **Q3.2** were addressed by conducting a brief literature review of 20 publications (**Appendix Table G1**) to overview the various approaches to model spatio-temporal point event data of any natural

hazard. As an example, for wildfires [**Serra *et al.*, 2014**; **Xu and Schoenberg, 2011**; **Møller and Díaz-Avalos, 2010**], the Cox point process was one of the point process modelling approaches. A similar approach was initially attempted for lightning strikes with modifications to the point process, but was not pursued due to complexity and lack of interpretability.

A characterisation and modelling study was conducted for the physical variables of lightning strikes and the thunderstorm as the moving source producing them (**Section 3.3**) to address **Q3.3** and **Q3.4**. Two case studies of increased lightning activity were selected: 77,252 events as recorded on 28 June 2012 and 12,521 events as recorded on 1 July 2015 over the UK. Subsets of these datasets were assigned to three storms in each case study based on the synoptic analyses [**Lewis and Silkstone, 2017**; **Clark and Webb, 2013**] and auxiliary data. Three physical variables – inter-event time distribution, movement speed and spatial spread – are characterised and representative ranges are found for their values. These ranges correspond to values identified in other studies (**Section 3.3.6**).

After the characterisation results (**Section 3.3.4**), the modelling aspect extended the work of [**Finke, 1999**] by explicitly including lightning strike and thunderstorm physical variables into the model framework. After presenting the modelling approach (**Section 3.3.5**), single storm and three storm reference models are generated with their input parameters and their values based on the physical variables identified by the characterisation study. By varying these values and generating contrasting variations in these reference models, a simulation study consisting of 10 datasets (for point events) was created and presented (**Section 3.4.1**), addressing **Q3.5**. No characterisation was done for trajectories, but an approach to generating homogeneous and clustered trajectory datasets was presented (**Section 3.4.2**).

The research outcomes of objective **(O.2)** directly contribute to the following objectives **(O.3)** and **(O.4)**. These next objectives use the synthetic point event and trajectory datasets (also real-world data) to apply the various clustering methodologies and the general clustering analysis framework from objective **(O.1)**.

(O.3) Application of the spatio-temporal clustering methodology framework.

The third objective of this thesis follows from objectives **(O.1)** and **(O.2)**. Research conducted in **Chapters 1** and **2** allows for an immediate application of the proposed spatio-temporal

clustering analysis framework (**Section 2.5**) to the simulation study (**Section 3.4**) and real-world data (**Section 3.2**). The application results are presented, analysed and evaluated in **Chapter 4** with **Chapter 5** and objective **(O.4)** further contributing to objective **(O.3)**. For **Chapter 4**, the research questions were as follows:

Q4.1. If presented with a point event dataset produced by a moving source, what framework (i.e. combination and sequence of spatio-temporal clustering methodologies) would provide the most computationally efficient and complete clustering analysis?

Q4.2. How do the spatio-temporal clustering methodologies, whether global or local measures, compare between themselves in terms of their clustering accuracy and general purpose of application to point events produced by a moving source?

Q4.3. How sensitive are individual methodologies to changes in input parameters (in particular, the characterised physical variables) of the lightning strike spatio-temporal point event model?

The research in **Chapter 4** was focused on addressing **Q4.2** and **Q4.3**. When applying the spatio-temporal clustering methodology analysis framework to the simulation study, the results of the K-function analysis (**Section 4.2.1**) is used to demonstrate the purpose of a global clustering measure as it provides a single quantitative measure of clustering present, given spatial and temporal ranges around an arbitrary point event. A local clustering measure, the scan statistic, then identifies particular set of point events that are contained within a volume and assigned to a common cluster. This addresses **Q4.2**. A useful extension of the analysis would be to select multiple local and multiple global clustering measures. The accuracy in the clustering results within each measure group would aid in further developing **Q4.2**.

For addressing **Q4.3**, the single storm (Model 3) and three storm (Model 4) reference models and their variations are useful for understanding the effect of varying movement speed and spatial spread parameters (as generated for the simulation study in **Section 3.4**). The K-function analysis results (**Figures 4.3** and **4.4**) clearly show the effect of changing the physical variable values as selected for the model input parameters. When applied to Models 4.1 and 4.3 (particularly **Figures 4.4b** and **4.4d**), the K-function analysis framework returns regions with significant negative values due to the storm sources having significantly varying movement speeds, making point events close in time be far apart in space (**Figure 4.5**).

Another set of results that addressed **Q4.3**, were the elliptic space-time scan statistics results. For the scanning volume, the methodology is sensitive to what ranges of input parameters (**Section 2.3.1.3**) can be used for selection. At its implementation stage in **Chapter 4**, the scan statistic is also unable to account for a moving source producing the point events, combining point events from various storms into an individual cluster (**Figure 4.14**) if the temporal separation between storms is relatively small (0.5 h for Model 4 and its variations).

Q4.1 becomes computationally complex when using the spatio-temporal clustering methodology analysis framework (**Section 2.5**). There are certainly reductions in complexity when working with a limit set of voxels as centroid locations for the scan statistic, which would be beneficial as future work. As for the framework's completeness, the research in this thesis gives a clear and accurate (after improvements to the space-time can statistic in **Chapter 5**) way of understanding both the global and local clustering patterns of a point event datasets produced by a moving source.

(O.4) Improving on the limitations of the applied spatio-temporal methodologies from objective (O.3).

The final objective of this thesis follows the evaluation of the spatio-temporal clustering framework application results from completing objective **(O.3)**. The proposed improvement to the scanning volume of the space-time scan statistic in **Chapter 5** mostly completes this objective. The improvement is implemented and applied to the simulation study and real-world data, similar to **Chapter 4**. Research questions **Q4.1**–**Q4.3** are still applicable for this objective due to the repeated application of the methodology framework. In addition, **Chapter 5** proposes these research questions:

Q5.1. How can the performance shortcomings of the already implemented methodologies be addressed in the context of points events produced by a moving source?

Q5.2. Are improvements on these methodologies only in terms of the clustering accuracy or also in computational complexity, adjustments to the general methodology framework?

Q5.3. Can these improvements on the methodologies bring more detailed characterisation of the physical processes and variables?

Of the two proposed improvements in **Chapter 5**, the changes to the scanning volume of the space-time scan statistic (**Section 5.3**) had a significant positive impact on the clustering performance when applied to Model 4 and its variations (**Figure 5.7** compared to **Figure 4.14**). While the input parameter ranges had to be selected carefully, the inclined space-time volume assigned point events from the same storm source into an individual cluster. This improvement overcame the performance issue of the methodology when applied to point events produced by moving sources with small temporal separation; it worked well for temporally well-separated real-world data in **Section 4.4.3**. This addressed research question **Q5.1**.

The inclined elliptic space-time scan statistic (**Appendix Figure R2**), introduced two additional input parameters (**Table 5.1** compared to **Table 4.6**) relative to the elliptic space-time scan statistic, by splitting its 'Height' parameter into 'Inclination height' and 'Height Inclination height'. The 'Inclined height' linked to the movement speed of the moving source producing the point events and 'Height – Inclination height' indicating the temporal duration of the storm at a particular spatial location (averaged over the total duration of the storm). This aspect addresses **Q5.3**.

The research question **Q5.2** was addressed in part, as the overall framework was improved but not substantively changed. A standardisation approach to the K-functions was presented and analysed (**Section 5.2**), but without an immediate purpose within the general methodology framework as it was used more visually. The inclined elliptic space-time scan statistic improves an individual methodology and therefore the performance of the overall framework, but no direct adjustments are made to how the methodologies are used between each other.

6.3. Further research directions

Several research directions were explored during this thesis, which could be more fully expanded upon in future research. This section addresses these future directions while simultaneously discussing some limitations, complementing those discussed within each research chapter. Three research directions are presented along with analysis on what was initiated and the implications for future research:

- **Section 6.3.1.** The results of applying a uniform filter of varying window sizes are presented and discussed as a methodology for characterising the distance to the movement track variables Δx and Δy (**Section 3.3.3.4**).
- **Section 6.3.2.** A discussion on the various already presented and potential extensions of the point event clustering methodologies (**Section 2.3.1**).
- **Section 6.3.3.** For trajectory clustering methodology Trajectory-OPTICS, the potential extensions to this methodology are presented and discussed.
- **Section 6.3.4.** Further research directions from the research undertaken in this thesis as applicable to other natural hazard processes.

6.3.1. Thunderstorm movement tracking to characterise spatial spread

In **Section 3.3**, one of the selected physical variables for the characterisation study was the spatial spread of lightning strikes with respect to the thunderstorm track, which is not immediately obvious from a dataset of lightning strikes. The exact methodology for identifying the movement track is detailed in **Section 3.3.3.4** where a uniform filter is applied.

The movement track is assumed to be non-linear. It is estimated by applying a uniform linear filter with given window size Δw on the easting/northing (x, y) coordinates (in the OSGB 1936 / British National Grid) of the lightning strike dataset in the order of occurrence k instead of real-time. When window size causes the average calculation at a given successive k to exceed the length of the northing and easting data vectors, the approach uses duplicates of the first or the last values accordingly, introducing an additional uncertainty in the movement track estimate. To illustrate the variations in the estimated movement track based on the window size Δw, the lightning strikes with uniform filters of varying window size Δw values and an orthogonal distance regression (ODR) fit to the lightning strike data are shown for Storm 1A in **Figure 6.1**.

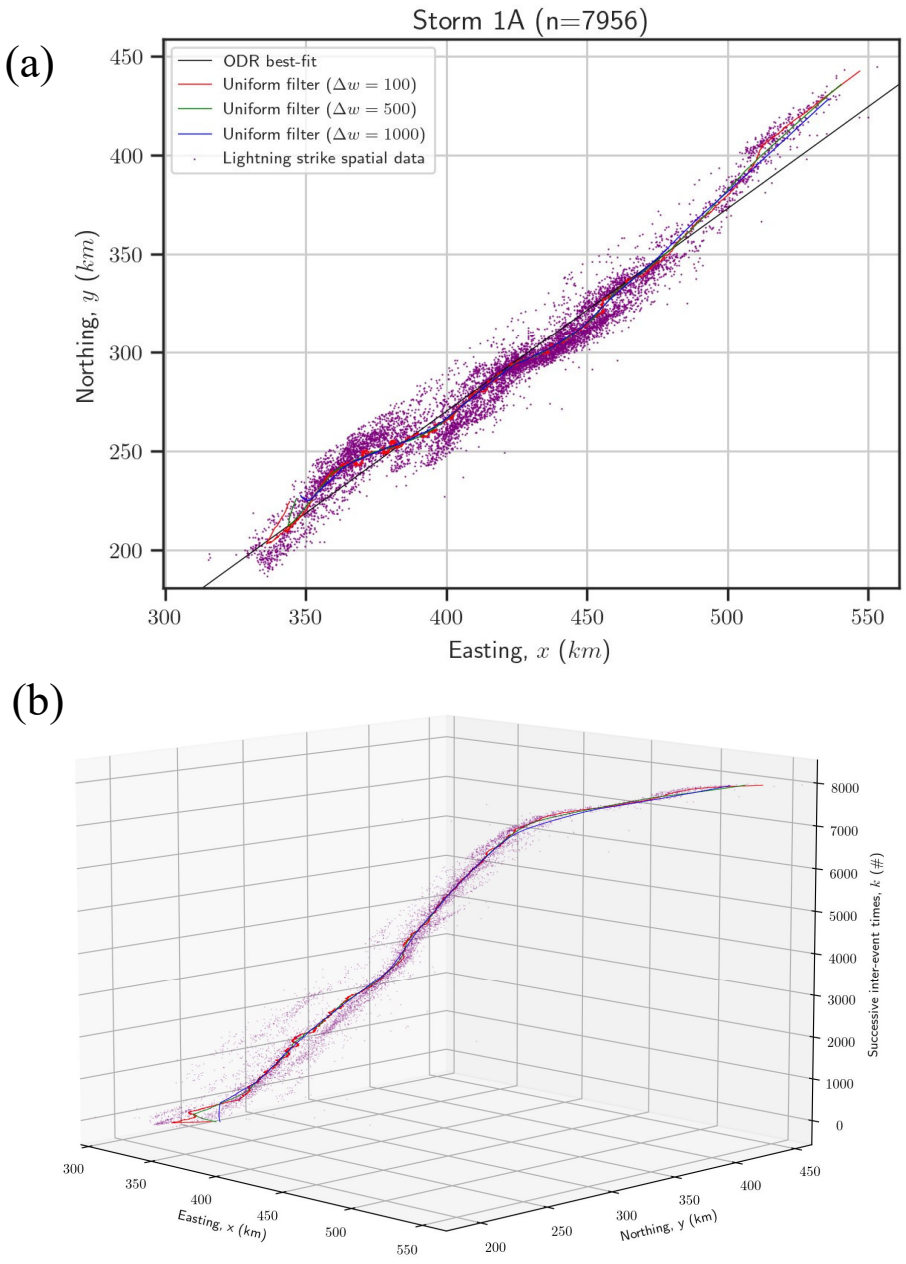

Figure 6.1. Illustration of selecting a movement track. **(a)** 2D spatial view (with events collapsed onto the spatial dimensions) and **(b)** 3D view of the lightning strikes of Storm 1 from storm-system recorded on 28 June 2012 indicated as purple point events. The orthogonal distance regression fit to the point events is indicated by the dark grey line, while the uniform filters with varying window size Δw values are marked as per the legend.

In **Figure 6.1**, the ability of this initial approach to select the movement track is evident as, depending on the window size Δw values, the track estimated by the filter can follow the more dense regions of the storm. As mentioned in the discussion on the limitations of the characterisation approach (**Section 3.3.6**), the movement track estimates are usually done by

temporal windows. In [**Strauss *et al.*, 2013**], additional precipitation data allowed to estimate storm cells and their centroid locations over 10-minute intervals were used as points along the movement track. In [**Kunz *et al.*, 2020**], 26,000 severe convective storm tracks between 2005–2014 were estimated from two-dimensional radar reflectivity data. They estimated the storm tracks by using a modified version of the cell-tracking algorithm TRACE3D [**Handwerker, 2002**]. There is an additional simplification of working with successive point events, therefore losing the information on their temporal structure.

This particular research direction of using the selected uniform filter approach produced reasonable results for the movement track. However, the approach was considered as too complex as the objective of the characterisation process was to provide an estimate for the spatial spread parameter for the models. Nonetheless, a potential further research direction would be a more detailed study on estimating the movement track for the characterisation of spatial spread of the lightning strikes.

6.3.2. K-function spatial and temporal weights

The clustering measure D (**Section 2.2.1**), which was derived from estimating both the separate spatial and temporal and the combined spatio-temporal K-functions, was used to establish performance benchmarks and variations due to different sets of model parameters for a global clustering measure as applied to the point event simulation study in **Section 4.2**. To enable an initial assessment of the scale of spatio-temporal clustering, continuous temporal weights had to be evaluated and their effect assessed to understand the limitations of the current implementation of this methodology. Temporal weights implemented in this thesis are discrete (**Section 2.3.1.1** and **Section 4.2**). When using $\{X = 20, Y = 20, T = 20\}$ space-time volume for Models 1 and 2, if the temporal range from an arbitrary point event is within the space-time cube (temporal dimension $T = 20$), a unit weight (1.0) is assigned. Otherwise, the weight is assigned a value of 2.0. Such a simplification was justified [**Diggle *et al.*, 1995**; **Ripley, 1982**] as spatial edge effects have a more pronounced effect on the results.

Continuous temporal weights were assigned to evaluate this claim further for Models 1 and 2 (point event simulation study, **Section 3.4.1**). Their value is the reciprocal of the proportion of the line segment, formed by the spatial range value extending from a given point event, that

lies within the space-time cube. The results for clustering measure D over the same spatial and temporal range values as in **Section 4.2.1** for Models 1 and 2 are shown in **Figure 6.2**.

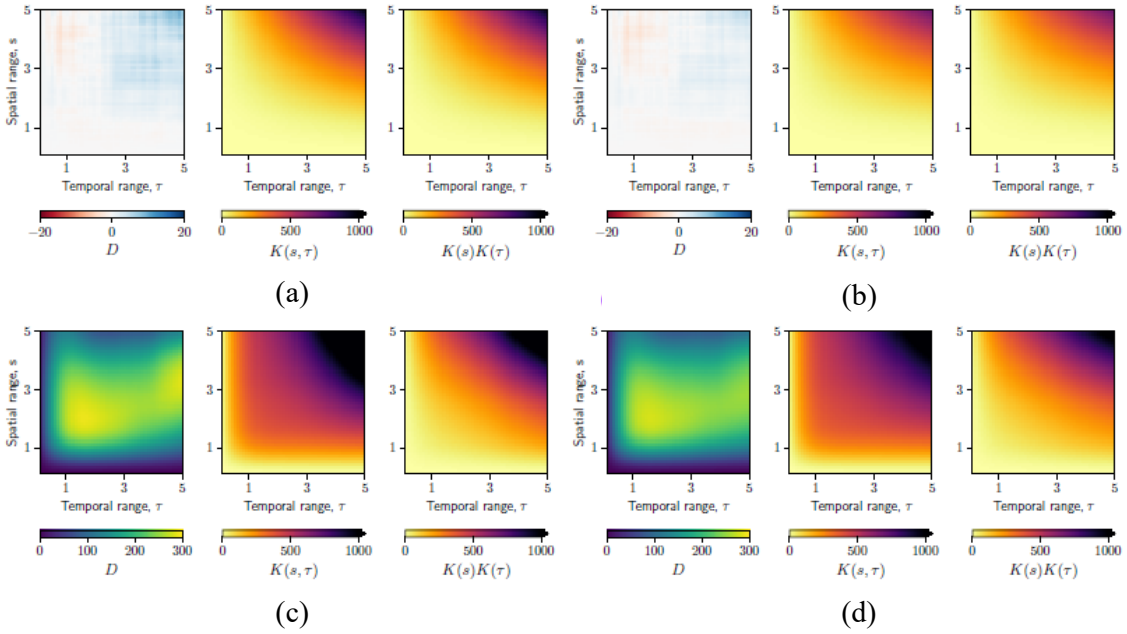

Figure 6.2 Three subplots of the clustering measure $D = K(s,\tau)–K(s)K(\tau)$, spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$ as applied to **[(a),(b)]** homogeneous and **[(c),(d)]** clustered point event datasets. Their values are calculated for various combinations of spatial range s and temporal range τ with **[(a),(c)]** discrete or **[(b),(d)]** continuous temporal weights and are indicated by their respective colour scales. Note the various colour scales for the clustering measure D depending on the dataset.

When compared to the results of the clustering measure D and the K-functions with discrete temporal weights (**Figures 4.1** and **4.2**), introducing continuous temporal weights (**Figures 6.2b** and **6.2d**) has a minor effect. For the homogeneous point event dataset Model 1 (**Figures 6.2a** and **6.2b**), the clustering measure D has values closer to zero as the pattern of positive values for $s > 3$ and $\tau > 3$ is no longer as pronounced. While the distribution patterns of D values for Model 1 is dependent on the particular initialisation, the edge effects have an effect at these spatial and temporal ranges. For the clustered point event dataset Model 2 (**Figures 6.2c** and **6.2d**), the distribution of D values exhibits similar patterns with a relative decrease in its value for larger spatial and temporal range values, adding further evidence for edge effects.

In this thesis, the analysis of the K-function results, depending on the approaches for calculating the spatial and temporal weights, was only done for the case of continuous temporal weights presented in this section. While for Models 3 and 4, the space-time volume was extended to dimensions of $\{X = 1000$ km, $Y = 1000$ km, $T = 24$ h$\}$. As the storm sources are initiated well within this volume, the weight variations would have almost no effect. But this is not the case

with real-world data, where further research would be required to establish the significance level for varying weight calculation approaches.

6.3.3. Trajectory-OPTICS density-based clustering methodology

The trajectory clustering results of the Trajectory-OPTICS methodology (**Figure 4.24** in **Section 4.5**) showed sensitivity to the duration threshold T applied to real-world trajectory data (136 real-world hurricane trajectories from 1998 to 2017 over the North Atlantic basin with alignment at $t = 0$). Depending on this threshold, there is a trade-off between the completeness of the information on the propagation of the trajectories and the size of the dataset. The threshold is introduced due to the spatio-temporal distance used as the distance measure for the current methodology implementation. This spatio-temporal distance (**Appendix Equation E.1**) calculates the average Euclidean distance over all time steps and requires the trajectories to be of the same duration (set as $T = 7.5$ d in **Section 4.5**) and have equally spaced points in the temporal domain. This further research direction can be separated into three different approaches on how to attempt to solve this particular issue with the Trajectory-OPTICS methodology.

Approach 1, one-to-many alignment of trajectories with varying duration: A duration threshold is not applied to the trajectories and they are aligned. One-to-many alignment allows a single point along one trajectory to be aligned with multiple points along another. This property allows for complete matching between two trajectories as all points are aligned. An example of this approach is the dynamic time warping (hereafter, DTW) approach [**Salvador and Chan, 2007**; **Kruskal *et al.*, 1999**]. The implementation of DTW as a distance function between two trajectories X and Y with lengths $|X| = n$ and $|Y| = m$, without considering improvements on computational complexity, can be summarised in the following steps:

A. **Calculate the cost for alignments**: the first (X_1 and Y_1) and the last (X_n and Y_m) points are aligned. There are $n \times m$ combinations of aligning the points between the two trajectories. For any given combination of the points (i, j) for $i = 1, ..., n$ and $j = 1, ..., m$ along the trajectories, the warping distance D_W is calculated for $D_W(1,1)$ as the Euclidean distance and then for any subsequent combination of points as

$$D_W(i,j) = d_{ED}(X_i, Y_i) + \min[D_W(i-1,j), D_W(i,j-1), D_W(i-1,j-1)] \qquad (6.1)$$

where d_{ED} denotes the Euclidean distance between the given points. For the aligned last points, $D_W(n, m)$ represents the sum of Euclidean distances along the warping (i.e. alignment) paths.

B. **Find the optimal warping path:** to add meaningful interpretation for the warping distance, the optimal warping path is found. The optimal warp path is calculated starting from $D_W(n,m)$. Three values are evaluated: $D_W(n-1, m)$, $D_W(n, m-1)$ and $D_W(n-1, m-1)$. Whichever of these three has the smallest D_W value is then added to the beginning of the warp path, and the search continues until $D_W(1,1)$ is reached.

C. **Normalise the warping path:** the final value $D(n, m)$ is normalised by the length (i.e. number of point combinations) of the optimal warping path.

Approach 2, **one-to-one alignment of trajectories with varying duration**: While the duration threshold is not applied and trajectories are aligned, one-to-one alignment does not allow for a single point of one trajectory to be aligned with multiple points along another. One-to-one alignment discards some points along the trajectories as outliers, creating a common subsequence between the trajectories. An example of this approach is the longest common subsequence (hereafter, LCSS) similarity measure [**Vlachos *et al.*, 2006**].

Approach 3, **distance measures for trajectories with duration threshold**: A duration threshold is still applied, but a different distance measure is used. Depending on the real-world trajectory dataset that is used, the approaches vary. For the real-world hurricane trajectory dataset (**Section 3.2.2**), the time steps between consecutive points along the hurricane trajectories are equal. A potential approach is to use a distance function that would not directly measure some distance between individual points along the trajectories, but incorporate additional information about the line segments between them. An example of this approach is the line segment distance function as used in the subtrajectory clustering methodology TRACLUS [**Lee *et al.*, 2007**]. This function has three components that describe the similarity between two line segments in a two-dimensional spatial domain: perpendicular, parallel and angle distances. A visual representation of all three components between two line segments is presented in **Figure 6.3**.

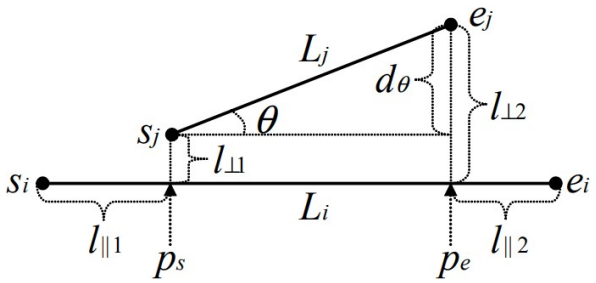

Figure 6.3. The three components of the distance function used in measuring separation between two line segments L_i (defined by two points in spatial domain (s_i, e_i) occurring at the same point in time) and L_j (defined by points (s_j, e_j) occurring at the same point in time) in two–dimensional space: perpendicular distance $d_\perp$, parallel distance d_l and angle distance d_θ. Figure from **[Lee *et al.*, 2007]** (Copyright © 2007 ACM).

The final expression for the distance function $dist(L_i, L_j)$ between two line segments L_i and L_j can be expressed as a linear combination of the components with coefficients $\{w_\perp, w_{||}, w_\theta\}$

$$dist(L_i, L_j) = w_\perp d_\perp + w_{||} d_{||} + w_\theta d_\theta \tag{6.2}$$

While the coefficients for the components can be set as $\{w_\perp = 1, w_{||} = 1, w_\theta = 1\}$ for most applications [**Lee *et al.*, 2007**], implementation results of various sets of coefficients will be considered. For the scope of this thesis, spatio-temporal clustering of subtrajectories was not considered as a research objective and the TRACLUS methodology itself was not implemented. Without implementation of the full methodology, the distance function as given in **Equation 6.2** can still be applied to all line segments from the real-world hurricane trajectories. Instead of the spatio-temporal distance function as defined for the implementation of Trajectory- OPTICS, an average of the distance function from **Equation 6.2** over all line segments can be used for trajectories of equal duration.

6.3.4. Further applications of thesis research to natural hazard processes

The research undertaken in this thesis consists of various methodologies and their extensions applied to two types of geographical data – point events and trajectories (**Section 1.1**). For the point event data, two lightning strike case studies over the UK were characterised and the hazard process modelled in **Chapter 3**, identifying various physical variables representing any point event dataset generated by a moving source. The spatio-temporal clustering methodology framework (**Section 2.5**) was applied to the aforementioned characterised lightning strike data in **Chapter 4**.

Similarly, for the trajectory data type, a density-based spatio-temporal clustering methodology Trajectory-OPTICS (**Section 2.3.2.1**) was applied to synthetic trajectory (**Section 3.4.2**) and North-Atlantic hurricane track (**Section 3.2.2**) datasets. The spatio-temporal clusterings results and analyses from **Chapter 4** and the analysed improvements on these methodologies as presented in **Chapter 5** can be applied not only to lightning and hurricane track data but also to other natural hazard processes such as tornado touchdown locations, which can be expressed as either point event or trajectory data type.

From these research results, there are five major research directions proposed here that can use these results to advance research for natural hazard processes:

1. **Spatio-temporal clustering methodology framework**.
   This thesis extended the research on global and local clustering measure categories in the literature [e.g. **Tonini *et al.*, 2022**; **Wang *et al.*, 2021**; **Tonini and Cama, 2019**; **Tonini *et al.*, 2017**], defining and justifying the combined use of these clustering measure categories on point event data (**Section 2.5**). Such a combined approach is not particular for any specific natural hazard process but can be taken for any point event dataset. Global clustering measures within this framework can provide additional analysis on the global spatial and temporal scales of clustering present in the data. These analyses can then inform local measures, improving their computational complexity and accuracy. As an example of a real-world application of such a clustering methodology framework for point event data, global clustering measures can be applied to tornado touchdown location point data as they are point events produced by a moving source. These global clustering measures characterise physical variables of the natural hazard process and their value ranges. From a tornado touchdown location point event dataset, variables such as tornado track spatial length and temporal duration can be characterised, providing an initial geophysical characterisation of the dataset and indicating the relevant spatial and temporal scales for further identification of individual clusters. Local clustering measures can then indicate particular space-time volumes with higher density of touchdown locations.
2. **Characterisation of point events produced by a moving source**.
   In **Section 3.3**, a total of three physical variables of the lightning strike natural hazard process were characterised – inter-event time, movement speed and spatial spread. These physical variables, while characterised from lightning strike case study datasets in particular, can be characterised from any natural hazard point event dataset produced

by a moving source and extended for other natural hazards (e.g. hail observations, tornado touchdown locations).

3. **Spatio-temporal model of point events produced by a moving source**.

   In **Section 3.3.5,** a modelling framework based on the physical variable characterisation results was presented. This model is an extension of work in [**Finke, 1999**] by explicitly including these physical variables. As for characterisation, this model is not restricted to being applied to just lightning strike data, rather it includes input parameters (i.e. characterised physical variables) that can characterise any natural hazard point event dataset produced by a moving source. As an atmospheric related natural hazard, hail can be modelled in an identical approach by including the same three characterised physical variables. The values of the variables will vary and additional parameters like hail size can be added to the modelling framework.

4. **Density-based clustering analysis of trajectories**.

   In **Sections 4.3** and **4.5**, the density-based clustering methodology Trajectory-OPTICS was applied to synthetic and real-world North Atlantic hurricane track datasets respectively. The methodology was evaluated, examining the sensitivity of its input parameters. When applied to real-world data, the methodology identified two distinct clusters with similar propagation speeds and directions, but varying areas of origin. Further research is necessary to provide geophysical interpretation to these two clusters. It is noted that for the trajectory data type, there were no identified global clustering measures, which is a potential research direction for the trajectory data type.

5. **Inclined elliptic space-time scan statistic**.

   In **Chapter 4**, the elliptic scan statistic did not identify individual point events clusters of individual storms for Model 4 (and variations 4.1, 4.2, 4.3), identifying simply the most point-event dense space-time volume. The inclination aspect of the scanning volume geometry was presented in **Chapter 5**. The improved methodology was able to now correctly identify individual clusters containing point events even with a small temporal separation between them. As with research directions 2 and 3, the inclination of the space-time scan statistic can identify the most-dense cluster volumes for any natural hazard point event dataset produced by a moving source. This could be extended by considering additional geometries of the space-time volume. An example is given in **Figure 5.1**, a space-time volume with changing height value, capturing the persistence of the natural hazard process. For a natural hazard process such as earthquakes, the main shock and aftershock sequences could be captured by the current implementation

of the inclined elliptic scanning volume and allow for earthquakes to persist at a given spatial region with the changing scanning volume height volume.

The clustering methodology application, variable characterisation and modelling frameworks presented in this thesis and their potential extensions discussed in this chapter, when applied to a given hazard or other phenomena over a given region and period of time allows us to further examine and understand various natural hazard processes. Commonalities or differences for this given natural hazard or other phenomena with other regions or periods of time can be found, aiding in understanding what the underlying processes might be. The underlying behaviour can be characterised with the existing physical variables and additional equations of the physical processes from which we might model these processes to produce similar statistics to what has been observed not only for the same natural hazard process, but for similar natural hazards to potentially investigate multi-hazard interactions. This could add further complexity to the production of synthetic benchmark natural hazard datasets.

6.4. Concluding remarks

This thesis (i) synthesises into a common framework and applies spatio-temporal clustering analysis methodologies and (ii) develops spatio-temporal models to reproduce observed natural hazard clustering patterns. **Chapter 2** defined and extended the research on global and local clustering measure categories in [**Tonini *et al.*, 2022**; **Wang *et al.*, 2021**; **Tonini and Cama, 2019**; **Tonini *et al.*, 2017**], focusing on the combined use of these measures in a common spatio-temporal clustering methodology application framework. **Chapter 3** defined spatio-temporal point events models representative of lightning strikes. Three physical variables were used as input parameters to produce these models: inter-event time, movement speed and distance to the movement track. These models are an extension of work in [**Finke, 1999**] by explicitly including lightning strike physical variables into the model. **Chapter 4** presented the analyses of the common framework results after applying them to a simulation study consisting of various synthetic datasets generated by the lightning strike model. In the simulation study, the elliptic space-time scan statistic methodology [**Kulldorff *et al.*, 2006**] could not identify point events clusters produced by a moving source due to the relatively small temporal separation. **Chapter 5** extended this methodology to incorporate additional inclination parameters in its input parameter set, resulting in immediate improvements and additional physical interpretation of the results.

References

[Adelfio and Ogata, 2010] Adelfio, G., and Ogata, Y. (2010). Hybrid kernel estimates of space–time earthquake occurrence rates using the epidemic-type aftershock sequence model. Annals of the Institute of Statistical Mathematics, 62, 127-143.

[Adepeju *et al.*, 2016] Adepeju, M., Rosser, G., and Cheng, T. (2016). Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions-a crime case study. International Journal of Geographical Information Science, 30(11):2133–2154.

[Aftergood and Flannigan, 2022] Aftergood, O. S., and Flannigan, M. D. (2022). Identifying and analyzing spatial and temporal patterns of lightning-ignited wildfires in Western Canada from 1981 to 2018. *Canadian Journal of Forest Research*, *5*2(11), 1399-1411.

[Anderson and Klugmann, 2014] Anderson, G. and Klugmann, D. (2014). A european lightning density analysis using 5 years of atdnet data. Nat. Hazards Earth Syst. Sci, 14(4):815–829.

[Ankerst *et al.*, 1999] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. In ACM Sigmod record, volume 28, pages 49–60. ACM.

[Artigas, 2012] Artigas, S. E. S. (2012). Stochastic modeling of lightning occurrence by nonhomogeneous poisson process. In 2012 International Conference on Lightning Protection (ICLP) (pp. 1-7). IEEE.

[Assuncao *et al.*, 2006] Assuncao, R., Costa, M., Tavares, A., and Ferreira, S. (2006). Fast detec- tion of arbitrarily shaped disease clusters. Statistics in medicine, 25(5):723–742.

[Atluri *et al.*, 2018] Atluri, G., Karpatne, A., and Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR), 51(4), 1-41.

[Ballarotti *et al.*, 2012] Ballarotti, M. G., Medeiros, C., Saba, M. M., Schulz, W., and Pinto Jr, O. (2012). Frequency distributions of some parameters of negative downward lightning flashes based on accurate-stroke-count studies. Journal of Geophysical Research: Atmospheres, 117(D6).

[Beirle *et al.*, 2014] Beirle, S., Koshak, W., Blakeslee, R., and Wagner, T. (2014). Global patterns of lightning properties derived by OTD and LIS. Natural Hazards and Earth System Sciences, 14(10), 2715-2726.

[Berdahl, 2016] Berdahl, J. S. (2016). *Geological Effects on Lightning Strike Distributions* (Doctoral dissertation).

[Besag, 1977] Besag, J. E. (1977). Contribution to the discussion on Dr. Ripley's paper. JR Stat. Soc. Ser, 39, 193-195.

[Blender *et al.*, 1997] Blender, R., Fraedrich, K., and Lunkeit, F. (1997). Identification of cyclone- track regimes in the north atlantic. Quarterly Journal of the Royal Meteorological Society, 123(539):727–741.

[Brunsdon *et al.*, 2007] Brunsdon, C., Corcoran, J., and Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. Computers, environment and urban systems, 31(1):52–75.

[Camargo *et al.*, 2007] Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P., and Ghil,
M. (2007). Cluster analysis of typhoon tracks. part i: General properties. Journal of Climate, 20(14):3635–3653.

[Chowdhuri *et al.*, 2005] Chowdhuri, P., Anderson, J. G., Chisholm, W. A., Field, T. E., Ishii, M., Martinez, J. A., ... and Short, T. A. (2005). Parameters of lightning strokes: a review. IEEE Transactions on Power Delivery, 20(1), 346-358.

[Chudova *et al.*, 2003] Chudova, D., Gaffney, S., Mjolsness, E., and Smyth, P. (2003). Translation- invariant mixture models for curve clustering. In Proceedings of the ninth ACM SIGKDD inter- national conference on Knowledge discovery and data mining, pages 79–88.

[Costafreda-Aumedes *et al.*, 2016] Costafreda-Aumedes, S., Comas, C., and Vega-Garcia, C. (2016). Spatio-temporal configurations of human-caused fires in Spain through point patterns. Forests, 7(9), 185.

[Dale *et al.*, 2002] Dale, M. R., Dixon, P., Fortin, M.-J., Legendre, P., Myers, D. E., and Rosenberg, M. S. (2002). Conceptual and mathematical relationships among methods for spatial analysis. Ecography, 25(5):558–577.

[Demšar and Virrantaus, 2010] Demšar, U. and Virrantaus, K. (2010). Space–time density of tra- jectories: exploring spatio-temporal patterns in movement data. International Journal of Geo- graphical Information Science, 24(10):1527–1542.

[Diggle *et al.*, 1995] Diggle, P. J., Chetwynd, A. G., Häggkvist, R., and Morris, S. E. (1995). Second-order analysis of space-time clustering. Statistical methods in medical research, 4(2):124– 136.

[Duczmal and Assuncao, 2004] Duczmal, L. and Assuncao, R. (2004). A simulated annealing strat- egy for the detection of arbitrarily shaped spatial clusters. Computational Statistics & Data Analysis, 45(2):269–286.

[Dwyer and Uman, 2014] Dwyer, J. R., and Uman, M. A. (2014). The physics of lightning. Physics Reports, 534(4), 147-241.

[Eckert *et al.*, 2010] Eckert, N., Parent, E., Kies, R., and Baya, H. (2010). A spatio-temporal modelling framework for assessing the fluctuations of avalanche occurrence resulting from climate change: application to 60 years of data in the northern French Alps. Climatic change, 101(3), 515-553.

[Economou *et al.*, 2014] Economou, T., Stephenson, D. B., and Ferro, C. A. (2014). Spatio-temporal modelling of extreme storms. The Annals of Applied Statistics, 8(4), 2223-2246.

[Ellis *et al.*, 2016] Ellis, K. N., Trepanier, J. C., and Hodges, R. E. (2016). Using synthetic tropical cyclones to characterize extreme hurricanes affecting charleston, south carolina. Journal of Applied Meteorology and Climatology, 55(4):883–892.

[Emanuel *et al.*, 2006] Emanuel, K., Ravela, S., Vivant, E., and Risi, C. (2006). A statistical deterministic approach to hurricane risk assessment. Bulletin of the American Meteorological Society, 87(3):299–314.

[Emanuel *et al.*, 2008] Emanuel, K., Sundararajan, R., and Williams, J. (2008). Hurricanes and global warming: Results from downscaling ipcc ar4 simulations. Bulletin of the American Me- teorological Society, 89(3):347–368.

[Emanuel, 2013] Emanuel, K. A. (2013). Downscaling cmip5 climate models shows increased trop- ical cyclone activity over the 21st century. Proceedings of the National Academy of Sciences, 110(30):12219–12224.

[Enno *et al.*, 2020] Enno, S. E., Sugier, J., Alber, R., and Seltzer, M. (2020). Lightning flash density in Europe based on 10 years of ATDnet data. Atmospheric Research, 235, 104769.

[Ester *et al.*, 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.* (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, volume 96, pages 226–231.

[Estivill-Castro, 2002] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. SIGKDD explorations, 4(1):65–75.

[Finke, 1999] Finke, U. (1999). Space–time correlations of lightning distributions. Monthly weather review, 127(8), 1850-1861.

[Gaffard *et al.*, 2008] Gaffard, C., Nash, J., Atkinson, N., Bennett, A., Callaghan, G., Hibbett, E., Taylor, P., Turp, M., and Schulz, W. (2008). Observing lightning around the globe from the surface. In the Preprints, 20th International Lightning Detection Conference, Tucson, Arizona, pages 21–23.

[Galanaki *et al.*, 2015] Galanaki, E., Kotroni, V., Lagouvardos, K., and Argiriou, A. (2015). A ten-year analysis of cloud-to-ground lightning activity over the Eastern Mediterranean region. Atmospheric Research, 166, 213-222.

[Gao, 2015] Gao, S. (2015). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. Spatial Cognition & Computation, 15(2):86–114.

[Giannotti *et al.*, 2005] Giannotti, F., Mazzoni, A., Puntoni, S., and Renso, C. (2005). Synthetic generation of cellular network positioning data. In Proceedings of the 13th annual ACM inter- national workshop on Geographic information systems, pages 12–20.

[Gill and Malamud, 2014] Gill, J. C. and Malamud, B. D. (2014). Reviewing and visualizing the interactions of natural hazards. Reviews of Geophysics, 52(4):680–722.

[González *et al.*, 2016] González, J. A., Rodríguez-Cortés, F. J., Cronie, O., and Mateu, J. (2016). Spatio-temporal point process statistics: a review. Spatial Statistics, 18:505–544.

[González *et al.*, 2020] González, J. A., Hahn, U., and Mateu, J. (2020). Analysis of tornado reports through replicated spatiotemporal point patterns. Journal of the Royal Statistical Society: Series C (Applied Statistics), 69(1), 3-23.

[Hallegatte, 2007] Hallegatte, S. (2007). The use of synthetic hurricane tracks in risk analysis and climate change damage assessment. Journal of applied meteorology and climatology, 46(11):1956– 1966.

[Handwerker, 2002] Handwerker, J. (2002). Cell tracking with TRACE3D—A new algorithm. Atmospheric Research, 61(1), 15-34.

[Hering *et al.*, 2009] Hering, A. S., Bell, C. L., and Genton, M. G. (2009). Modeling spatio-temporal wildfire ignition point patterns. Environmental and Ecological Statistics, 16, 225-250.

[Hernandez-Magallanes and Genton, 2019] Hernandez-Magallanes, I., & Genton, M. G. (2019). A point process analysis of cloud-to-ground lightning strikes in urban and rural Oklahoma areas. *Environmetrics*, *30*(1), e2535.

[Hu *et al.*, 2018] Hu, Y., Wang, F., Guin, C., and Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. Applied geography, 99:89–97.

[Ivezić *et al.*, 2014] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. (2014). Statistics, data mining, and machine learning in astronomy. In Statistics, Data Mining, and Machine Learning in Astronomy. Princeton University Press.

[Iyengar, 2004] Iyengar, V. S. (2004). On detecting space-time clusters. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 587–592. ACM.

[Jaquet *et al.*, 2012] Jaquet, O., Lantuéjoul, C., and Goto, J. (2012). Probabilistic estimation of long-term volcanic hazard with assimilation of geophysics and tectonic data. Journal of volcanology and geothermal research, 235, 29-36.

[Jacquez, 1996] Jacquez, G. M. (1996). A k nearest neighbour test for space–time interaction. Statistics in medicine, 15(18), 1935-1949.

[Kisilevich *et al.*, 2009] Kisilevich, S., Mansmann, F., Nanni, M., and Rinzivillo, S. (2009). Spatio- temporal clustering. In Data mining and knowledge discovery handbook, pages 855–874. Springer.

[Knox and Bartlett, 1964] Knox, E. and Bartlett, M. (1964). The detection of space-time interac- tions. Journal of the Royal Statistical Society. Series C (Applied Statistics), 13(1):25–30.

[Kotroni and Lagouvardos. 2016] Kotroni, V., and Lagouvardos, K. (2016). Lightning in the Mediterranean and its relation with sea-surface temperature. Environmental Research Letters, 11(3), 034006.

[Kruskal *et al.*, 1999] Kruskal, J. B., Liberman, M., and Kruskal, J. (1999). The symmetric time- warping problem: from continuous to discrete.

[Kulldorff, 1997] Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics-Theory and methods, 26(6):1481–1496.

[Kulldorff *et al.*, 1998] Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A., and Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico. American journal of public health, 88(9):1377–1380.

[Kulldorff *et al.*, 2005] Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., and Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. PLoS medicine, 2(3):e59.

[Kulldorff *et al.*, 2006] Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. Statistics in medicine, 25(22):3929–3943.

[Kunz *et al.*, 2020] Kunz, M., Wandel, J., Fluck, E., Baumstark, S., Mohr, S., & Schemm, S. (2020). Ambient conditions prevailing during hail events in central Europe. Natural Hazards and Earth System Sciences, 20(6), 1867-1887.

[Landsea and Franklin, 2013] Landsea, C. W. and Franklin, J. L. (2013). Atlantic hurricane database uncertainty and presentation of a new database format. Monthly Weather Review, 141(10):3576–3592.

[Lang, 2001] Lang, P. (2001, July). Cell tracking and warning indicators derived from operational radar products. In Proceedings of the 30th International Conference on Radar Meteorology, Munich, Germany (pp. 245-247).

[Lee *et al.*, 2007] Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: a partition- and-group framework. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 593–604.

[Leśniak and Isakow, 2009] Leśniak, A. and Isakow, Z. (2009). Space–time clustering of seismic events and hazard assessment in the zabrze-bielszowice coal mine, poland. International Journal of Rock Mechanics and Mining Sciences, 46(5):918–928.

[Lewis and Silkstone, 2017] Lewis, M. W. and Silkstone, N. (2017). Improvements in nowcasting capability: analysis of three structurally distinct severe thunderstorms across northern england on 1 july 2015. Weather, 72(4):91–98.

[Li and Racine, 2007] Li, Q. and Racine, J. S. (2007). Nonparametric econometrics: theory and practice. Princeton University Press.

[Li *et al.*, 2011] Li, X.-Z., Wang, J.-F., Yang, W.-Z., Li, Z.-J., and Lai, S.-J. (2011). A spatial scan statistic for multiple clusters. Mathematical biosciences, 233(2):135–142.

[Long and Nelson, 2013] Long, J. A. and Nelson, T. A. (2013). A review of quantitative methods for movement data. International Journal of Geographical Information Science, 27(2):292–318.

[Lovejoy *et al.*, 1986] Lovejoy, S., Schertzer, D., & Ladoy, P. (1986). Fractal characterization of inhomogeneous geophysical measuring networks. Nature, 319, 43-44.

[Malamud *et al.*, 2016] Malamud, B. D., Turcotte, D. L., and Brooks, H. E. (2016). Spatial–temporal clustering of tornadoes. Natural Hazards and Earth System Sciences, 16(12):2823– 2834.

[Mantel, 1967] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer research, 27(2_Part_1), 209-220.

[McNally and Colver, 2008] McNally, R. J., and Colver, A. F. (2008). Space-time clustering analyses of occurrence of cerebral palsy in Northern England for births 1991 to 2003. Annals of epidemiology, 18(2), 108-112.

[McNally *et al.*, 2009] McNally, R. J., Ducker, S., & James, O. F. (2009). Are transient environmental agents involved in the cause of primary biliary cirrhosis? Evidence from space–time clustering analysis. Hepatology, 50(4), 1169-1174.

[Mezuman *et al.*, 2014] Mezuman, K., Price, C., and Galanti, E. (2014). On the spatial and temporal distribution of global thunderstorm cells. Environmental Research Letters, 9(12), 124023.

[Mohee and Miller, 2010] Mohee, F. M., and Miller, C. (2010). Climatology of thunderstorms for North Dakota, 2002–06. Journal of Applied Meteorology and Climatology, 49(9), 1881-1890.

[Møller and Diaz-Avalos, 2010] Møller, J., and Díaz-Avalos, C. (2010). Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires. Scandinavian Journal of Statistics, 37(1), 2-25.

[Nakaya and Yano, 2010] Nakaya, T. and Yano, K. (2010). Visualising crime clusters in a space- time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. Transactions in GIS, 14(3):223–239.

[Nanni and Pedreschi, 2006] Nanni, M. and Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems, 27(3):267–289.

[Naus, 1965] Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association, 60:532–538.

[Orozco *et al.*, 2012] Vega Orozco, C., Tonini, M., Conedera, M., and Kanveski, M. (2012). Cluster recognition in spatial-temporal sequences: the case of forest fires. Geoinformatica, 16, 653-673.

[Palma *et al.*, 2008] Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. In Proceedings of the 2008 ACM symposium on Applied computing, pages 863–868. ACM.

[Parente *et al.*, 2016] Parente, J., Pereira, M. G., and Tonini, M. (2016). Space-time clustering analysis of wildfires: The influence of dataset characteristics, fire prevention policy decisions, weather and climate. Science of the total environment, 559, 151-165.

[Patil and Taillie, 2004] Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological statistics, 11(2):183–197.

[Pereira *et al.*, 2015] Pereira, M. G., Caramelo, L., Orozco, C. V., Costa, R., and Tonini, M. (2015). Space-time clustering analysis performance of an aggregated dataset: The case of wildfires in Portugal. Environmental Modelling & Software, 72, 239-249.

[Perry *et al.*, 2006] Perry, G. L., Miller, B. P., and Enright, N. J. (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. Plant ecology, 187(1):59–82.

[Priestley, 1981] Priestley, M. B. (1981). Spectral analysis and time series: Univariate series (Vol. 1). Academic press.

[Rakov and Uman, 2003] Rakov, V. A., and Uman, M. A. (2003). Lightning: physics and effects. Cambridge university press.

[Rakov, 2013] Rakov, V. A. (2013). The physics of lightning. Surveys in Geophysics, 34(6), 701-729.

[Richard *et al.*, 1986] Richard, P., Delannoy, A., Labaune, G., and Laroche, P. (1986). Results of spatial and temporal characterisation of the VHF-UHF radiation of lightning. Journal of Geophysical Research: Atmospheres, 91(D1), 1248-1260.

[Ripley, 1976] Ripley, B. D. (1976). The second-order analysis of stationary point processes. Journal of applied probability, 13(2):255–266.

[Ripley, 1982] Ripley, B. D. (1982). Edge effects in spatial stochastic processes. Statistics in theory and practice. Essays in honour of Bertil Matern.

[Saba *et al.*, 2006] Saba, M. M. F., Ballarotti, M. G., and Pinto Jr, O. (2006). Negative cloud-to-ground lightning properties from high-speed video observations. Journal of Geophysical Research: Atmospheres, 111(D3).

[Salvador and Chan, 2007] Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5), 561-580.

[Schulz *et al.*, 2005] Schulz, W., Cummins, K., Diendorfer, G., and Dorninger, M. (2005). Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system. Journal of Geophysical Research: Atmospheres, 110(D9).

[Serra *et al.*, 2014] Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., and Rue, H. (2014). Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008. Environmental and ecological statistics, 21(3), 531-563.Intelligent Data Analysis, 11(5):561–580.

[Shao and Krehbiel, 1996] Shao, X. M., and Krehbiel, P. R. (1996). The spatial and temporal development of intracloud lightning. Journal of Geophysical Research: Atmospheres, 101(D21), 26641-26668.

[Soriano *et al.*, 2005] Soriano, L. R., De Pablo, F., and Tomas, C. (2005). Ten-year study of cloud-to-ground lightning activity in the Iberian Peninsula. Journal of Atmospheric and Solar-Terrestrial Physics, 67(16), 1632-1639.

[Stall *et al.*, 2009] Stall, C. A., Cummins, K. L., Krider, E. P., and Cramer, J. A. (2009). Detecting multiple ground contacts in cloud-to-ground lightning flashes. Journal of Atmospheric and oceanic technology, 26(11), 2392-2402.

[Steiger *et al.*, 2002] Steiger, S. M., Orville, R. E., and Huffines, G. (2002). Cloud-to-ground lightning characteristics over Houston, Texas: 1989–2000. Journal of Geophysical Research: Atmospheres, 107(D11), ACL-2.

[Strauss *et al.*, 2013] Strauss, C., Rosa, M. B., and Stephany, S. (2013). Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events. Atmospheric research, 134:87–99.

[Tango and Takahashi, 2005] Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. International journal of health geographics, 4(1):11.

[Taszarek *et al.*, 2015] Taszarek, M., Czernecki, B., and Kozioł, A. (2015). A cloud-to-ground lightning climatology for Poland. Monthly weather review, 143(11), 4285-4304.

[Taylor *et al.*, 2001] Taylor, C. C., Dryden, I. L., and Farnoosh, R. (2001). The k-function for nearly regular point processes. Biometrics, 57(1):224–231.

[Telesca *et al.*, 2008] Telesca, L., Bernardi, M., and Rovelli, C. (2008). Time-scaling analysis of lightning in Italy. Communications in nonlinear science and numerical simulation, 13(7), 1384-1396.

[Tonini *et al.*, 2017] Tonini, M., Pereira, M. G., Parente, J., and Orozco, C. V. (2017). Evolution of forest fires in portugal: from spatio-temporal point events to smoothed density maps. Natural Hazards, 85(3):1489–1510.

[Tonini and Cama, 2019] Tonini, M. and Cama, M. (2019). Spatio-temporal pattern distribution of landslides causing damage in switzerland. Landslides, 16(11):2103–2113.

[Tonini *et al.*, 2022] Tonini, M., Pecoraro, G., Romailler, K., amd Calvello, M. (2022). Spatio-temporal cluster analysis of recent Italian landslides. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(3), 536-554.

[Tuia *et al.*, 2008] Tuia, D., Ratle, F., Lasaponara, R., Telesca, L., and Kanevski, M. (2008). Scan statistics analysis of forest fire clusters. Communications in Nonlinear Science and Numerical Simulation, 13(8), 1689-1694.

[UNISDR, 2009] UNISDR (2009). 2009 unisdr terminology on disaster risk reduction.

[Van Lieshout and Stein, 2012] van Lieshout, M. N. M., and Stein, A. (2012). Earthquake modelling at the country level using aggregated spatio-temporal point processes. Mathematical geosciences, 44(3), 309-326.

[Vlachos *et al.*, 2006] Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., and Keogh, E. (2006). Indexing multidimensional time-series. The VLDB Journal, 15(1):1–20.

[Wang and Anderson, 2010] Wang, Y., and Anderson, K. R. (2010). An evaluation of spatial and temporal patterns of lightning-and human-caused forest fires in Alberta, Canada, 1980–2007. *International Journal of Wildland Fire*, 19(8), 1059-1072.

[Wang *et al.*, 2021] Wang, N., Lombardo, L., Tonini, M., Cheng, W., Guo, L., and Xiong, J. (2021). Spatiotemporal clustering of flash floods in a changing climate (China, 1950–2015). Natural Hazards and Earth System Sciences, 21(7), 2109-2124.

[Wapler and James, 2015] Wapler, K., & James, P. (2015). Thunderstorm occurrence and characteristics in Central Europe under different synoptic conditions. Atmospheric Research, 158, 231-244.

[Wiegand and Moloney, 2004] Wiegand, T. and Moloney, K. (2004). Rings, circles, and null-models for point pattern analysis in ecology. Oikos, 104(2):209–229.

[Xu and Schoenberg, 2011] Xu, H., and Schoenberg, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. The Annals of Applied Statistics, 5(2A), 684-704.

[Yair *et al.*, 2009] Yair, Y. Y., Aviv, R., and Ravid, G. (2009). Clustering and synchronisation of lightning flashes in adjacent thunderstorm cells from lightning location networks data. Journal of Geophysical Research: Atmospheres, 114(D9).

[Yair *et al.*, 2014] Yair, Y., Shalev, S., Erlich, Z., Agrachov, A., Katz, E., Saaroni, H., ... and Ziv, B. (2014). Lightning flash multiplicity in eastern Mediterranean thunderstorms. Natural Hazards and Earth System Sciences, 14(2), 165-173.

[Yuan *et al.*, 2017] Yuan, G., Sun, P., Zhao, J., Li, D., and Wang, C. (2017). A review of moving object trajectory clustering algorithms. Artificial Intelligence Review, 47(1):123–144.

[Zaliapin *et al.*, 2008] Zaliapin, I., Gabrielov, A., Keilis-Borok, V., and Wong, H. (2008). Clustering analysis of seismicity and aftershock identification. Physical review letters, 101(1), 018501.

[Zaliapin and Ben-Zion, 2013] Zaliapin, I., and Ben-Zion, Y. (2013). Earthquake clusters in southern California I: Identification and stability. Journal of Geophysical Research: Solid Earth, 118(6), 2847-2864.

[Zhang *et al.*, 2010] Zhang, Z., Assunção, R., and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters. Journal of Probability and Statistics, 2010.

[Zhang *et al.*, 2011] Zhang, Z., Chen, D., Liu, W., Racine, J. S., Ong, S., Chen, Y., Zhao, G., and Jiang, Q. (2011). Nonparametric evaluation of dynamic disease risk: a spatio-temporal kernel approach. PloS one, 6(3).

[Zhang *et al.*, 2013] Zhang, J. L., Bi, W., Wang, X. H., Wang, Z. B., and Li, D. F. (2013). Lightning-caused fire, its affecting factors and prediction: A review. *Ying Yong Sheng tai xue bao= The Journal of Applied Ecology*, *24*(9), 2674-2684.

[Zheng, 2015] Zheng, Y. (2015). Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology (TIST), 6(3):29.

Appendix A. Critically reviewed publications

This appendix consists of the results of a critical literature review (as described in **Section 2.2**) of spatio-temporal clustering methodologies. **Appendix Table A1** presents those 24 journals with 2 or more publications included in the final critical literature review database of 143 publications (**Appendix Table A2**).

Table A1. Scientific journals with multiple critically reviewed publications.

Journal	# publications
Journal of Geophysical Research	6
PLOS One	4
Environmental and Ecological Statistics	3
ACM Sigmod Record	2
Bulletin of the Seismological Society of America	2
Tectonophysics	2
Journal of the American Statistical Association	2
Physical Review Letters	2
International Biometric Society	2
Statistical Methods in Medical Research	2
Natural Hazards	2
Natural Hazards and Earth System Sciences	2
Environmental Modelling & Software	2
Journal of the Royal Statistical Society	2
Computational Statistics and Data Analysis	2
International Journal of Geographical Information Science	2
Annals of the Institute of Statistical Mathematics	2
Applied Geography	2
Preventive Veterinary Medicine	2
Statistics in Medicine	2
Journal of Quantitative Criminology	2
Ecology	2
The International Journal on Very Large Data Bases	2
Computers, Environment and Urban Systems	2
[88 journals with 1 publication only]	88
Total	**143**

Table A2. Critical literature review database of 143 publications on spatio-temporal clustering methodologies as described in **Section 2.2**. The below is a partial inclusion of the full critical review with the following categories not included from **Table 2.1**: "Pre-Processing", "Dataset Embedding", "Clustering Shape", "Clustering separation", "Magnitude Threshold", "Region", "Period". "Clustering interpretation" below indicates whether the authors provide an interpretation of the clustering results ('+' for 'Yes', '–' for 'No').

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
1	Spatio-temporal clustering	S. Kisilevich, F. Mansmann, M. Nanni, S. Rinzivillo	Data Mining and Knowledge Discovery Handbook	2010	Statistics	-	Review	-	–
2	Trajectory Clustering with Mixtures of Regression Models	S. Gaffney, P. Smyth	Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery in Data Mining	1999	Statistics	-	Mixture of regression models	-	–
3	Translation-Invariant Mixture Models for Curve Clustering	D. Chudova, S. Gaffney, E. Mjolsness, P. Smyth	Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery in Data Mining	2003	Statistics	Various	Mixture of regression models	Both	+
4	Discovering Clusters in Motion Time-Series Data	J. Alon, S. Sclaroff, G. Kollios	CAS: Computer Science: Technical Reports	2003	Movement Science	-	Mixture of hidden Markov models	Both	+
5	OPTICS: Ordering Points To Identify the Clustering Structure	M. Ankerst, M. M. Breunig, H. P. Kriegel, J. Sander	ACM Sigmod Record	1999	Statistics	-	Density-based	-	–
6	Time-focused clustering of trajectories of moving objects	M. Nanni, D. Pedreschi	Journal of Intelligent Information Systems	2006	Statistics	-	Density-based	Synthetic	–
7	Discovery of Convoys in Trajectory Databases	H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, H. T. Shen	Proceedings of the VLDB Endowment	2008	Statistics	Various	Density-based	Real	–
8	Visually driven analysis of movement data by progressive clustering	S. Rinzivillo, D. Pedreschi, M. Nanni, F. Gianotti, N. Andrienko, G. Andrienko	Information Visualization	2008	Statistics	-	Visual-aided	Real	+
9	A review of moving object trajectory clustering algorithms	G. Yuan, P. Sun, J. Zhao, D. Li, C. Wang	Artifical Intelligence Review	2016	Statistics	-	Review	-	–
10	ST-DBSCAN: An algorithm for clustering spatial-temporal data	D. Birant, A. Kut	Data & Knowledge Engineering	2006	Statistics	Climate	Density-based	Real	+
11	Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980-1995	J. Le Gallo, C. Ertur	Papers in Regional Science	2003	Economics	Economics	Local indicators of spatial association	Real	+
12	Epidemiological Analysis, Detection, and Comparison of Space-Time	J. Wang, Z. Cao, D. D. Zeng, Q. Wang, X. Wang, H. Qian	PLOS One	2014	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	Patterns of Beijing Hand-Foot-Mouth Disease (2008?2012)								
13	Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico	M. Kulldorff, W. F. Athas, B. A. Miller, C. R. Key	American Journal of Public Health	1998	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
14	Actions as Space-Time Shapes	L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri	IEEE Transactions on Pattern Analysis and Machine Intelligence	2007	Movement Science	Movement Science	Interest point operators	Synthetic	–
15	Upper level set scan statistic for detecting arbitrarily shaped hotspots	G. P. Patil, C. Taillie	Environmental and Ecological Statistics	2004	Statistics	Epidemiology	Maximum likelihood ratio	Real	–
16	Multiple Spatio-Temporal Cluster Detection for Case Event Data: An Ordering-Based Approach	C. Demattei, L. Cucala	Communications in Statistics - Theory and Methods	2010	Statistics	Earthquakes	Ordering-based spatio-temporal method	Real	–
17	Space-time clustering of childhood malaria at the household level: a dynamic cohort in a Mali village	J. Gaudart, B. Poudiougou, A. Dicko, S. Ranque, O. Toure, I. Sagara, M. Diallo, S. Diawara, A. Ouattara, M. Diakite, O. K. Doumbo	BMC Public Health	2006	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
18	Space-time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland	A. Lesniak, Z. Isakow	International Journal of Rock Mechanics & Mining Sciences	2009	Earthquakes	Earthquakes	Mean minimum distance clustering	Real	+
19	Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics	T. Nakaya, K. Yano	Transactions in GIS	2010	Statistics	Crime	Maximum likelihood ratio and space-time kernel estimation	Real	+
20	Space-Time Clustering of Large Thrust Earthquakes along the Mexican Subduction Zone: An Evidence of Source Stress Interaction	M. A. Santoyo, S. K. Singh, T. Mikumo, M. Ordaz	Bulletin of the Seismological Society of America	2005	Geophysics	Earthquakes	x2 goodness-of-fit test	Real	+
21	Refining earthquake clustering models	R. Console, M. Murru, A. M. Lombardi	Journal of Geophysical Research	2003	Geophysics	Earthquakes	Space-time brancing process (ETAS) model	Real	+
22	Identifying space-time clustering properties of the 1983?1997 Irpinia-Basilicata	L. Telesca, V. Cuomo, V. Lapenna, M. Macchiato	Tectonophysics	2000	Geophysics	Earthquakes	Fractal tools	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	Southern Italy seismicity								
23	A Space-Time Permutation Scan Statistic for Disease Outbreak Detection	M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao, F. Mostashari	PLoS Medicine	2005	Epidemiology	Epidemiology	Poisson generalized likelihood ratio (GLR)	Real	+
24	Spatial-Temporal Clusters and Risk Factors of Hand, Foot, and Mouth Disease at the District Level in Guangdong Province, China	T. Deng, Y. Huang, S. Yu, J. Gu, C. Huang, C. Xiao, Y. Hao	PLoS One	2013	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
25	Stochastic Declustering of Space-Time Earthquake Occurrences	J. Zhuang, Y. Ogata, D. Vere-Jones	Journal of the American Statistical Association	2002	Statistics	Earthquakes	Space-time brancing process (ETAS) model	Real	+
26	Long-range correlations in two-dimensional spatio-temporal seismic fluctuations	L. Telesca, M. Lovallo, V. Lapenna, M. Macchiato	Physica	2007	Geophysics	Earthquakes	Detrended fluctuation analysis	Real	+
27	Measuring multifractality in seismic sequences	L. Telesca, V. Lapenna	Tectonophysics	2006	Geophysics	Earthquakes	Multifractal detrended fluctuation analysis	Real	+
28	Spatial variability of the time-correlated behaviour in Italian seismicity	L. Telesca, V. Lapenna, M. Macchiato	Earth and Planetary Science Letters	2003	Geophysics	Earthquakes	Detrended fluctuation analysis	Real	+
29	Space-Time Clustering and Correlations of Major Earthquakes	J. R. Holliday, J. B. Rundle, D. L. Turcotte, W. Klein, K. F. Tiampo, A. Donnellan	Physical Review Letters	2006	Physics	Earthquakes	Peirce difference function	Real	–
30	Spatial and space-time clustering of childhood acute leukaemia in France from 1990 to 2000: a nationwide study	S. Bellec, D. Hemon, J. Rudant, A. Goubin, J. Clavel	British Journal of Cancer	2006	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
31	Research The spatial-temporal clustering of Plasmodium falciparum infection over eleven years in Gezira State, The Sudan	S. E. Mirghani, B. Y. M. Nour, S. M. Bushra, I. M. Elhassan, R. W. Snow, A. M. Noor	Malaria Journal	2010	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
32	Exploratory space-time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991-1992	A. C. Morrison, A. Getis, M. Santiago, J. G. Rigau-Perez, P. Reiter	The American journal of tropical medicine and hygiene	1998	Epidemiology	Epidemiology	Knox test	Real	+
33	Testing Separability in Spatial-Temporal Marked Point Processes	F. P. Schoenberg	International Biometric Society	2004	Statistics	Wildfires	Cramer-von Mises-type statistic	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
34	Self-Exciting Point Process Modeling of Crime	G. O. Mohler, M. B. Short, P. J. Brantingham	Journal of the American Statistical Association	2011	Statistics	Crime	Variable bandwidth kernel density estimation	Real	+
35	Point process methodology for on-line spatio-temporal disease surveillance	P. J. Diggle, B. S. Rowlingson, T. Su	Environmetrics	2005	Epidemiology	Epidemiology	Posterior exceedance probabilities	Both	–
36	Partial-Likelihood Analysis of Spatio-Temporal Point-Process Data	P. J. Diggle, I. Kaimi, R. Abellana	International Biometric Society	2010	Epidemiology	Epidemiology	Partial likelihood approach	Both	–
37	Spatio-temporal point processes, partial likelihood, foot and mouth disease	P. J. Diggle	Statistical Methods in Medical Research	2006	Epidemiology	Epidemiology	Partial likelihood approach	Real	–
38	Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST	G. Falcone, R. Console, M. Murru	Annals of Geophysics	2010	Geophysics	Earthquakes	Purely stochastic epidemic type earthquake sequence (ETES) model, Epidemic rate-state (ERS) forecast and Long-term stress transfer (LTST) model	Real	–
39	Large Earthquake Triggering, Clustering, and the Synchronization of Faults	C. H. Scholz	Bulletin of the Seismological Society of America	2010	Geophysics	Earthquakes	Kuramoto synchronisation model	-	–
40	A constitutive law for rate of earthquake production and its application to earthquake clustering	J. Dieterich	Journal of Geophysical Research	1994	Geophysics	Earthquakes	Normalised number of pairs	Real	+
41	A mechanism for spatial and temporal earthquake clustering	E. A. Jagla, A. B. Kolton	Journal of Geophysical Research	2010	Geophysics	Earthquakes	Gutenberg?Richter and Omori laws	Both	+
42	Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology	A. C. Gatrell, T. C. Bailey, P. J. Diggle, B. S. Rowlingson	Transactions of the Institute of British Geographers	1996	Epidemiology	Epidemiology	Difference between space-time and absence of space-time interaction K-functions	Real	+
43	A comparison of methods for the statistical analysis of spatial point patterns in plant ecology	G. L. W. Perry, B. P. Miller, N. J. Enright	Plant Ecology	2006	Ecology & Environment	Ecology	Review	Both	–
44	Spatial-temporal characteristics of temperature variation in China	W. Qian, A. Qin	Meteorology and Atmospheric Physics	2006	Climate	Climate	T-test score	Real	+
45	Potential source zones for Himalayan earthquakes: constraints from spatial?temporal clusters	B. Mukhopadhyay, A. Acharyya, S. Dasgupta	Natural Hazards	2011	Geophysics	Earthquakes	Moving time-distance window technique	Real	+
46	Cluster recognition in spatial-temporal	C. V. Orozco, M. Tonini, M.	Geoinformatica	2012	Wildfires	Wildfires	Poisson generalized	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	sequences: the case of forest fires	Conedera, M. Kanveski					likelihood ratio (GLR)		
47	RainSim: A spatial?temporal stochastic rainfall modelling system	A. Burton, C. G. Kilsby, H. J. Fowler, P. S. P. Cowpertwait, P. E. O'Connell	Environmental Modelling & Software	2008	Climate	Climate	Spatio-temporal Neyman?Scott Rectangular Pulses model	-	–
48	Analyzing earthquake clustering features by using stochastic reconstruction	J. Zhuang, Y. Ogata	Journal of Geophysical Research	2004	Geophysics	Earthquakes	Stochastic reconstruction of ETAS model	Both	–
49	Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes	A. M. Lombardi, W. Marzocchi	Journal of Geophysical Research	2007	Geophysics	Earthquakes	Declustering of ETAS model and K-mean cluster analysis algorithm for non-stationary ETAS model	Real	+
50	Monitoring Point Patterns for the Development of Space-Time Clusters	P. A. Rogerson	Journal of the Royal Statistical Society	2001	Statistics	Epidemiology	Local Knox statistic with cumulative sum	Real	–
51	Nonparametric Approach to Analysis of Space Time Data on Earthquake Occurrences	E. Choi, P. Hall	Journal of Computational and Graphical Statistics	1999	Statistics	Earthquakes	Coupled form of near-neighbour method with linked smoothing parameters	Real	+
52	Surveillance to detect emerging space-time clusters	R. Assun‡?oa, T. Correa	Computational Statistics and Data Analysis	2009	Statistics	Epidemiology	Maximum likelihood ratio	Both	–
53	Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise	T. Pei, C. Zhou, A. X. Zhu, B. Li, C. Qin	International Journal of Geographical Information Science	2011	Statistics	Earthquakes	Windowed nearest neighbour (WNN) method	Both	+
54	Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town	L. H. Duczmal, G. J. P. Moreira, D. Burgarelli, R. H. C. Takahashi, F. C. O. Magalhaes, E. C. Bodevan	International Journal of Health Geographics	2011	Epidemiology	Epidemiology	Weighted complete graph	Both	+
55	Hybrid kernel estimates of space-time earthquake occurrence rates using the epidemic-type aftershock sequence model	G. Adelfio, Y. Ogata	Annals of the Institute of Statistical Mathematics	2010	Statistics	Earthquakes	Back-transforming the estimated intensity	Real	–
56	Novel evaluation metrics for sparse spatiotemporal point process hotspot predictions - a crime case study	M. Adepeju, G. Rosser, T. Cheng	International Journal of Geographical Information Science	2016	Statistics	Crime	Review	Real	–
57	Spatio-temporal analysis of wildfire ignitions in the St Johns River Water	M. G. Genton, D. T. Butry, M. L. Gumpertz, J. P. Prestemon	International Journal of Wildland Fire	2006	Wildfires	Wildfires	Relative clustering index	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	Management District, Florida								
58	Spatio-temporal hazard estimation in the Auckland Volcanic Field, New Zealand, with a new event-order model	M. S. Bebbington, S. J. Cronin	Bulletin of Volcanology	2011	Geophysics	Earthquakes	Age-order model	Real	+
59	Modeling spatio-temporal wildfire ignition point patterns	A. S. Hering, C. L. Bell, M. G. Genton	Environmental and Ecological Statistics	2009	Statistics	Wildfires	L-function	Both	+
60	A web-based geospatial toolkit for the monitoring of dengue fever	E. M. Delmelle, H. Zhu, W. Tang, I. Casas	Applied Geography	2014	Epidemiology	Epidemiology	Accelerated kernel density estimation	Real	–
61	Spatio-temporal clustering of overall and serovar-specific Leptospira microscopic agglutination test seropositivity among dogs in the United States	R. Gautam, L. F. Guptill. C. C. Wu, A. Potter, G. E. Moore	Preventive Veterinary Medicine	2010	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
62	Clustering of suicides among people with mental illness	N. McKenzie, S. Landau, J. Robinson, H. Bickley	British Journal of Psychiatry	2005	Epidemiology	Epidemiology	Knox test	Real	+
63	Spatial-temporal clustering of tornadoes	B. D. Malamud, D. L. Turcotte, H. E. Brooks	Natural Hazards and Earth System Sciences	2016	Geophysics	Tornadoes	Non-dimensional velocity ratio	Both	+
64	Clustering analysis of seismicity and aftershock identification	I. Zaliapin, A. Gabrielov, V. Keilis-Borok, H. Wong	Physical Review Letters	2008	Physics	Earthquakes	Bimodal distribution of nearest-neighbour distance	Both	+
65	Earthquake clusters in southern California I: Identification and stability	I. Zaliapin, Y. Ben-Zion	Journal of Geophysical Research	2013	Geophysics	Earthquakes	Spanning network	Both	+
66	A model-adjusted space-time scan statistic with an application to syndromic surveillance	K. P. Kleinman, A. M. Abrams, M. Kulldorff, R. Platt	Epidemiology and Infection	2005	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
67	Second-order analysis of space-time clustering	P.J. Diggle, A. G. Chetwynd, R. Haggkvist, S. E. Morris	Statistical Methods in Medical Research	1995	Statistics	Epidemiology	Difference between space-time and absence of space-time interaction K-functions	Both	+
68	Periodicity and space-time clustering of severe childhood malaria on the coast of Kenya	R. W. Snow, J. R. M. Armstrong Schellenberg, N. Peshu, D .Forster, C. R. J. C. Newton, P. A. Winstanley, I. Mwangi, C. Waruiru, P. A.Warn, C. Newbold, K.Marsh	Transactions of the Royal Society of Tropical Medicine and Hygiene	1993	Epidemiology	Epidemiology	Knox test	Real	+
69	Increased Incidence Rates	U. Hjalmars, M. Kulldorff, Y.	Cancer	1999	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	but No Space-time Clustering of Childhood Astrocytoma in Sweden, 1973?1992	Wahlqvist, B. Lannering							
70	A space-time clustering model for historical earthquakes	F. Musmeci, D. Vere-Jones	Annals of the Institute of Statistical Mathematics	1992	Statistics	Earthquakes	Space-time branching process with immigration	Real	–
71	Are Transient Environmental Agents Involved in the Cause of Primary Biliary Cirrhosis? Evidence from Space-time Clustering Analysis	R. J. Q. McNally, S. Ducker, O. F. W James	Hepatology	2009	Medicine	Medicine	Maximum likelihood ratio	Real	+
72	The stability of space-time clusters of burglary	S. D. Johnson, K. J. Bowers	British Journal of Criminology	2004	Crime	Crime	Mantel's test	Real	+
73	A K-Nearest Neighbor Test for Space-Time Interaction	G. M. Jacquez	Statistics in Medicine	1996	Statistics	Wildfires	k-NN cumulative test statistic	Both	–
74	Analysis of time-space clustering in veterinary epidemiology	M. P. Ward, T. E. Carpenter	Preventive Veterinary Medicine	2000	Epidemiology	Epidemiology	Barton's method	Real	–
75	A Hypothesis-Free Multiple Scan Statistic with Variable Window	L. Cucala	Biometrical Journal	2008	Statistics	Epidemiology	Largest concentration index	Both	–
76	A flexible spatial scan test for case event data	L. Cucala	Computational Statistics and Data Analysis	2009	Statistics	Epidemiology	Union of cluster zones	Both	+
77	On Detecting Space-Time Clusters	V. S. Iyengar	Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery in Data Mining	2004	Statistics	Epidemiology	Heuristic search algorithm	Real	+
78	The Detection of Space-Time Interactions	E. G. Knox, M. S. Bartlett	Journal of the Royal Statistical Society	1964	Statistics	Epidemiology	Knox test	-	–
79	Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors	A. C. Klassen, M. Kulldorff, F. Curriero	International Journal of Health Geographics	2005	Epidemiology	Epidemiology	Maximum likelihood ratio	Real	+
80	An elliptic spatial scan statistic	M. Kulldorff, L. Huang, L. Pickle, L. Duczmal	Statistics in Medicine	2006	Statistics	Epidemiology	Maximum likelihood ratio with eccentricity penalty function	Real	–
81	Space-Time Clustering Analyses of Occurrence of Cerebral Palsy in Northern England for Births 1991 to 2003	R. J. Q. McNally, A. F. Colver	Annals of Epidemiology	2008	Epidemiology	Epidemiology	Maximum likelihood ratio, K-functions, k-NN cumulative test statistic	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
82	Spatial Scan Statistics Adjusted for Multiple Clusters	Z. Zhang, R. Assuncao, and M. Kulldorff	Journal of Probability and Statistics	2010	Statistics	Epidemiology	Maximum likelihood test statistic for multiple clusters	Real	–
83	Multiple Temporal Cluster Detection	N. Molinari, C. Bonaldi, and J.-P. Daures	Biometrics	2001	Statistics	Epidemiology	Stepwise regression function	Both	+
84	Detection of Emerging Space-Time Clusters	D. B. Neill, A. W. Moore, M. Sabhnani, K. Daniel	Proceedings of the eleventh ACM SIGKDD international conference on Knowledge Discovery in Data Mining	2005	Statistics	Epidemiology	Maximum likelihood ratio for count data	Both	–
85	A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation	Y. Hua , F. Wangb , C. Guinc , H. Zhu	Applied Geography	2018	Statistics	Crime	Space-time kernel esimation	Real	–
86	Modelling of the spatio-temporal distribution of rat sightings in an urban environment	I. Tamayo-Uria, J. Mateuc , P. J. Diggle	Spatial Statistics	2014	Statistics	Epidemiology	Difference between space-time and absence of space-time interaction K-functions and partial likelihood approach	Real	+
87	PULSE, progressive upper level set scan statistic for geospatial hotspot detection	G. P. Patil, S. W. Joshi, R. E. Koli	Environmental and Ecological Statistics	2010	Statistics	Ecology & Environment	Maximum likelihood ratio	Real	–
88	Crime Pattern Detection Using Data Mining	S. V. Nath	2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology	2006	Statistics	Crime	Semi-supervised k-means clustering	Real	–
89	Spatio-Temporal Interaction of Urban Crime	T. H. Grubesic, E. A. Mack	Journal of Quantitative Criminology	2008	Crime	Crime	Knox test and k-NN cumulative test statistic	Real	+
90	Improving the Robustness and Accuracy of Crime Prediction with the Self-Exciting Point Process Through Isotropic Triggering	G. Rosser, T. Cheng	Applied Spatial Analysis	2016	Statistics	Crime	Triggering kernel density estimation	Both	+
91	The Afghanistan-Pakistan Wars, 2008-2009: Microgeographies, Conflict Diffusion, and Clusters of Violence	J. O'Loughlin, F. D. W. Witmer, A. M. Linke	Eurasian Geography and Economics	2010	Crime	Crime	Maximum likelihood ratio	Real	+
92	On The Application of Fuzzy Clustering for Crime Hot Spot Detection	T. H. Grubesic	Journal of Quantitative Criminology	2006	Crime	Crime	Generalized partitioning (fuzzy clustering)	Real	–

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
93	Crime Data Mining: A General Framework and Some Examples	H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, M. Chau	IEEE Computer Society	2004	Statistics	Crime	Hierarchical clustering	Real	–
94	Modelling and Estimation of Multi-Source Clustering in Crime and Security Data	G. Mohler	The Annals of Applied Statistics	2013	Statistics	Crime	Metropolis adjusted Langevin algorithm	Both	+
95	Scan statistics analysis of forest fire clusters	D. Tuia, F. Ratle, R. Lasaponara, L. Telesca, M. Kanevski	Communications in Nonlinear Science and Numerical Simulation	2008	Statistics	Wildfires	Poisson generalized likelihood ratio (GLR)	Real	–
96	The Impact of Hurricanes on Crime: A Spatio-Temporal Analysis in the City of Houston, Texas	M. Leitner, M. Helbich	Cartography and Geographic Information Science	2011	Statistics	Crime	Poisson generalized likelihood ratio (GLR) and kernel density estimation	Real	+
97	Rings, Circles, and Null-Models for Point Pattern Analysis in Ecology	T. Wiegand, K. A. Moloney	Oikos	2004	Ecology & Environment	Ecology & Environment	L-function and O-ring statistic	Synthetic	+
98	Population Growth in Space and Time: Spatial Logistic Equations	R. Law, D. J. Murrell, U. Dieckmann	Ecology	2003	Ecology & Environment	Ecology & Environment	Dynamical system of spatial logistic equations	Synthetic	–
99	Conceptual and mathematical relationships among methods for spatial analysis	M. R. T. Dale, P. Dixon, M.-J. Fortin, P. Legendre, D. E. Myers, M. S. Rosenber	Econography	2002	Ecology & Environment	Ecology & Environment	Review	-	–
100	American cutaneous leishmaniasis in Southeast Brazil: space-time clustering	G. L. L. Machado-Coelho, R. Assuncao, W. Mayrink, W. T. Caiaffa	International Journal of Epidemiology	1999	Epidemiology	Epidemiology	Knox	Real	+
101	Extending point pattern analysis for objects of finite size and irregular shape	T. Wiegand, W. D. Kissling, P. A. Cipriotti, M. R. Aguiar	Journal of Ecology	2006	Ecology & Environment	Ecology & Environment	Pair correlation function	Real	+
102	Analyzing the spatial structure of a Sri Lankan Tree Species with multiple scales of clustering	T. Wiegand, S. Gunatilleke, N. Gunatilleke, T. Okuda	Ecology	2007	Ecology & Environment	Ecology & Environment	Pair correlation function	Both	+
103	Index-based Most Similar Trajectory Search	E. Frentzos, K. Gratsias, Y. Theodoridis	IEEE 23rd International Conference on Data Enginering	2007	Statistics	-	k-MST (most similar trajectory) search	Synthetic	–
104	DB-SMoT: A Direction-Based Spatio-Temporal Clustering Method	J. A. M. R. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, V. Bogorny	2010 5th IEEE International Conference Intelligent Systems	2010	Statistics	Various	DB-SMoT	Real	–
105	Multivariate Spatio-Temporal Clustering (MSTC) as a Data Mining Tool for Environmental Applications	F. M. Hoffman, W. W. Hargrove, R. T. Mills, S. Mahajan, D. J. Erickson, R. J. Oglesby	Proceedings of the IESMs Fourth Biennial Meeting: International Congress on Environmental	2008	Ecology & Environment	Ecology & Environment	k-means clustering	Real	–

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
			Modeling and Software						
106	Arbitrarily shaped multiple spatial cluster detection for case event data	C. Dematte‹, N. Molinari, Jean-Pierre DaurŠs	Computational Statistics & Data Analysis	2007	Statistics	Various	Test statistic (Bai and Perron)	Both	–
107	Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering	R. C. de Amorim, B. Mirkin	Pattern Recognition	2012	Statistics	-	k-MST (most similar trajectory) search	Synthetic	–
108	Similarity Search in Trajectory Databases	N. Pelekis,ÿI. Kopanakis,ÿG. Marketos, I. Ntoutsi, G. Andrienko. Y. Theodoridis	14th International Symposium on Temporal Representation and Reasoning	2007	Statistics	-	Review	Both	–
109	A clustering-based approach for discovering interesting places in trajectories	A. T. Palma, V. Bogomy, B. Kuijpers, L. O. Alvares	Proceedings of the 2008 ACM symposium on Applied computing	2008	Statistics	Urban Studies	Review	-	–
110	Unveiling the complexity of human mobility by querying and mining massive trajectory data	F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti	The International Journal on Very Large Data Bases	2011	Statistics	Urban Studies	Density-based (OPTICS)	Real	+
111	Trajectory data mining: an overview	Y. Zheng	ACM Transactions on Intelligent Systems and Technology	2015	Statistics	-	Review	-	–
112	Trajectory clustering: a partition-and-group framework	J.-G. Lee, J. Han, K.-Y. Whang	Proceedings of the 2007 ACM SIGMOD international conference	2007	Statistics	Ecology & Environment	TRACLUS	Real	–
113	Spatio-temporal pattern distribution of landslides causing damage in Switzerland	M. Tonini, M. Cama	Landslides	2019	Landslides	Landslides	Poisson generalized likelihood ratio (GLR), K-functions and space-time kernel density estimation	Real	+
114	Bandwidth Selection for Fixed-Kernel Analysis of Animal Utilization Distributions	R. A. Gitzen, J. J. Millspaugh,, B. J. Kernohan	The Journal of Wildlife Management	2006	Ecology & Environment	-	Space-time kernel density estimation	Synthetic	–
115	Evolution of forest fires in Portugal: from spatiotemporal point events to smoothed density maps	M. Tonini, M. Gonzalez Pereira, J. Parente, C. V. Orozco	Natural Hazards	2017	Geophysics	Wildfires	Poisson generalized likelihood ratio (GLR), K-functions and space-time kernel density estimation	Real	+
116	Spatial distribution of wild boar population abundance: Basic information for spatial epidemiology and wildlife management	P. Acevedoa, F. Quirǿs-Fern ndezc, J. Casal, J. Vicente	Ecological Indicators	2014	Ecology & Environment	Ecology & Environment	Mantel's test	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
117	Fractal characterization of the spatial distribution of geological point processes	R. Zuo, F. P. Agterberg, Q. Cheng, L. Yao	International Journal of Applied Earth Observation and Geoinformation	2009	Geophysics	Various	L-function and Fry analysis	Both	–
118	Clustering and aggregating clues of trajectories for mining trajectory patterns and routes	C.-C. Hungl W.-C. Peng, W.-C. Lee	The International Journal on Very Large Data Bases	2015	Statistics	Urban Studies	Clustering and Aggregating Clues of Trajectories (CACT) using OPTICS	Both	–
119	Incremental clustering for trajectories	Z. Li, J. G. Lee, X. Li, J. Han	International Conference on Database Systems for Advanced Applications	2010	Statistics	Various	Density-based	Both	–
120	BIRCH: an efficient data clustering method for very large databases	T. Zhang, R. Ramakrishnan, M. Livny	ACM Sigmod Record	1996	Statistics	Various	BIRCH	Both	–
121	Finding long and similar parts of trajectories	K. Buchin, M. Buchin, M. van Kreveld, J. Luo	Computational Geometry: Theory and Applications	2011	Statistics	-	Density-based	-	–
122	Football analysis using spatio-temporal tools	J. Gudmundsson, T. Wolle	Computers, Environment and Urban Systems	2014	Statistics	Sports	Density-based	Real	–
123	Spatial and Temporal Characterization of Travel Patterns in a Traffic Network Using Vehicle Trajectories	J. Kima, H. S. Mahmassani	Transportation Research Procedia	2015	Urban Studies	Urban Studies	Density-based	Real	–
124	Detecting Commuting Patterns by Clustering Subtrajectories	K. Buchin, M. Buchin, J. Gudmundsson, M. Loffler, J. Luo	International Journal of Computational Geometry & Applications	2011	Statistics	Urban Studies	Density-based	Both	+
125	Clustering Sequences of Complex Objects	A. Ketterlin	KDD-97 Proceedings	1997	Statistics	-	Hierarchical	-	–
126	A review of quantitative methods for movement data	J. A. Long, T. A. Nelson	International Journal of Geographical Information Science	2012	Statistics	-	Review	-	–
127	Space-time clustering analysis performance of an aggregated dataset: The case of wildfires in Portugal	M. G. Pereira, L. Caramelo, C. V. Orozco, R. Costa, M. Tonini	Environmental Modelling & Software	2015	Ecology & Environment	Wildfires	Poisson generalized likelihood ratio (GLR)	Both	+
128	Space-time clusters and co-occurrence of chikungunya and dengue fever in Colombia from 2015 to 2016	M.R. Desjardins, A. Whiteman, I. Casas, E. Delmelle	Acta Tropica	2018	Epidemiology	Epidemiology	Poisson generalized likelihood ratio (GLR)	Real	+
129	Spatial and spatiotemporal clustering methods for detecting elephant poaching hotspots	P. Rashidi, T. Wang, A. Skidmore, A. Vrieling, R. Darvishzadeh, B. Toxopeus, S.	Ecological Modelling	2015	Ecology & Environment	Ecology & Environment	Poisson generalized likelihood ratio (GLR)	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
		Ngene, P. Omondi							
130	Geographic Distribution and Expansion of Human Lyme Disease, United States	K. J. Kugeler, G. M. Farley, J. D. Forrester, P. S. Mead	Emerging Infectious Diseases	2015	Epidemiology	Epidemiology	Poisson generalized likelihood ratio (GLR)	Real	+
131	Detecting hotspots of urban residents? behaviours based on spatio-temporal clustering techniques	P. Zhang, M. Deng, Y. Shi, L. Zhao	GeoJournal	2017	Urban Studies	Urban Studies	Density-based	Real	+
132	An Eigenspace approach for detecting multiple space-time disease clusters: Application to measles hotspots detection in Khyber-Pakhtunkhwa, Pakistan	S. Ullah, H. Daud, S. C. Dass, H. Fanaee-T, A. Khalil	PLOS One	2018	Epidemiology	Epidemiology	Multi-EigenSpot, EigenSpot, Poisson generalized likelihood ratio (GLR)	Real	+
133	Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events	C. Strauss, M. B. Rosa, S. Stephany	Atmospheric Research	2013	Atmospheric Science	Atmospheric Science	Space-time kernel density estimation	Real	+
134	Nonparametric evaluation of dynamic disease risk: A spatio-temporal kernel approach	Z. Zhang, D. Chen, W. Liu., J. S. Racine, S. Ong, Y. Chen, G. Zhao, Q. Jiang	PLOS One	2011	Epidemiology	Epidemiology	Space-time kernel density estimation	Both	–
135	Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age	S. Gao	Spatial Cognition & Computation	2015	Statistics	Urban Studies	Space-time kernel density estimation	Real	+
136	Visualising space and time in crime patterns:A comparison of methods	C. Brunsdon, J. Corcoran, G. Higgs	Computers, Environment and Urban Systems	2007	Statistics	Crime	Space-time kernel density estimation	Real	–
137	Spatio-temporal cluster analysis of recent Italian landslides	M. Tonini , G. Pecoraro , K. Romailler & M. Calvello	Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards	2022	Statistics	Landslides	Poisson generalized likelihood ratio (GLR)	Real	+
138	Space-time clustering analysis of wildfires: The influence of dataset characteristics, fire prevention policy decisions, weather and climate	J. Parente, M. G. Pereira, M. Tonini	Science of the Total Environment	2016	Ecology & Environment	Wildfires	Poisson generalized likelihood ratio (GLR)	Real	+
139	Recent evolution of spatial and temporal patterns of burnt areas and	T. Calheiros, J.P. Nunes, M.G. Pereira	Agricultural and Forest Meteorology	2020	Ecology & Environment	Wildfires	Monthly k-means clustering	Real	+

No.	Title	Authors	Journal	Year	Field of Study	Field of Application	Clustering measure	Applied dataset	Clustering interpretation
	fire weather risk in the Iberian Peninsula								
140	Modelling wildfire activity in Iberia with different atmospheric circulation weather types	R. M. Trigo, P. M. Sousa, M. G. Pereira, D. Rasillad and C. M. Gouveia	International Journal of Climatology	2013	Climate	Wildfires	Monthly k-means clustering	Real	+
141	Spatiotemporal clustering of flash floods in a changing climate (China, 1950–2015)	N. Wang, L. Lombardo, M. Tonini, W. Cheng, L. Guo, and J. Xiong	Natural Hazards and Earth System Sciences	2021	Geophysics	Floods	K-functions, Poisson generalized likelihood ratio (GLR)	Real	+
142	Spatio-Temporal Configurations of Human-Caused Fires in Spain through Point Patterns	S. Costafreda-Aumedes, C. Comas and C. Vega-Garcia	Forests	2022	Ecology & Environment	Wildfires	K-functions	Real	+
143	Detection of clusters using space–time scan statistics	Tonini, M., Tuia, D., and Ratle, F.	International Journal of Wildland Fire	2009	Geophysics	Wildfires	Poisson generalized likelihood ratio (GLR)	Real	+

Appendix B. K-function derivation

For implementing the K-functions, the necessary additional parameters that need to be introduced are listed below.

Table B.1: Variables and their definitions for K-function derivation.

Variable	Definition
s	spatial range value
τ	temporal range value
N	total number of point events
n_i	a given point event for $i = 1, ..., N$
d_{ij}	Euclidean distance between two point events n_i and n_j
u_{ij}	temporal distance between point events n_i and n_j
w_{ij}	weight in the spatial domain for the corresponding d_{ij} value
v_{ij}	weight in the temporal domain for the corresponding u_{ij} value
μ	underlying rate of intensity (number of events per unit volume)
$E[\cdot]$	expected value
$I(\cdot)$	binary indicator function

Intuitively, the spatio-temporal and the individual spatial and temporal K-functions can be expressed as

$$K_{ST}(s,\tau) = \mu^{-1} E\left[\#\text{ of point events within } s \text{ and } \tau \text{ from an arbitrary event}\right] \tag{B.1}$$

$$K_S(s) = (\mu T)^{-1} E\left[\#\text{ of point events within } s \text{ from an arbitrary event}\right] \tag{B.2}$$

$$K_T(\tau) = (\mu XY)^{-1} E\left[\#\text{ of point events within } \tau \text{ from an arbitrary event}\right] \tag{B.3}$$

When calculating the expected number of point events, weights are introduced as both the spatial and temporal distance values might extend past the defined space-time domain of dimensions $\{X, Y, T\}$. If not introduced, the counted number of events will underestimate the true values. Taking this into account and using all the previouly defined variables, the estimators for the three necessary K-functions become

$$K_{ST}(s,\tau) = \frac{XYT}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{ij} v_{ij} I(d_{ij} \leq s) I(u_{ij} \leq \tau) \tag{B.4}$$

$$K_S(s) = \frac{T}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{ij} I(d_{ij} \leq s) \tag{B.5}$$

$$K_T(\tau) = \frac{XY}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} v_{ij} I(u_{ij} \leq \tau) \tag{B.6}$$

The final clustering measure D can then be defined as

$$D = K_{ST}(s,\tau) - K_S(s) K_T(\tau) \tag{B.7}$$

Appendix C. Kernel density estimation derivation

To keep the expression for the kernel density estimation general (i.e. consider each dimension separately, thus not using a bivariate kernel for the two spatial dimensions), the following parameters are introduced below.

Table C.1: Variables and their definitions for kernel density estimation derivation.

Variable	Definition
h_x	bandwidth value of the first spatial dimension
h_y	bandwidth value of the second spatial dimension
h_t	bandwidth value of the temporal dimension
k	kernel function
N	total number of point events
(x_i, y_i, t_i)	coordinates of a given point event i for $i = 1, ..., N$
$\hat{f}(x, y, t)$	density function estimate at a given point event
$E[\cdot]$	expectation value

The expression for the estimate of the probability density function is then

$$\hat{f}(x,y,t) = \frac{1}{Nh_xh_yh_t}\sum_{i=1}^{N} k\left(\frac{x-x_i}{h_x}\right) k\left(\frac{y-y_i}{h_y}\right) k\left(\frac{t-t_i}{h_t}\right) \tag{C.1}$$

The kernel function can be defined in several ways depending on the purpose of application, but here it is given by the Gaussian kernel defined as

$$k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{C.2}$$

While direction estimation is possible by using any set of $\{h_x, h_y, h_t\}$ values, a data-driven way of choosing the optimal set is chosen. To this end, consider minimising the Kullback-Leibler divergence, which measures the difference between the true probability density $f(x, y, t)$ and the one estimated by the kernel density estimation $\hat{f}(x, y, t)$ and is given as

$$d_{KL}(f, \hat{f}) = \int_x \int_y \int_t f(x,y,t) \ln f(x,y,t) dx dy dt - \underbrace{\int_x \int_y \int_t f(x,y,t) \ln \hat{f}(x,y,t) dx dy dt}_{\text{aim to maximise}} \tag{C.3}$$

It can be seen that, in order to minimise $d_{KL}(f, \hat{f})$, the second term must be maximised. Furthermore, this term is just the expectation of $\ln \hat{f}(x, y, t)$. If this is directly maximised by using all the data points, it will return zero values due to the kernel being placed on the point itself. Therefore, the leave-one-out estimator at every point event is selected and given as

$$\hat{f}_{-i}(x_i, y_i, t_i) = \frac{1}{(N-1)h_xh_yh_t} \sum_{j=1, j\neq i}^{N} k\left(\frac{x_i - x_j}{h_x}\right) k\left(\frac{y_i - y_j}{h_y}\right) k\left(\frac{t_i - t_j}{h_t}\right) \tag{C.4}$$

It follows that, when using all the point events, the estimate for the expectation that is to be maximised can be expressed as

$$E\left[\ln \hat{f}(x,y,t)\right] = \frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}_{-i}(x_i, y_i, t_i) \tag{C.5}$$

Appendix D. Space-time scan statistic derivation

For defining the homogeneous and imhomogeneous Poisson processes, the only assumption made for implementation of the scan statistic must be noted. The inhomogeneous Poisson process assumes a smaller volume within the general study area with a higher density of point events, while the homogeneous Poisson process assumes an equal rate of intensity throughout the general study area. This difference then is taken into account when defining the average number of events λ within the space-time volume XYT. To this end, the parameters relevant to this particular methodology are introduced and listed below.

Table D.1: Variables and their definitions for space-time scan statistic derivation.

Variable	Definition
Z	volume of the elliptic cylinder within the space-time volume XYT
μ	underlying rate of intensity (number of events per unit volume)
p	probability that a volume unit **within** volume Z is an 'event'
q	probability that a volume unit **outside** volume Z is an 'event'
N	total number of point events
n_i	a given point event for $i = 1, ..., N$
$f(n_i)$	density function at a given point event
N_Z	number of point events within Z
L_Z	likelihood for the inhomogeneous Poisson process and volume Z
L_0	likelihood for the homogeneous Poisson process

The average number of events within the space-time volume XYT is then

$$\lambda = \underbrace{p\mu Z}_{\substack{\text{average number} \\ \text{of events within Z}}} + \underbrace{q\mu(XYT - Z)}_{\substack{\text{average number} \\ \text{of events outside Z}}} \tag{D.1}$$

It follows that for the assumed higher density of points events within volume Z for the inhomogeneous Poisson process, $p > q$, while for the homogeneous Poisson process $p = q$. Furthermore, the density function at a given point event can be defined as

$$f(n_i) = \begin{cases} \frac{p\mu}{\lambda}, & \text{if } n_i \in Z \\ \frac{q\mu}{\lambda}, & \text{if } n_i \notin Z \end{cases} \tag{D.2}$$

Having defined λ and $f(n_i)$, the likelihood follows as

$$L = \frac{e^{-\lambda}\lambda^N}{N!} \prod_{n_i \in Z} \frac{p\mu}{\lambda} \prod_{n_i \notin Z} \frac{q\mu}{\lambda} = \frac{e^{-\lambda}\lambda^N}{N!} \frac{p^{Nz} q^{(N-Nz)} \mu^N}{\lambda^N} = \frac{e^{-p\mu Z - q\mu(XYT-Z)} p^{Nz} q^{(N-Nz)} \mu^N}{N!} \tag{D.3}$$

For the homogeneous Poisson process, the likelihood expression follows from $p = q$ and the assumption that the total number of point events within the space-time volume can be approximated by the expected average number of events given as

$$N = \lambda_{p=q} = p\mu XYT \Longrightarrow p = \frac{N}{\mu XYT} \tag{D.4}$$

and then setting $p = q$ in the expression for L

$$L_0 = \frac{e^{-p\mu XYT} p^N \mu^N}{N!} = \frac{e^{-N} \left(\frac{N}{\mu XYT}\right)^N \mu^N}{N!} = \frac{e^{-N}\mu^N}{N!} \tag{D.5}$$

For the likelihood expression of the inhomogeneous Poisson process, the values of $\{p, q\}$ that would maximise it can be found as

$$\begin{aligned} \frac{\partial L}{\partial p} &= -\mu Z L + N_Z \frac{L}{p} = 0 \\ N_Z \frac{L}{p} &= \mu Z L \\ p &= \frac{N_Z}{\mu Z} \end{aligned} \tag{D.6}$$

$$\begin{aligned} \frac{\partial L}{\partial q} &= -\mu(XYT - Z)L + (N - N_Z)\frac{L}{p} = 0 \\ (N - N_Z)\frac{L}{p} &= \mu(XYT - Z)L \\ q &= \frac{N - N_Z}{\mu(XYT - Z)} \end{aligned} \tag{D.7}$$

The expression for the inhomogeneous Poisson process then follows as

$$L_Z = \frac{e^{-N}\left(\frac{N_Z}{\mu Z}\right)^{N_Z}\left(\frac{N-N_Z}{\mu(XYT-Z)}\right)^{(N-N_Z)}\mu^N}{N!} I\left(\frac{N_Z}{\mu Z} > 1\right) \tag{D.8}$$

Finally, the likelihood ratio (LR) expression then becomes

$$\begin{aligned} LR &= \frac{L_Z}{L_0} \\ &= \left(\frac{N_Z}{\mu Z}\right)^{N_Z}\left(\frac{N - N_Z}{\mu(XYT - Z)}\right)^{(N-N_Z)} I\left(\frac{N_Z}{\mu Z} > 1\right) \\ &= \left(\frac{N_Z}{\mu Z}\right)^{N_Z}\left(\frac{N - N_Z}{N - \mu Z}\right)^{(N-N_Z)} I\left(\frac{N_Z}{\mu Z} > 1\right) \end{aligned} \tag{D.9}$$

Appendix E. Trajectory-OPTICS implementation

For implementing the Trajectory-OPTICS algorithm, the necessary parameters and their definitions are listed below and further expanded afterwards.

Table E.1: Variables and their definitions for Trajectory-OPTICS implementation.

Variable	Definition
N	total number of trajectories
$\|T\|$	temporal length of trajectories
o_i	a given object (i.e. trajectory) for $i = 1, ..., N$
$d_{ED}(o_i(t), o_j(t))$	Euclidean distance between points at time t between trajectories o_i and o_j
$d(o_i, o_j)$	spatio-temporal distance between trajectories o_i and o_j
ϵ	neighbourhood distance
n_{pts}	minimum number of trajectories within neighbourhood distance ϵ
c_k	core trajectory for $k = 1, ..., K$ where $N \geq K \geq 1$
$d(c_k)$	core distance for core trajectory c_k
$D_R(o_i, o_j)$	reachability distance between trajectories o_i and o_j
ϵ'	reachability distance threshold

The algorithm initially establishes which of the trajectories can be classified as being core trajectories. Core trajectories are important as they are used as central trajectories in denser regions and spatio-temporal distances are measured with respect to them. To this end, the following three definitions are introduced:

Definition 1. Core trajectory - trajectory o_i is defined as a core trajectory if within a spatio-temporal neighbourhood distance ϵ there are at least n_{pts} other trajectories. The spatio-temporal distance used in this implementation is the one defined in [Nanni and Pedreschi, 2006] as

$$d(o_i, o_j) = \frac{\int_T d_{ED}(o_i(t), o_j(t))dt}{|T|} \tag{E.1}$$

Definition 2. Core distance - the spatio-temporal distance between a core trajectory and the trajectory which completes the n_{pts} requirement (e.g. if $n_{pts} = 3$ the core distance is between the core trajectory and the third furthest trajectory).

Definition 3. Reachability distance - the spatio-temporal distance between a core trajectory and any other trajectory unless it is smaller than the core distance, in which case it gets normalised to the core distance.

$$D_R(c_k, o_i) = \begin{cases} d(c_k, o_i), & \text{if } d(c_k, o_i) > d(c_k) \\ d(c_k), & \text{if } d(c_k, o_i) \leq d(c_k) \end{cases} \tag{E.2}$$

To further illustrate the principle of the reachability distance, consider a synthetic example for $n_{pts} = 3$ for a core trajectory c_1 and four other trajectories $\{o_1, o_2, o_3, o_4\}$ in Figure E.1 below. For this example, trajectories are considered as point objects for ease of visualisation.

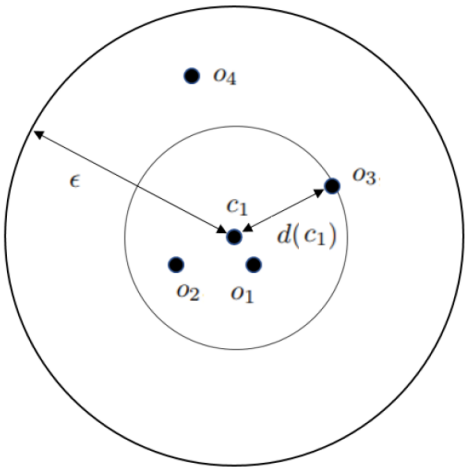

Figure E.1: An example of calculating the reachability distances between a core trajectory c_1 and four other trajectories $\{o_1, o_2, o_3, o_4\}$ given an arbitrary ϵ value and $n_{pts} = 3$. The core distance between c_1 and the third furthest trajectory is given as $d(c_1)$.

To illustrate the reachabiility distance principle, two examplery calculations can be done for $D_R(c_1, o_1)$ and $D_R(c_1, o_4)$. As $d(c_1, o_1) < d(c_1)$, the spatio-temporal distance is normalised to the core distance and thus $D_R(c_1, o_1) = d(c_1)$. For o_4, the spatio-temporal distance $d(c_1, o_4) > d(c_1)$ and thus $D_R(c_1, o_4) = d(c_1, o_4)$. This normalisation allows for relatively denser regions to not have a significant influence on the ordered reachability distance plot discussed in the next paragraph.

Having defined the process of identifying core trajectories and calculating reachability distance, the next step is the ordering process, which is detailed in the following steps:

1. *Creation of visited, unvisited and visited core trajectory lists* - an important aspect of the algorithm is the iteration through unvisited trajectories, thus three distinct lists are defined with all trajectories initially sorted into the unvisited list.

2. *Initial trajectory* - after establishing which trajectories are core, the algorithm selects a trajectory at random. This continues until this selected trajectory is core, at which point it can be used as a reference point for measuring reachability distances. This trajectory is then added to the visited core trajectory list. Previously selected trajectories that were not core are still left in the unvisited list.

3. *Finding the closest trajectory* - after establishing the first core trajectory, the trajectory with the smallest reachability distance is selected and the distance value stored. It is noted whether this trajectory is also a core trajectory and if so, transferred to the visited core trajectory list.

4. *Iterating through the dataset* - the previous step is continuously repeated and the returned reachability distance values stored. As the list of visited core trajectories grows, the reachability distances can be measured to any of them.

5. *Identifying sequences of low reachability distance values* - after all the trajectories are visited and their reachability distance values with respect to previously visited core trajectories are stored, the resultant reachability distance plot is visually examined. A threshold ϵ' on the reachability distance plot is set, separating the noise ($D_R \geq \epsilon'$) from the clustered ($D_R < \epsilon'$) trajectories. The clustered trajectories can further be separated into individual clusters if they are separated in the ordering by a non-clustered trajectory.

Appendix F. Synthetic data generator algorithms

Algorithm 1 Synthetic Point Event Data Generator

1) Regular grid data

Inputs:

1. $\{X, Y, T\}$ - the dimension lengths of the space-time rectangular prism.
2. s - spacing between consecutive iterations.

Outputs:

1. (x_i, y_i, t_i) for $i = 1, ..., N$ positions of the N point events.

set $(x_1, y_1, t_1) = (1, 1, 1)$ as the initial point event.
set $(n_x, n_y, n_t) = \left(\frac{X-1}{s}, \frac{Y-1}{s}, \frac{T-1}{s}\right)$ as the spatial and temporal steps.
for $j = 0, ..., n_x$ **do**
    for $k = 0, ..., n_y$ **do**
        for $m = 1, ..., n_t$ **do**
            store $(x_1 + js, y_1 + ks, t_1 + ms)$
        end for
    end for
end for

2) Homogeneous Poisson point process

Inputs:

1. $\{X, Y, T\}$ - the dimension lengths of the space-time rectangular prism.
2. λ - number of events per unit volume.

Outputs:

1. (x_i, y_i, t_i) for $i = 1, ..., N$ positions of the N point events.

set $N = \lambda XYT$ as the total number of point events.
for $i = 1, ..., N$ **do**
    draw x_i from $\mathcal{U}[0, X]$
    draw y_i from $\mathcal{U}[0, Y]$
    draw t_i from $\mathcal{U}[0, T]$
    store (x_i, y_i, t_i)
end for

3) Thomas point process

Inputs:

1. $\{X, Y, T\}$ - the dimension lengths of the space-time rectangular prism.
2. κ - number of parent events per unit volume.
3. σ - standard deviation of a Gaussian distribution.
4. γ - number of children events.

Outputs:

1. (x_k, y_k, t_k, p_k) for $k = 1, ..., \kappa\gamma XYT$ - positions and parent memberships of the $\kappa\gamma XYT$ point events.

set $P = \kappa XYT$ as the total number of parent events.
for $i = 1, ..., P$ **do**
    draw x_i from $\mathcal{U}[0, X]$
    draw y_i from $\mathcal{U}[0, Y]$
    draw t_i from $\mathcal{U}[0, T]$
    for $j = 1, ..., \gamma$ **do**
        draw x_{ij} from $\mathcal{N}(0, \sigma^2)$
        draw y_{ij} from $\mathcal{N}(0, \sigma^2)$
        draw t_{ij} from $\mathcal{N}(0, \sigma^2)$
        if $0 \leq x_{ij} \leq X \wedge 0 \leq y_{ij} \leq Y \wedge 0 \leq t_{ij} \leq T$ **then**
            store $(x_i + x_{ij}, y_i + y_{ij}, t_i + t_{ij})$
        else
            continue
    end for
end for

Algorithm 2 Synthetic Trajectory Data Generator

1) Trajectories with homogeneous IPs
Inputs:

1. $(x_{i1}, y_{i1}, 0)$ for $i = 1, ..., N$ - initial positions of the N trajectories.
2. n_t - number of time steps to be taken between $[0, T]$.
3. $\mu = [\mu_0, \mu_1]$ - interval of drift parameter values.
4. $\sigma = [\sigma_0, \sigma_1]$ - interval of standard deviation parameter values.

Outputs:

1. (x_{ij}, y_{ij}, t_j) for $i = 1, ..., N$ and $j = 1, ..., n_t$ - positions of the N trajectories over n_t time steps.

set $\delta t = \frac{T}{n_s}$ as the time step.
for $i = 1, ..., N$ **do**
  draw μ_k from $\mathcal{U}_\mu[\mu_0, \mu_1]$
  draw σ_k from $\mathcal{U}_\sigma[\sigma_0, \sigma_1)]$
  for $j = 1, \ldots, n_t$ **do**
    draw δx from $\mathcal{N}(0, 1)$
    draw δy from $\mathcal{N}(0, 1)$
    store $(x_{ij} + \delta x \sigma_k \sqrt{\delta t} + \mu_k \delta t, y_{ij} + \delta y \sigma_k \sqrt{\delta t} + \mu_k \delta t, 0 + j\delta t)$
  end for
end for

2) Trajectories with clustered IPs
Inputs:

1. $(x_{i1}, y_{i1}, 0, c_i)$ for $i = 1, ..., N$ - initial positions and cluster memberships of the N trajectories.
2. n_t - number of time steps to be taken between $[0, T]$.
3. $\mu = [\mu_0, \mu_1]$ - interval of drift parameter values.
4. $\sigma = [\sigma_0, \sigma_1]$ - interval of standard deviation parameter values.

Outputs:

1. $(x_{ij}, y_{ij}, t_j, c_i)$ for $i = 1, ..., N$ and $j = 1, ..., n_t$ - positions and cluster memberships of the N trajectories over n_t time steps.

set $\delta t = \frac{T}{n_s}$ as the time step.
set $C = \{c_i\}_{i \in \{1,...,N\}}$ as the set of distinct cluster membership numbers.
for $k = 1, \ldots, |C|$ **do**
  set $I = \{i\}$ for $i = 1, ..., N$ if $c_i = C_k$
  draw μ_k from $\mathcal{U}_\mu[\mu_0, \mu_1]$
  draw σ_k from $\mathcal{U}_\sigma[\sigma_0, \sigma_1]$
  for $\forall m \in I$ **do**
    for $j = 1, \ldots, n_t$ **do**
      draw δx from $\mathcal{N}(0, 1)$
      draw δy from $\mathcal{N}(0, 1)$
      store $(x_{mj} + \delta x \sigma_k \sqrt{\delta t} + \mu_k \delta t, y_{mj} + \delta y \sigma_k \sqrt{\delta t} + \mu_k \delta t, 0 + j\delta t, C_k)$
    end for
  end for
end for

Appendix G. List of publications on modelling natural hazards as point events

Table G1. Summary table of publications on using spatio-temporal point process models to model natural hazard processes.

No.	Title	Authors	Natural hazard	Mathematical model (extended description with how attributes are incorporated into the model)
1	Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008	Laura Serra, Marc Saez, Jorge Mateu, Diego Varga, Pablo Juan, Carlos Díaz-Ávalos, Håvard Rue	Wildfire	A log-Gaussian Cox process with its log-intensity function estimated from a linear predictor consisting of terms for expected number of events, covariates and their parameters and interaction between events.
2	Spatial–temporal clustering of tornadoes	Bruce D. Malamud, Donald L. Turcotte, Harold E. Brooks	Tornado	The number of lines and the spatial and temporal spacings between consecutive events are estimated from real-world data.
3	Structured Spatio-Temporal Shot-Noise Cox Point Process Models, with a View to Modelling Forest Fires	Jesper Møller, Carlos Díaz-Ávalos	Wildfire	A shot-noise Cox process with its intensity function inferred.
4	Spatio-temporal models for large-scale indicators of extreme weather	Matthew J. Heatona, Matthias Katzfuss, Shahla Ramachandar, Kathryn Pedings, Eric Gilleland, Elizabeth Mannshardt-Shamseldin, Richard L. Smith	Extreme weather event	A non-homogeneous Poisson process with the intensity function having parameters of a generalised extreme value distribution.
5	Is Timber Insurable? A Study of Wildfire Risks in the U.S. Forest Sector Using Spatio-temporal Models	Xuan Chen, Barry K. Goodwin, Jeffrey P. Prestemon	Wildfire	Various.
6	A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications	Alex Reinhart	Earthquake	Various modification of the self-exciting point process models are discussed.
7	Spatio-temporal point process statistics: A review	Jonatan A. González, Francisco J. Rodríguez-Cortés, Ottmar Cronie, Jorge Mateu	Tornado	For real-world data summary statistics were provided for various point processes.
8	Assessing spatio-temporal eruption forecasts in a monogenetic volcanic field	Mark S. Bebbington	Volcano eruption	Intensity at a given point is determined from the nearest neighbours with futher discussion on methods for determining their number.
9	Spatio-temporal avalanche forecasting with Support Vector Machines	A. Pozdnoukhov, G. Matasci, M. Kanevski, R. S. Purves	Avalanche release point	Avalanche forecast is made based on binary classication results from support vector machines.
10	Point process modeling of wildfire hazard in Los Angeles County, California	Haiyong Xu, Frederic Paik Schoenberg	Wildfire	Spatio-temporal point process models with their intensity functions consisting of varying sets of parameters.
11	A hierarchical Bayesian spatio-temporal model for extreme precipitation events	Souparno Ghosha, Bani K. Mallick	Extreme weather event	A non-homogeneous Poisson process with the intensity function having

No.	Title	Authors	Natural hazard	Mathematical model (extended description with how attributes are incorporated into the model)
				parameters of a generalised extreme value distribution.
12	Probabilistic estimation of long-term volcanic hazard with assimilation of geophysics and tectonic data	Olivier Jaquet, Christian Lantuéjoul, J. Goto	Volcano eruption	Cox processes are used as models for further implementation in a forecasting framework.
13	A spatio-temporal Poisson hurdle point process to model wildfires	Laura Serra, Marc Saez, Pablo Juan, Diego Varga, Jorge Mateu	Wildfire	Poisson hurdle process consists of two stages and specified in a way to gather together the two processes theoretically involved in the presence of wildfires: the occurrence of being a big wildfire and the frequency of big wildfires per spatial unit.
14	A spatio-temporal modelling framework for assessing the fluctuations of avalanche occurrence resulting from climate change: application to 60 years of data in the northern French Alps	Nicolas Eckert, E. Parent, R. Kies, H. Baya	Avalanche release point	A non-homogeneous Poisson process with the intensity function consisting of terms for expected number of events and a relative risk term.
15	Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm	Peter J. Diggle, Paula Moraga, Barry Rowlingson, Benjamin M. Taylor	-	For real-world data analyses were provided for various point processes.
16	Detecting and modeling multi-scale space-time structures: the case of wildfire occurrences	Edith Gabriel, Thomas Opitz, Florent Bonneu	Wildfire	A log-Gaussian Cox process with its log-intensity function estimated from a linear predictor consisting of terms for inhibition between events, covariates and their parameters and interaction between events.
17	Analysis of tornado reports through replicated spatiotemporal point patterns	Jonatan A. Gonzalez, Ute Hahn, Jorge Mateu	Tornado	A non-homogeneous Poisson point process with its intensity function estimated from an edge-corrected nonseparable estimator of Gaussian kernels.
18	Earthquake Modelling at the Country Level Using Aggregated Spatio-Temporal Point Processes	M.N.M. van Lieshout, A. Stein	Earthquake	A marked spatio-temporal clustered point process with incorporation of parents and offspring generation.
19	Second-order analysis of inhomogeneous spatio-temporal point process data	Edith Gabriel, Peter J. Diggle	-	For real-world data analyses were provided for various point processes.
20	Spatio-temporal modelling of extreme storms	Theodoros Economou, David B. Stephenson, Christopher A. T. Ferro	Hurricane	A non-homogeneous Poisson process models the high intensity points events along cyclone tracks.

Appendix H. Spatial extent and propagation of the lightning strike occurrences for both storm-systems

This appendix consists of additional angles of the three-dimensional view for the lightning strikes assigned to Storms 1A, 1B and 1C from the first case study (**Figure 3.7**) in **Figure H1** and for the strikes assigned to Storms 2A, 2B and 2C from the second case study (**Figure 3.10**) in **Figure H2**. The case studies are presented in **Section 3.3.2**.

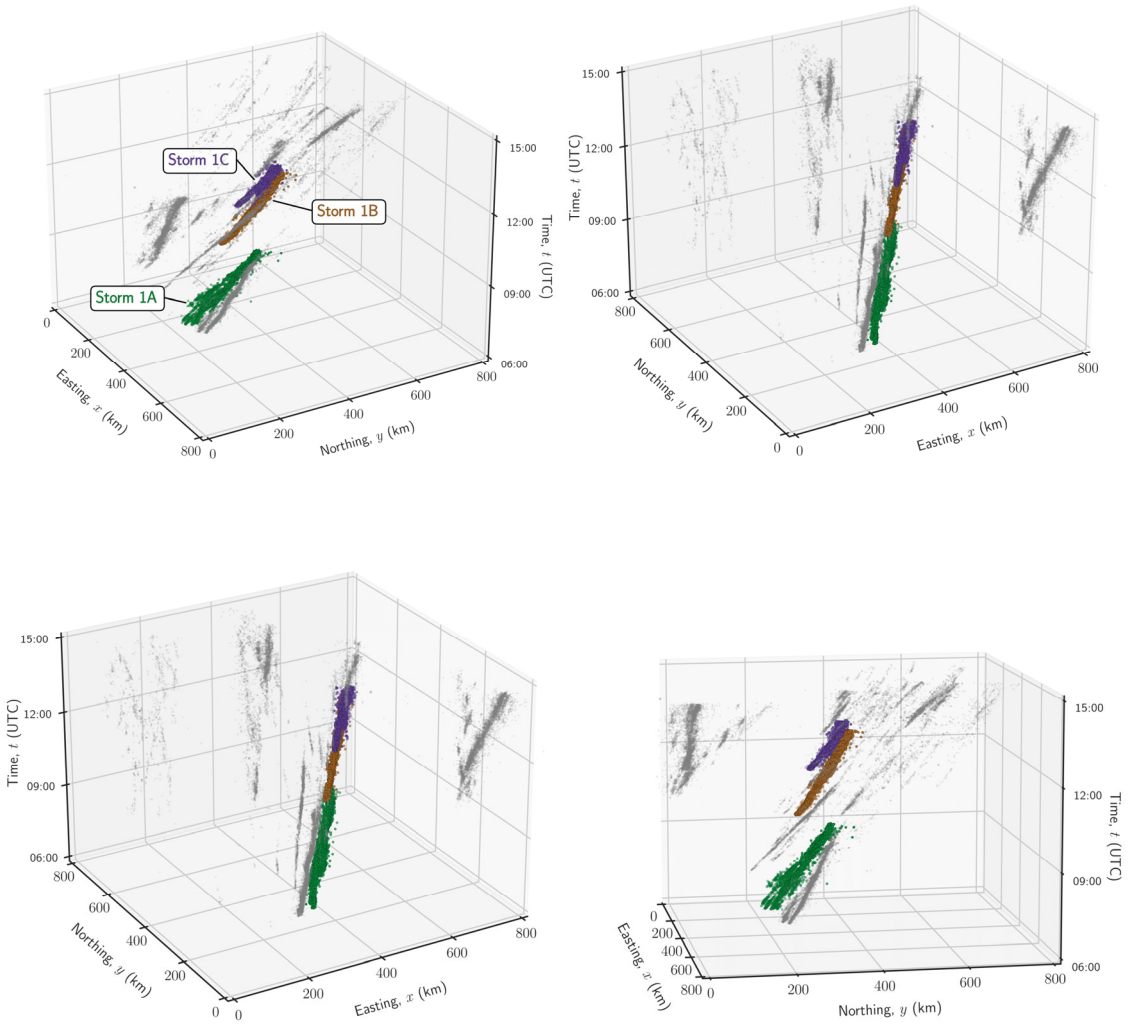

Figure H1. Additional angles of the 3D spatial extent and propagation of the lightning strike occurrences plots in **Figure 3.7**. Lightning strikes are projected onto the OSGB 1936 / British National Grid northing/easting coordinates, assigned to the identified thunderstorms corresponding to the relevant synoptical analysis.

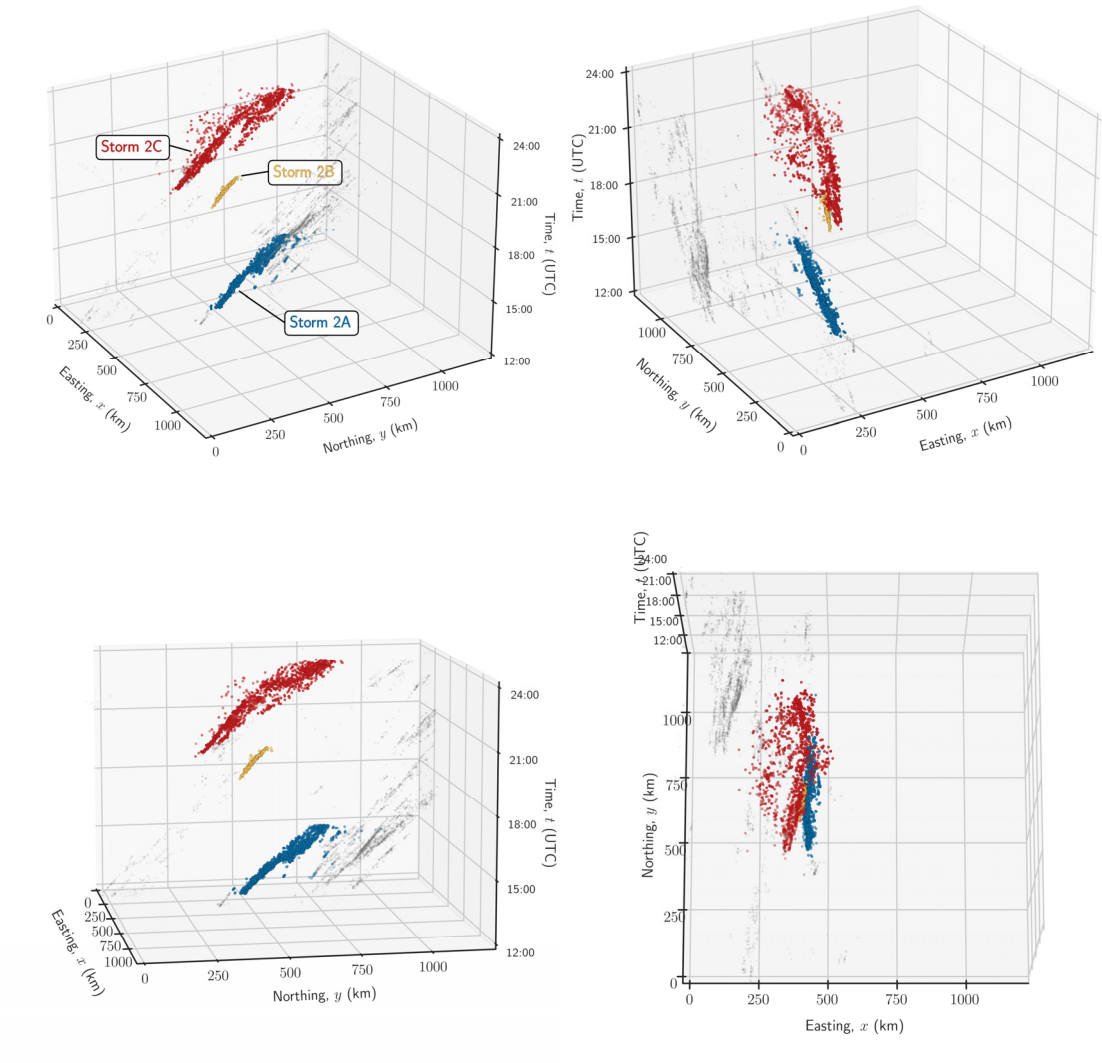

Figure H2. Additional angles of the 3D spatial extent and propagation of the lightning strike occurrences plots in **Figure 3.10**. Lightning strikes are projected onto the OSGB 1936 / British National Grid northing/easting coordinates, assigned to the identified thunderstorms corresponding to the relevant synoptical analysis.

Appendix I. 2D spatial count of lightning strikes in easting and northing

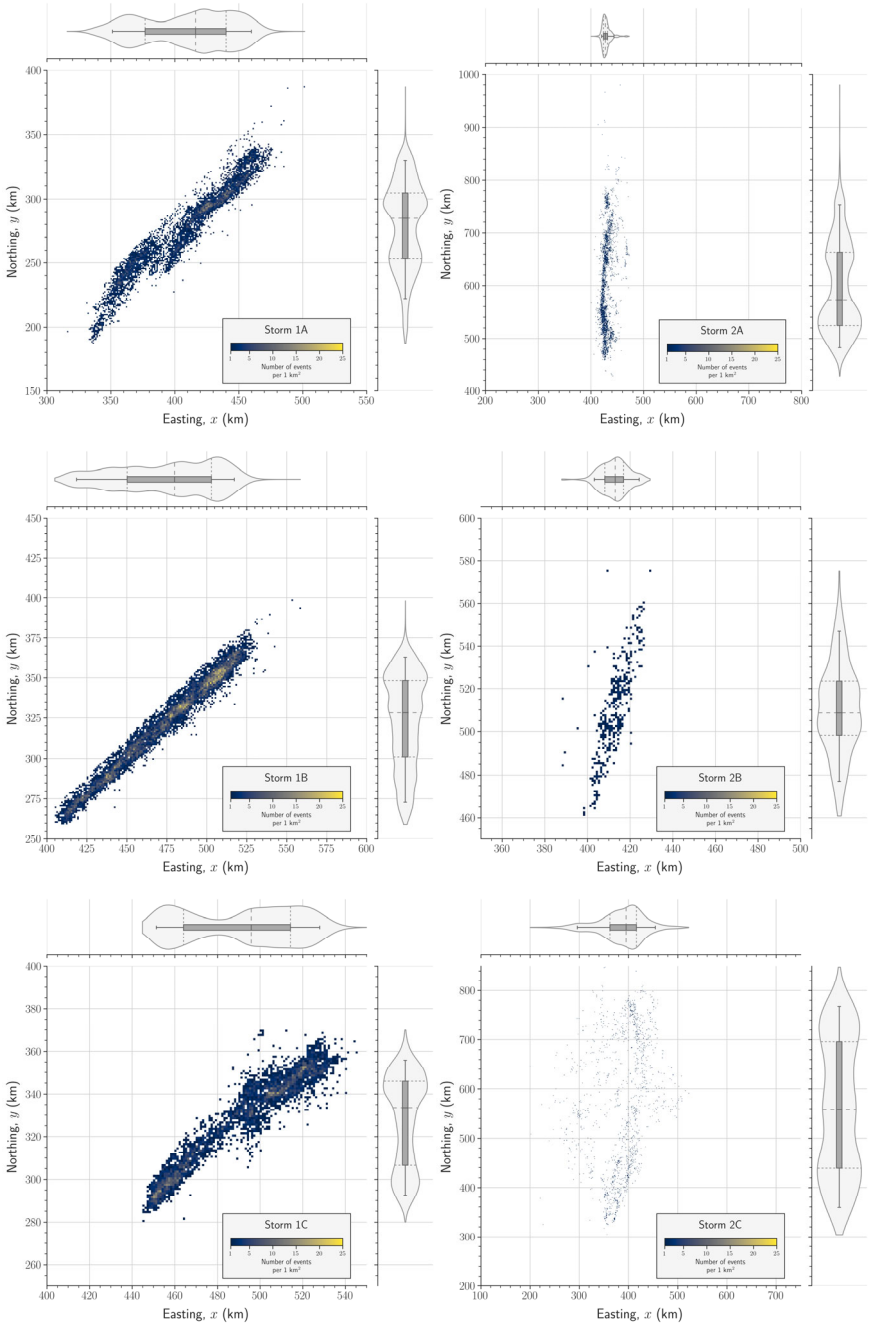

Figure I1. 2D spatial count of lightning strikes per 1 km^2 of the OSGB 1936 / British National Grid northing/easting coordinates for thunderstorms 1A, 1B, 1C, 2A, 2B and 2C as per the legend in each subplot. Only bins with more than 1 lightning strike are coloured. Note that the northing and easting extents are the same, but the axes limits differ. For both the easting and northing values, the marginal distribution is indicated by overlapping boxplots and violinplots. Boxplots indicate the quartiles, have whiskers for the 5th and 95th percentiles, while the violinplots are cut at the last value.

Appendix J. Cutting time series of inter-event times

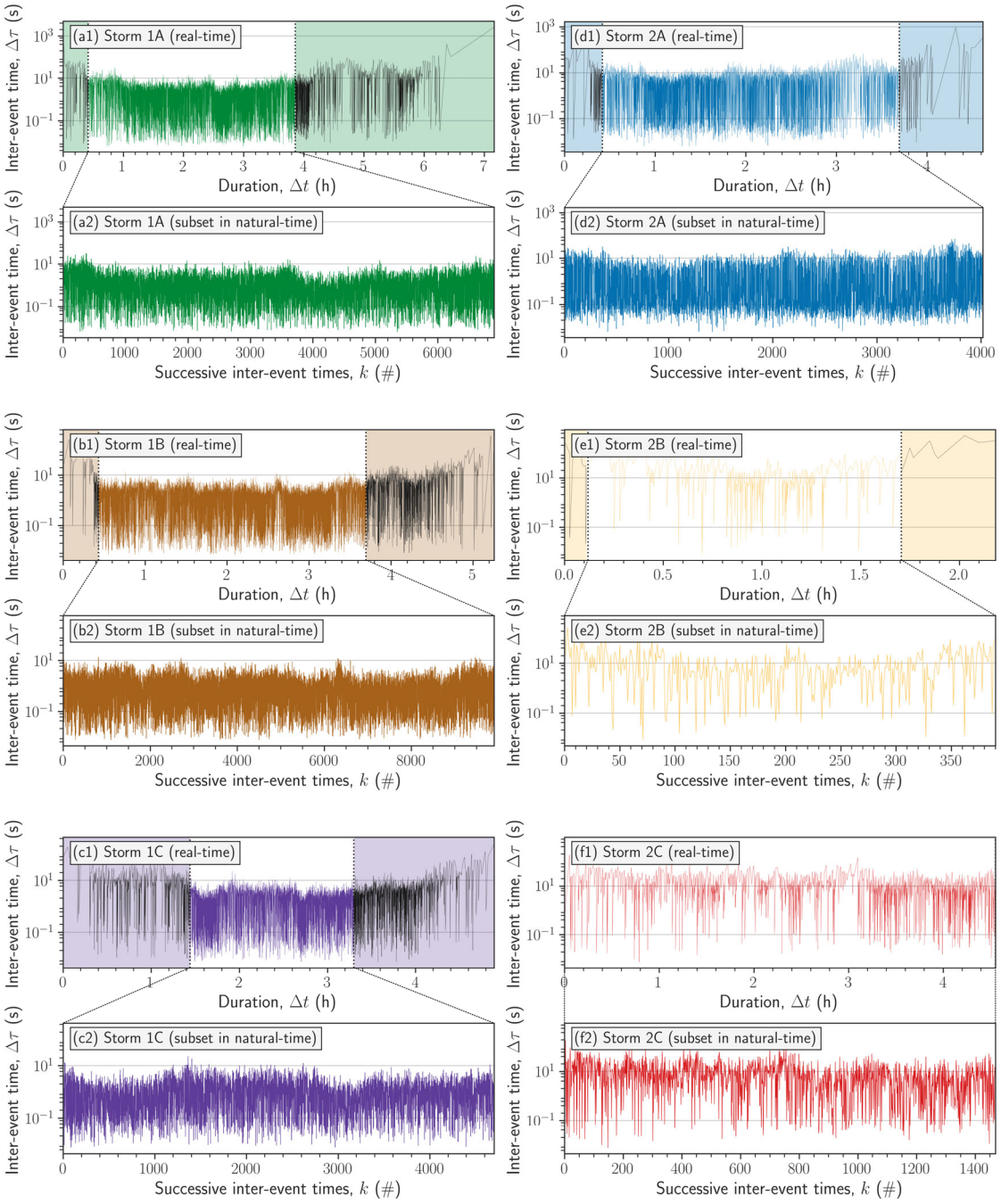

Figure J1. Subplots showing the cutting of inter-event time $\Delta\tau$ time-series in real-time t (removal of black line in upper subfigure) and the subsequent inter-event time time-series in natural time k (lower subfigure) for thunderstorms (a#) 1A, (b#) 1B, (c#) 1C, (d#) 2A, (e#) 2B and (f#) 2C.

Appendix K. Least-squares plane-fit solution to lightning datasets

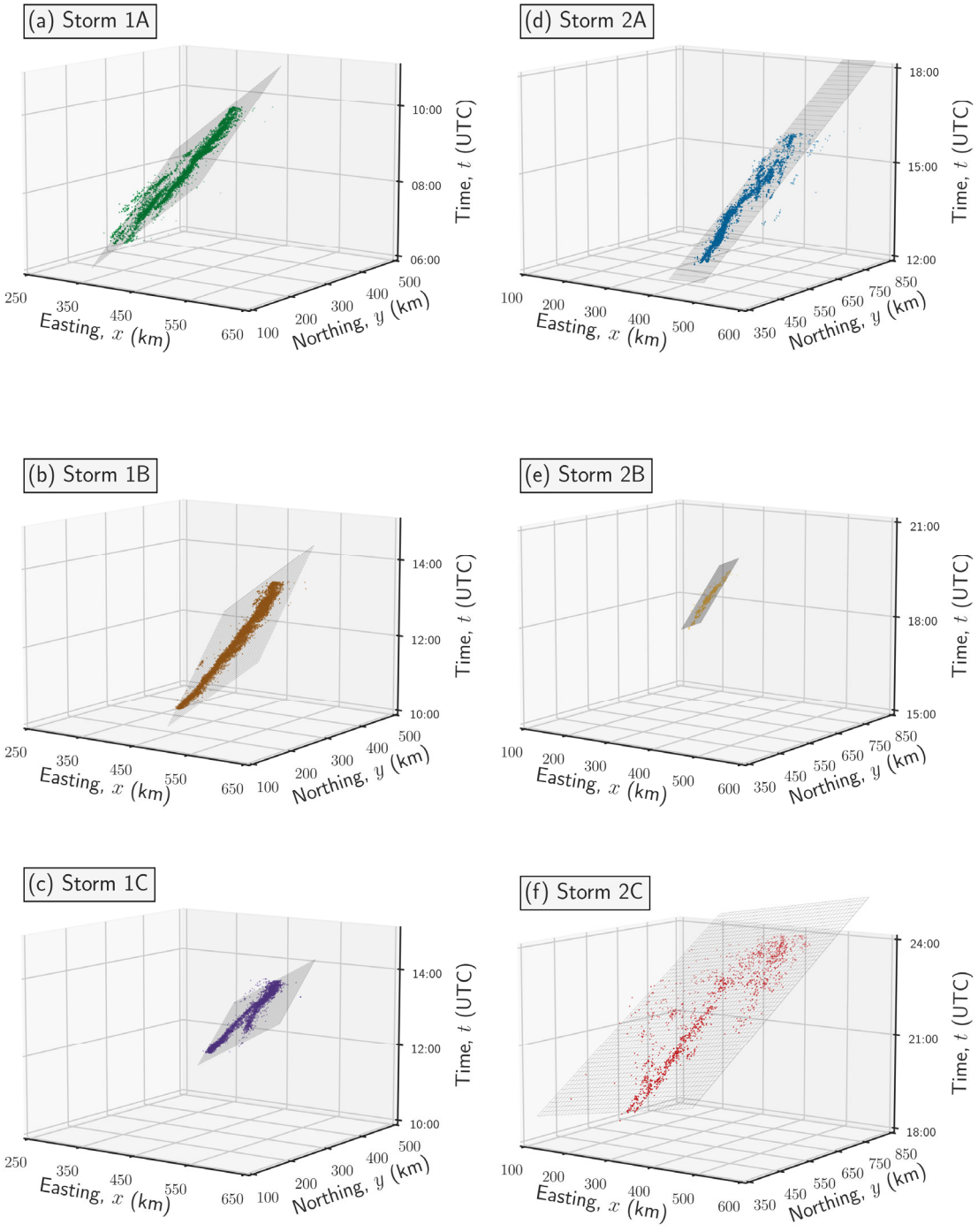

Figure K1. Least-squares plane-fit solution to the lightning strike datasets for thunderstorms **(a)** 1A, **(b)** 1C, **(c)** 2B and **(d)** 2C. The colour scheme for lightning strikes of each storm is consistent with **Figure 3.8-3.11** and the best-fit plane is drawn as grey wireframe. Note that all three axes have different extents for each storm, while the ratio between the axes is constant.

Appendix L. Linear least-squares fit to estimate storm movement speed in the spatio-temporal lag space

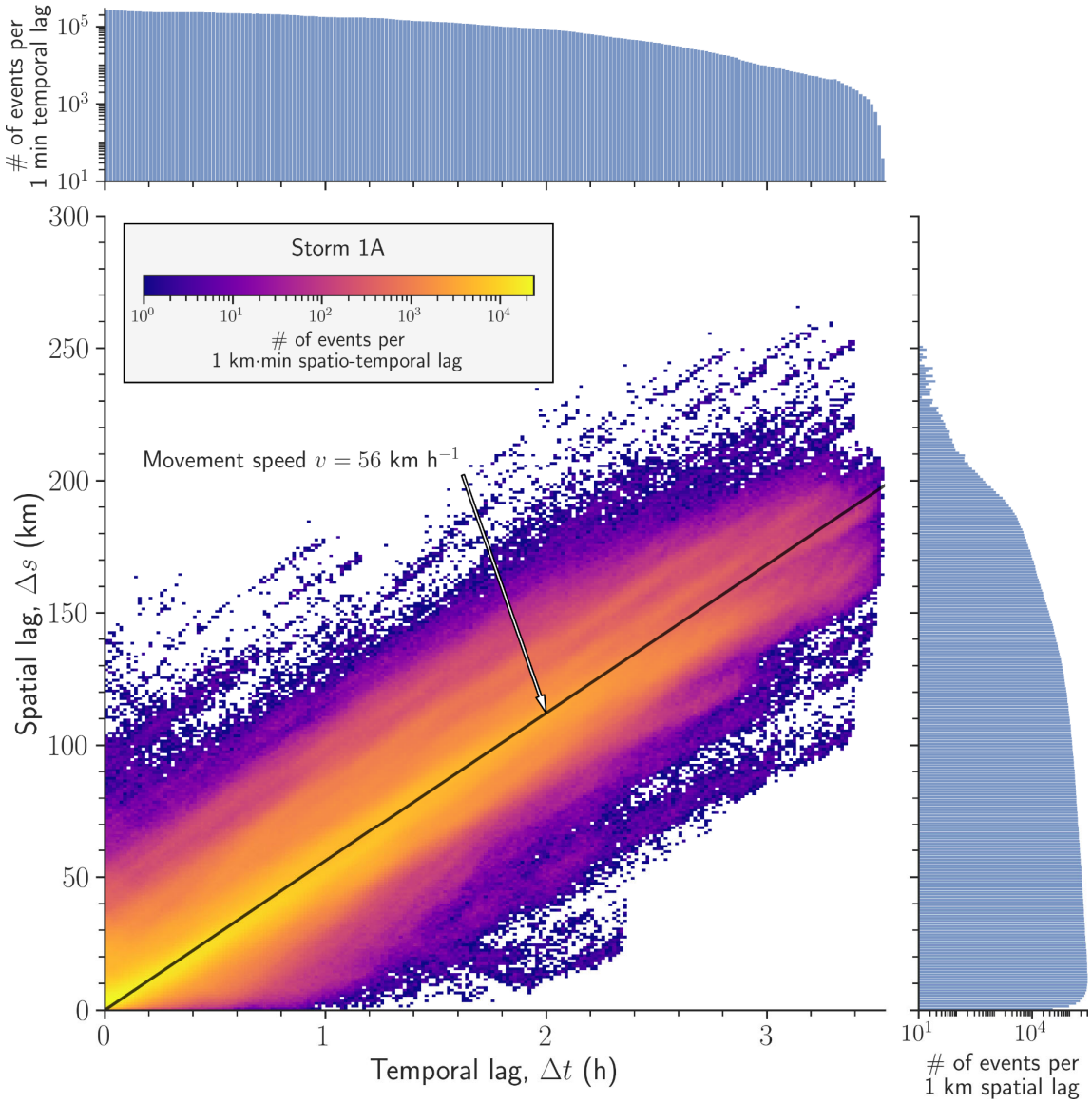

Figure L1. Storm 1A (28 June 2012, n = 7,001 events, v = 56 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The abrupt vertical end of lightning strike data at Δt = 3.45 h occurs due to removal of inter-event time values (see **Figure 3.13**). The solid black line indicates the best-fit linear line passing through the point of origin.

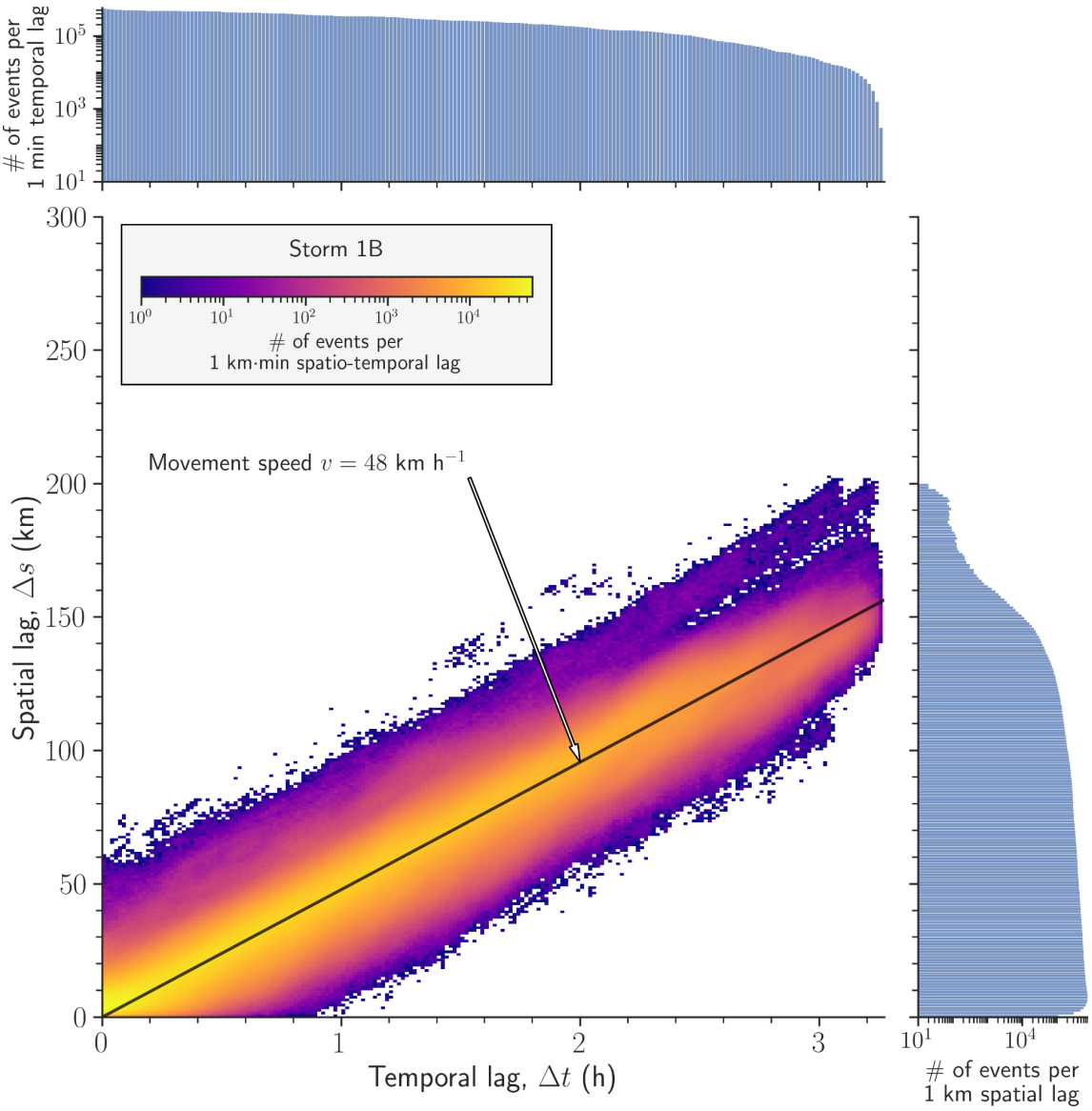

Figure L2. Storm 1B (28 June 2012, n = 9,901 events, v = 48 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The abrupt vertical end of lightning strike data at Δt = 3.25 h occurs due to removal of inter-event time values (see **Figure 3.13**). The solid black line indicates the best-fit linear line passing through the point of origin.

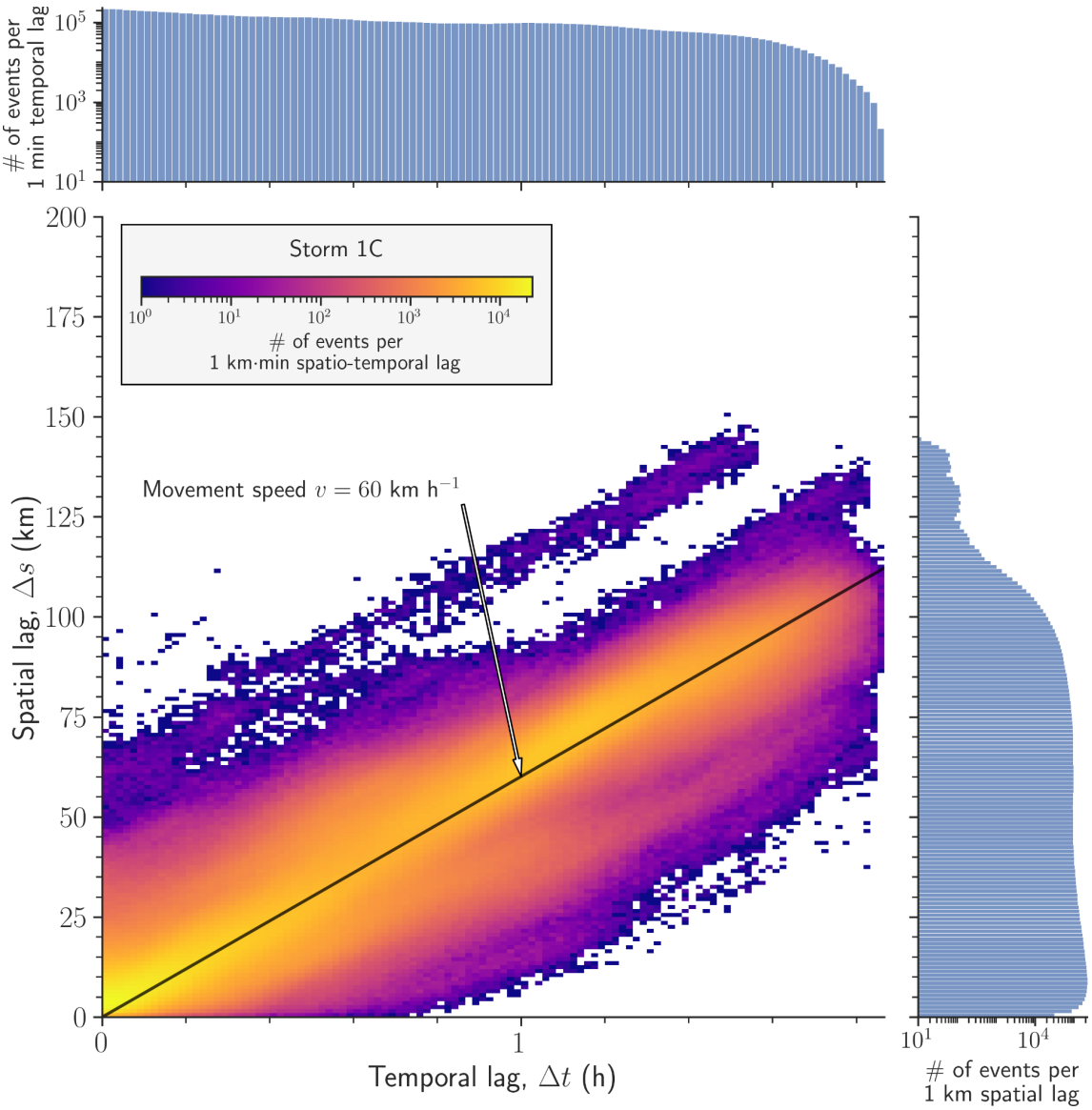

Figure L3. Storm 1C (28 June 2012, n = 4,701 events, v = 60 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The abrupt vertical end of lightning strike data at Δt = 1.9 h occurs due to removal of inter-event time values (see **Figure 3.13**). The solid black line indicates the best-fit linear line passing through the point of origin.

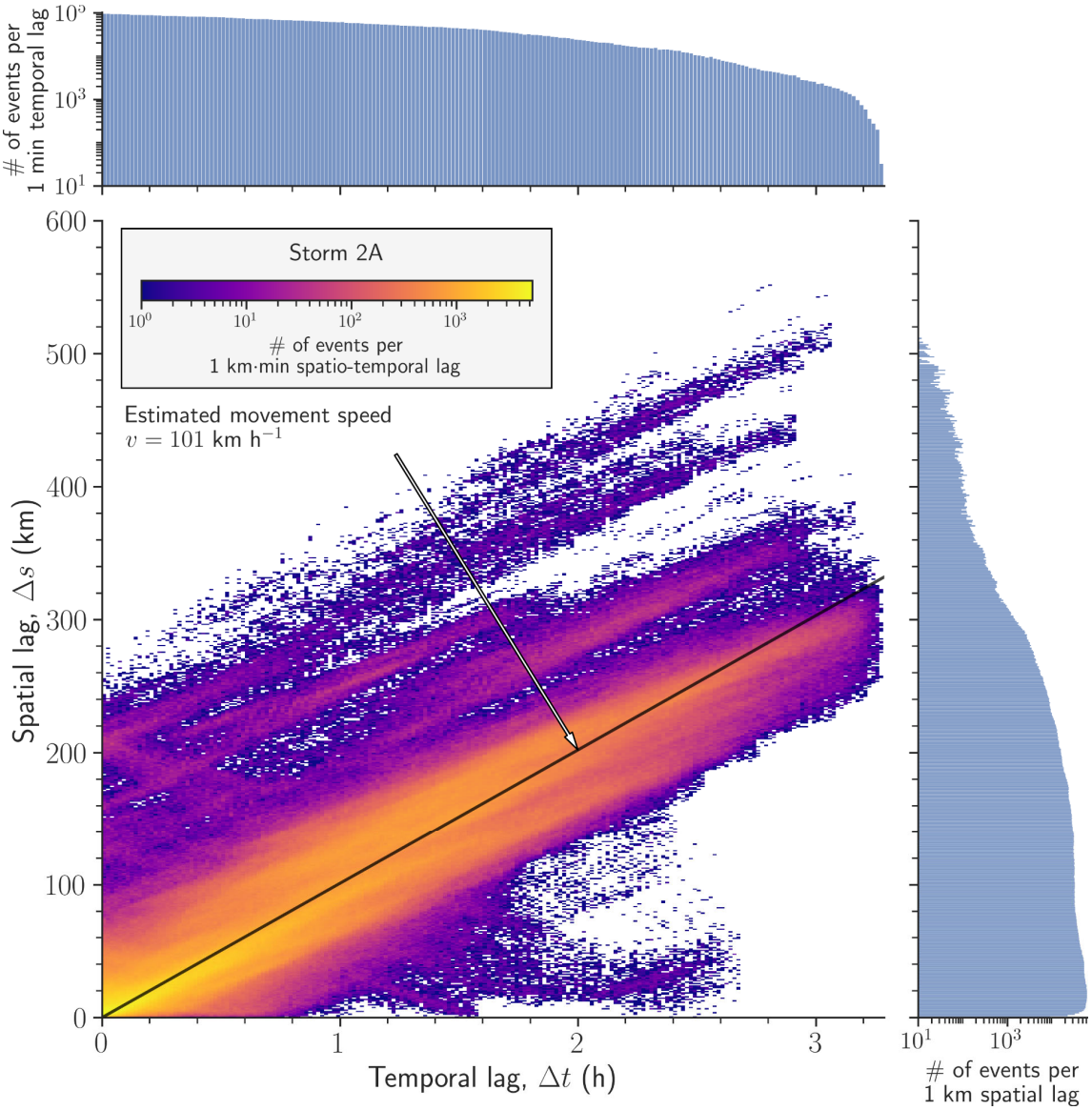

Figure L4. Storm 2A (1 July 2015, n = 4,019 events, v = 101 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The abrupt vertical end of lightning strike data at Δt = 3.25 h occurs due to removal of inter-event time values (see **Figure 3.13**). The solid black line indicates the best-fit linear line passing through the point of origin.

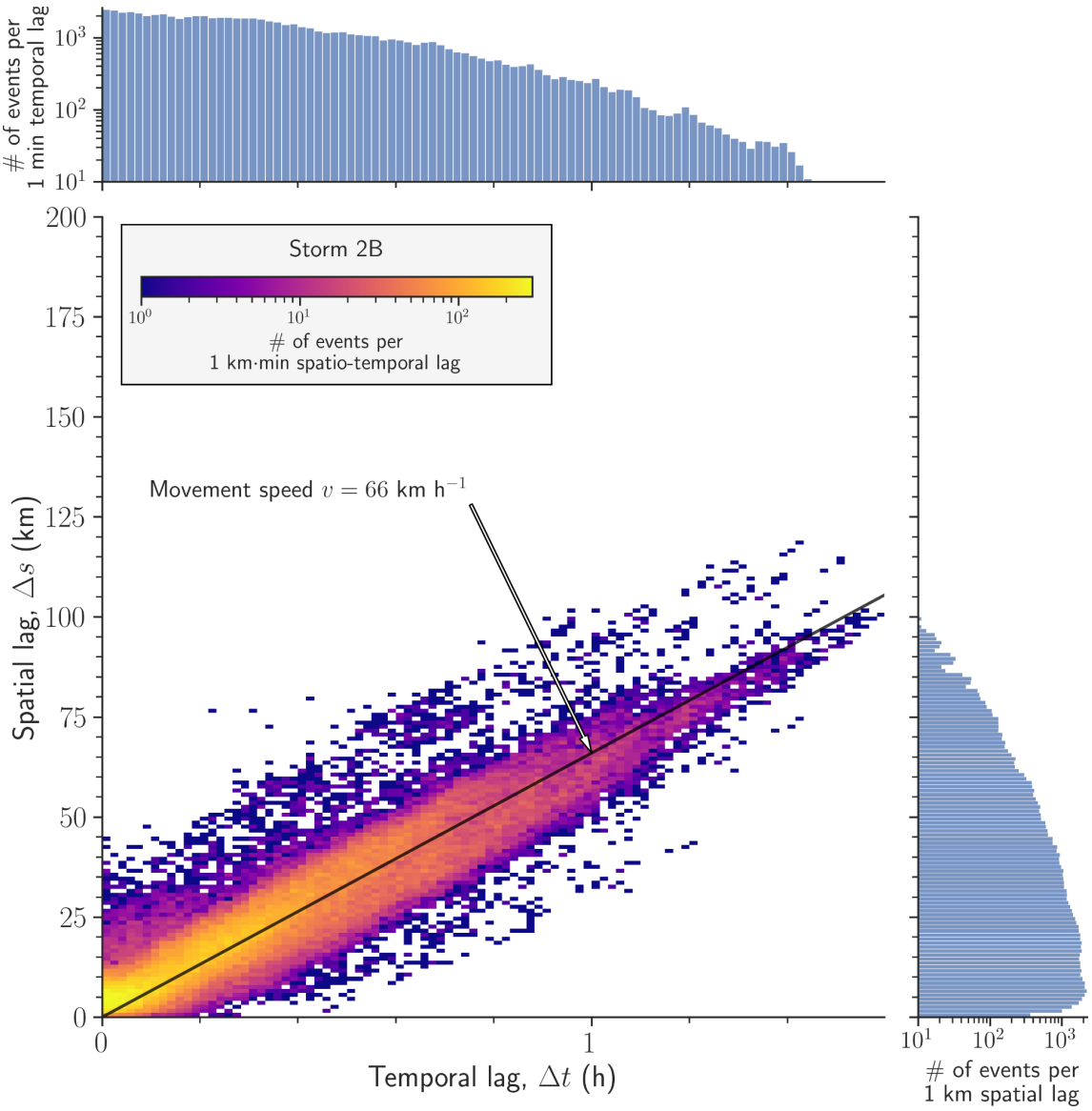

Figure L5. Storm 2B (1 July 2015, $n = 390$ events, $v = 66$ km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The solid black line indicates the best-fit linear line passing through the point of origin.

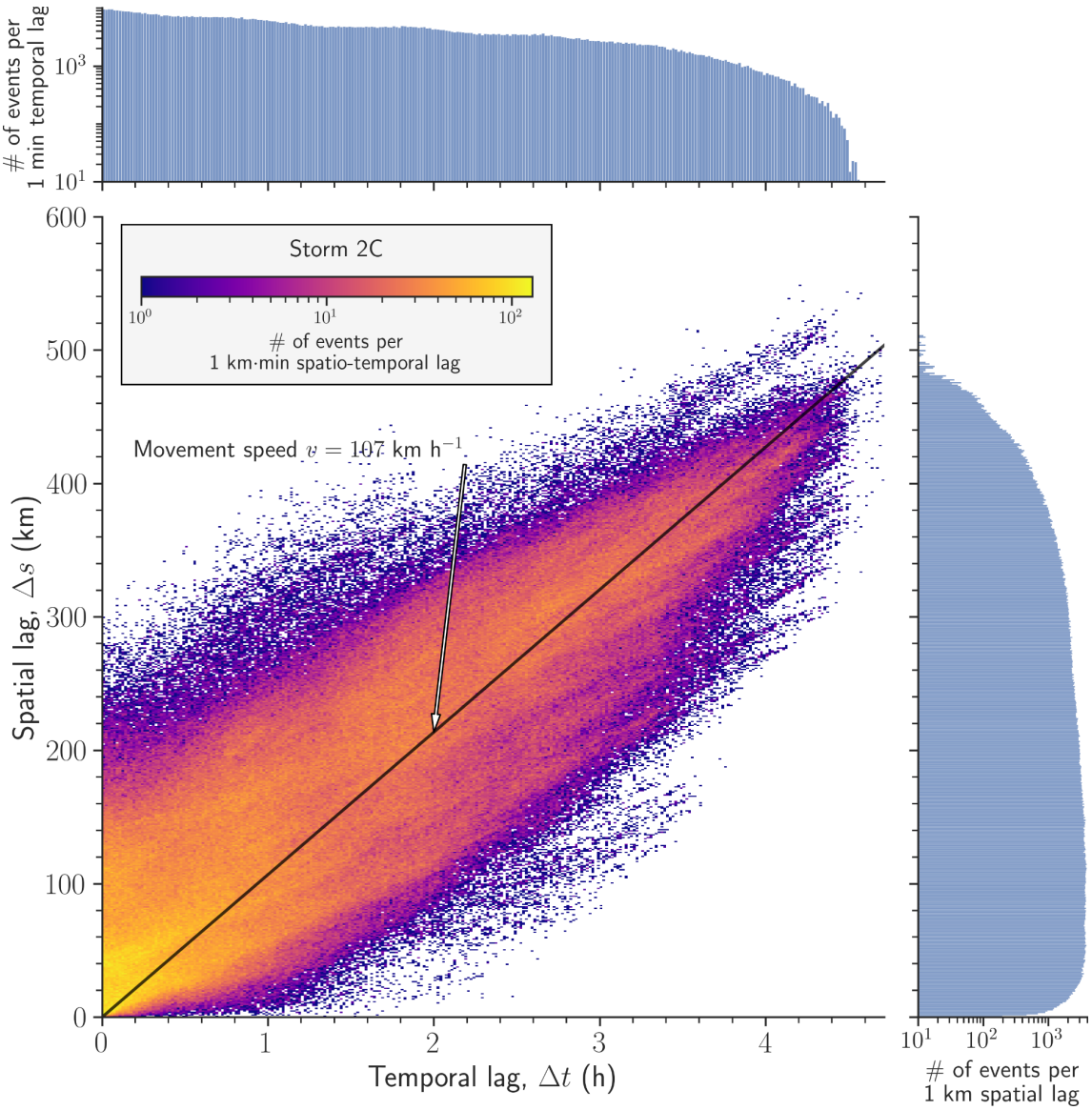

Figure L6. Storm 2C (1 July 2015, n = 1,469 events, v = 107 km h^{-1}) least-squares linear-fit solution and 2D count of lightning strikes per 1 km·min in the spatio-temporal lag space. Bins are coloured according to the given logarithmic colour scale. The marginal bin counts per 1 km spatial lag and 1 min temporal lag are given above and to the right of the main plot. The solid black line indicates the best-fit linear line passing through the point of origin.

Appendix M. Bivariate Gaussian fit to spatial spread variable to sample easting and northing distances

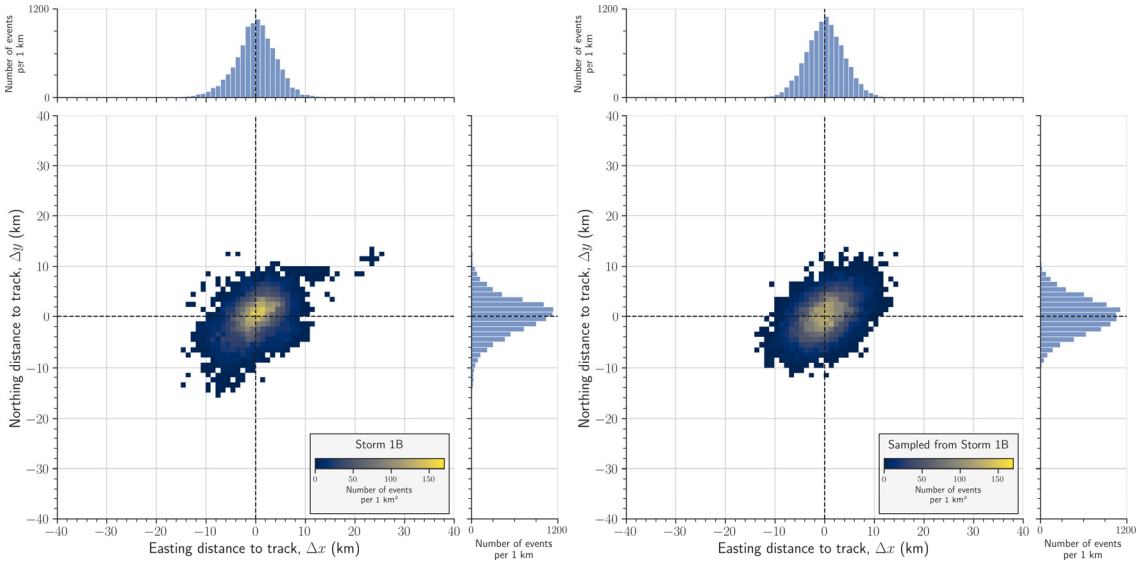

Figure M1. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1B (left subfigure) and for a bivariate Gaussian distribution based on thunderstorm 1B (right subfigure) for window sizes $\Delta w = 500$. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [-40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions.

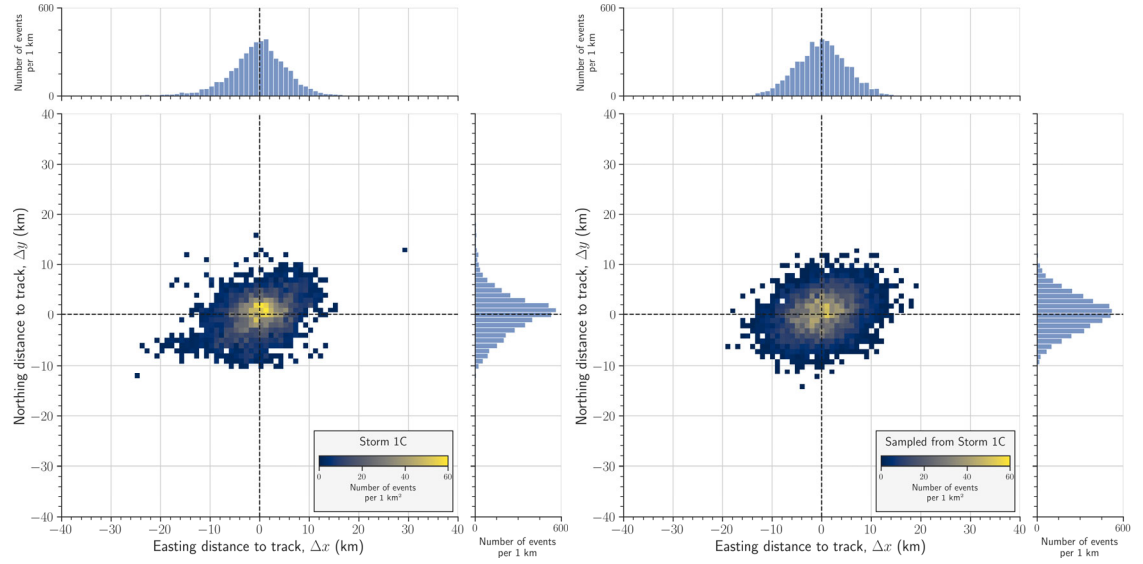

Figure M2. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1C (left subfigure) and for a bivariate Gaussian distribution based on thunderstorm 1C (right subfigure) for window sizes $\Delta w = 500$. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [-40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions.

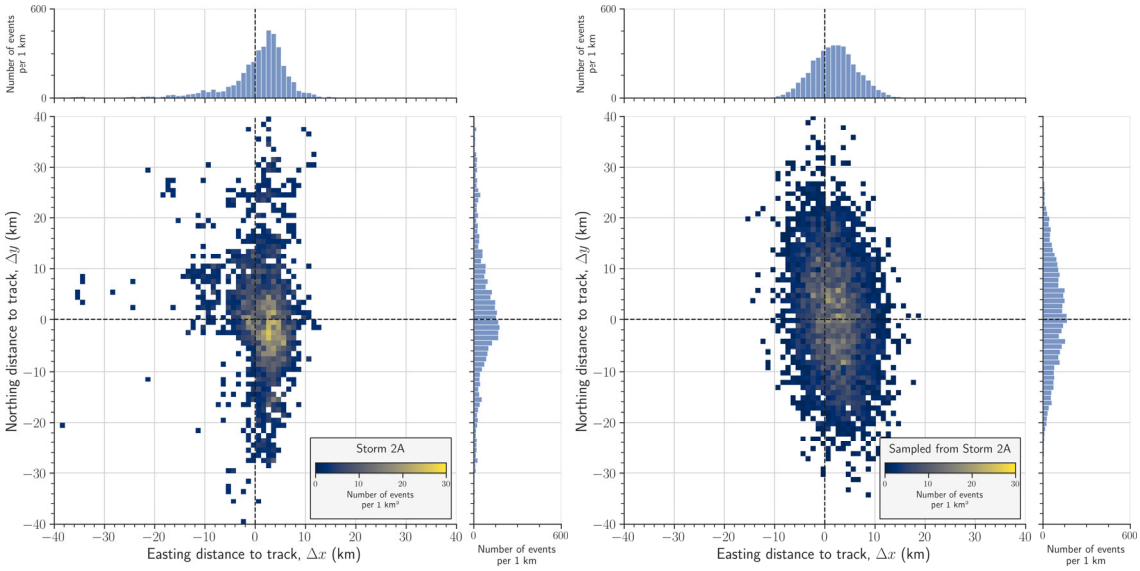

Figure M3. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 2A (left subfigure) and for a bivariate Gaussian distribution based on thunderstorm 2A (right subfigure) for window sizes $\Delta w = 500$. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [-40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions.

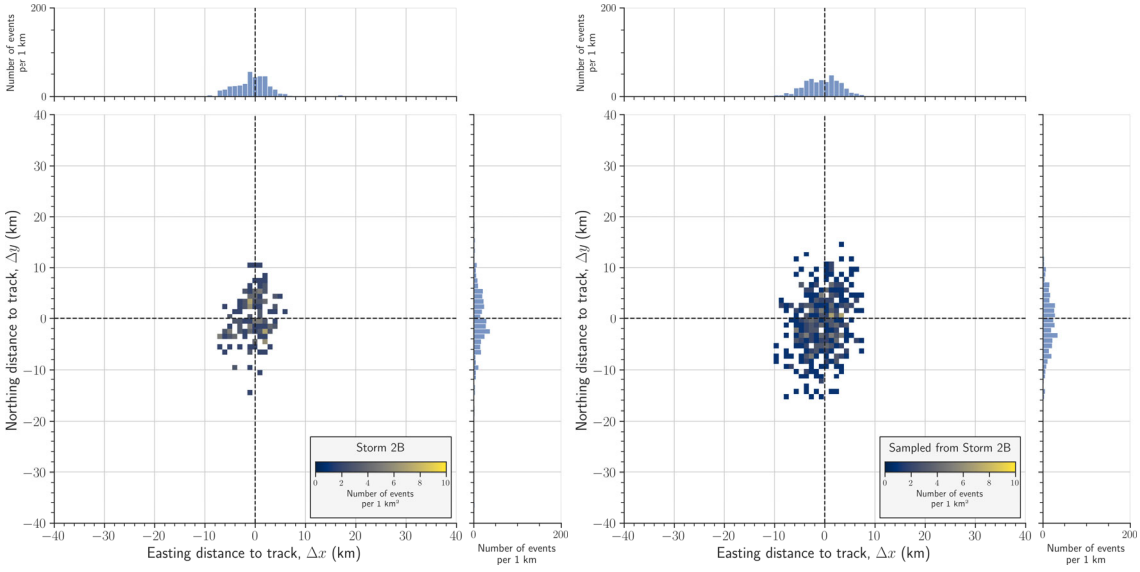

Figure M4. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 2B (left subfigure) and for a bivariate Gaussian distribution based on thunderstorm 2B (right subfigure) for window sizes $\Delta w = 500$. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [-40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions.

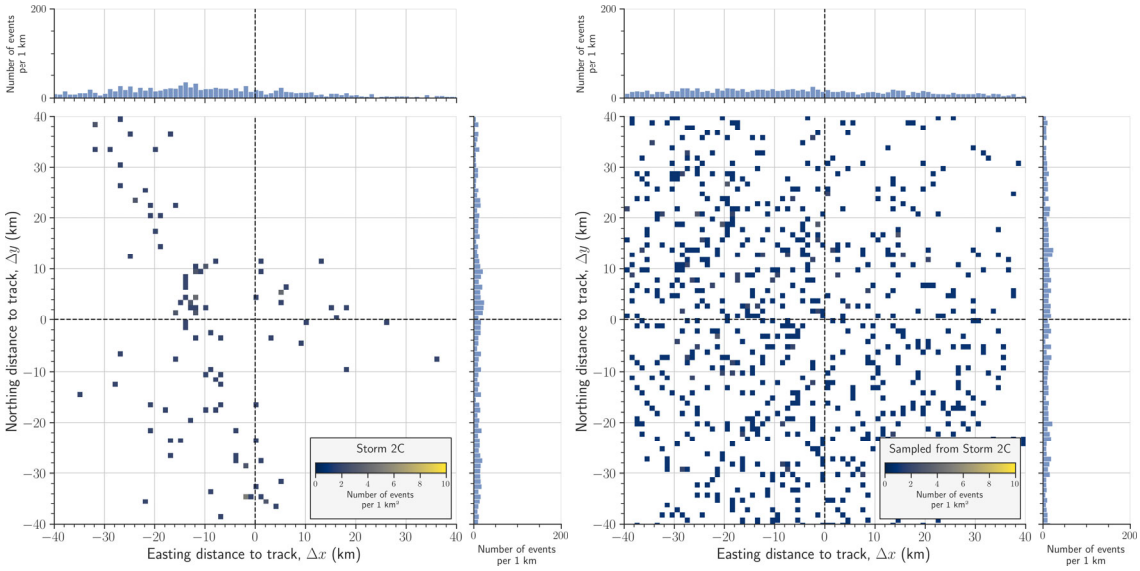

Figure M5. 2D spatial count of lightning strikes per 1 km^2 of easting and northing distances to the movement track in natural time k for thunderstorm 1C (left subfigure) and for a bivariate Gaussian distribution based on thunderstorm 1C (right subfigure) for window sizes $\Delta w = 500$. Only bins with more than 1 lightning strike are coloured and the plot is zoomed onto [-40,40] km intervals for both easting and northing distance. The marginal bin counts per 1 km are given for both dimensions.

Appendix N. Sampling, sorting and re-indexing procedure for inter-event time variable

This appendix shows the sampling, sorting and re-indexing procedure for the inter-event time data for all six thunderstorms from the two case studies in **Section 3.3.2**. It supplements the storms shown in **Figure 3.20**.

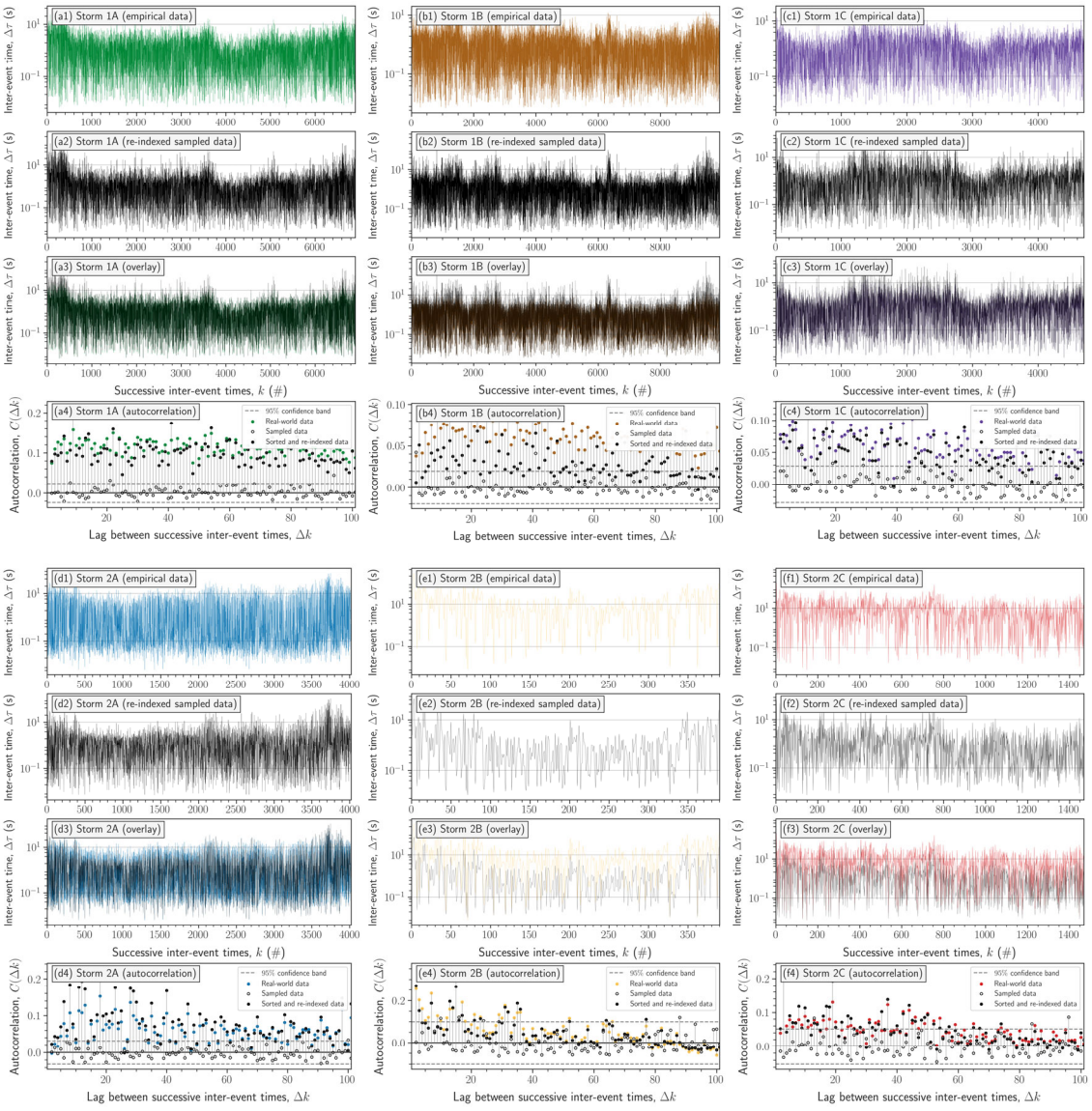

Figure N1. Sampling, sorting and re-indexing procedure based on the combined supercell inter-event time $\Delta\tau$ dataset and thunderstorm (a#) 1B, (b#) 1C, (c#) 2A and (d#) 2B. First subfigure shows the original cut inter-event time time-series in natural time k. Second subfigure shows this sampled time-series being sorted and re-indexed based on the particular thunderstorm, where the indices of sorted time-series coincide between the two time-series. The third subfigure shows the overlay between the original time-series and the sorted and re-indexed one in natural time k. The fourth subfigure shows the overlay between the original time-series and the sorted and re-indexed one in natural time k.

Appendix O. Spatio-temporal clustering results for regular grid data

This appendix consists of the regular grid dataset, its dimensions and bin counts as shown in **Figure O1**. The K-function analysis and the space-time scan statistic framework is applied to this dataset with results displayed in **Figures O2** and **O4** respectively. The initialisation of regular grid data was used to establish the intuition of the methodology performance on a dataset with set geometry. As all clustering methodologies were manually implemented, the results in this appendix aided the understanding of the methodologies and ensuring that the methodologies are performing as expected. It is intended as an additional material to **Chapter 4** as it was not included in the simulation study in **Section 3.4.1**.

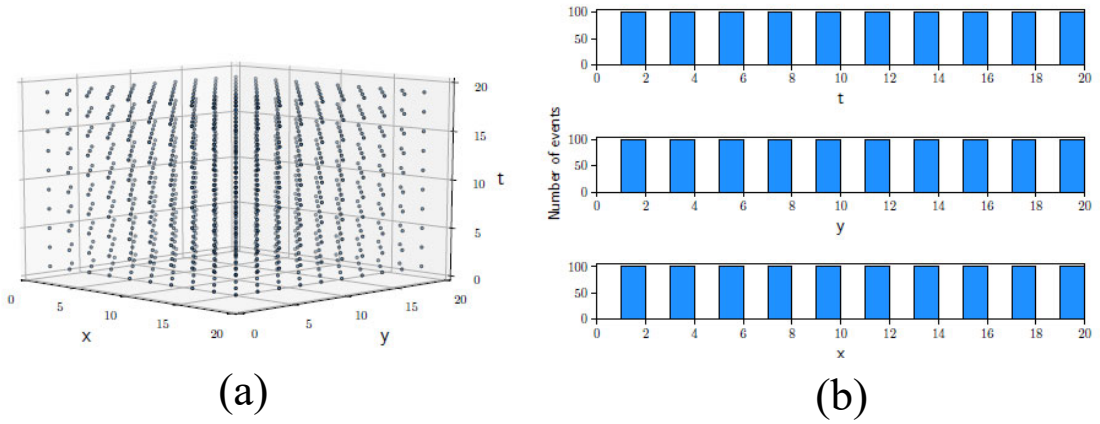

Figure O1. (a) The positions of $N = 1{,}000$ point events as a set of regular grid data in a $\{X = 20, Y = 20, T = 20\}$ space-time volume. **(b)** The unit bin bar charts (with the lower limit of each bin included) of the point event positions when collapsed on each of the three dimensions.

For the K-function analysis, the spatial and temporal ranges are set as $s = \{0.1, 0.2, ..., 5\}$ and $\tau = \{0.1, 0.2, ..., 5\}$. The results contain both the clustering measure D and its components: the spatio-temporal K-function $K(s,\tau)$ and the product of the spatial and temporal K-functions ($K(s)K(\tau)$). For the regular grid point event data, they are given in **Figure O2**.

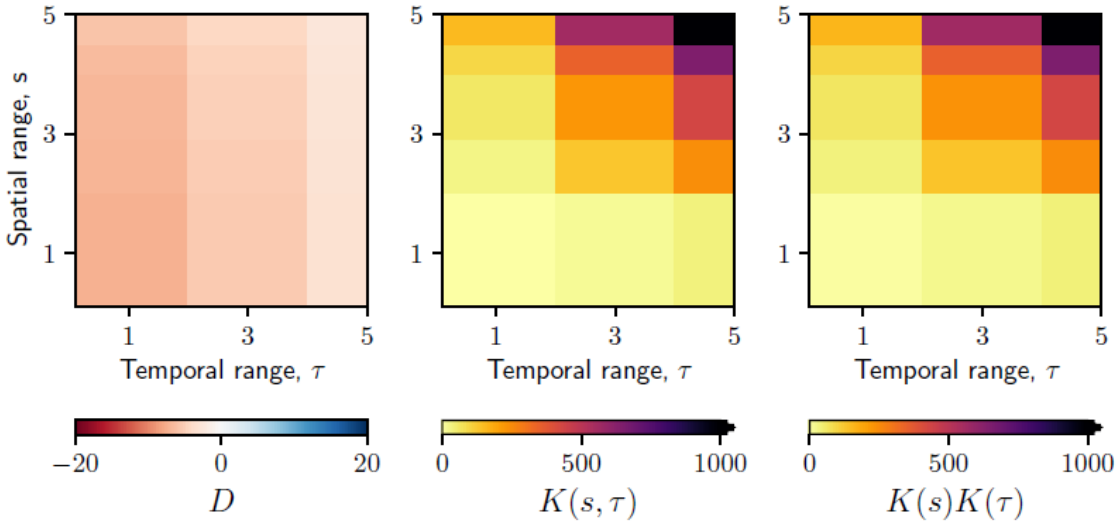

Figure O2. Three subplots of the clustering measure $D = K(s,\tau) - K(s)K(\tau)$, spatio-temporal K-function $K(s,\tau)$ and the product of spatial and temporal K-functions $K(s)K(\tau)$ as applied to regular grid point event data. Their values are calculated for various combinations of spatial range s and temporal range τ and are indicated by their respective colour scales.

As the regular grid dataset was created with equal spacing in all three dimensions, the expected effect of the spatial and temporal ranges on the values of measures D, $K(s,\tau)$ and $K(s)K(\tau)$ can be confirmed. For the spatial range s, changes in all three measures coincide with the chosen spacing parameter of 2 spatial units. An illustrative summary of all the possible spatial distances (within the selected maximum spatial range of $s = 5$) from an arbitrary point event within the regular grid data is given in **Figure O3**.

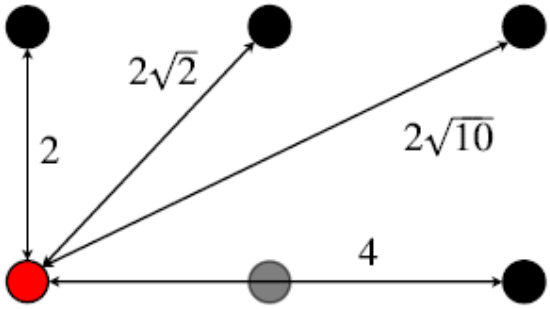

Figure O3. Various spatial distances that can be measured from an arbitrary point event (coloured in red) for the regular grid dataset with the set spacing of 2 spatial units.

Comparing the changes in all three measures along the spatial ranges in **Figure O2** with the spatial distances in **Figure O3** shows perfect agreement. In the temporal domain, the point events are separated by a constant spacing value of 2 temporal units, which is confirmed in the changes of all three measures along the temporal range τ.

Clustering measure D is negative throughout the selected spatial and temporal range values, indicating repulsion between events in space and time [**Diggle *et al.*, 1995; Diggle *et al.*, 1995**; **Tonini and Cama, 2019; Tonini and Cama, 2019**]. If events are close in space, they will tend to be further in time. This is an expected result for the regular grid data due to collapsing of the points events on the spatial or temporal domain. When measuring temporal separation, with the set spacing of 2 temporal units, there are 10 unique time coordinates $t = \{1, 3, ..., 19\}$ in the generated space time cube $\{X = 20, Y = 20, T = 20\}$. As there are $N = 1{,}000$ point events in the dataset, 100 point events will have the same unique time coordinate. For spatial separation, there are 100 unique spatial coordinate pairs $(x, y) = \{(1, 1), (1, 3), ..., (19, 17), (19, 19)\}$, each shared between 10 point events. Therefore, even for spatial and temporal range values below the set spacing of 2, measure $K(s)K(\tau)$ will have a non-zero value, while the point events collapsed onto the spatial domain will not share the same unique time coordinate and $K(s, \tau) = 0$. As both range values increase and cross the set spacing value, the effect of collapsing regular grid data decreases and the clustering measure D approaches zero.

The results of the space-time scan statistic for the regular grid data are shown in **Figure O4** and the input parameters and their values that define the elliptic cylinder encompassing the identified cluster of point events are listed in **Table O1**.

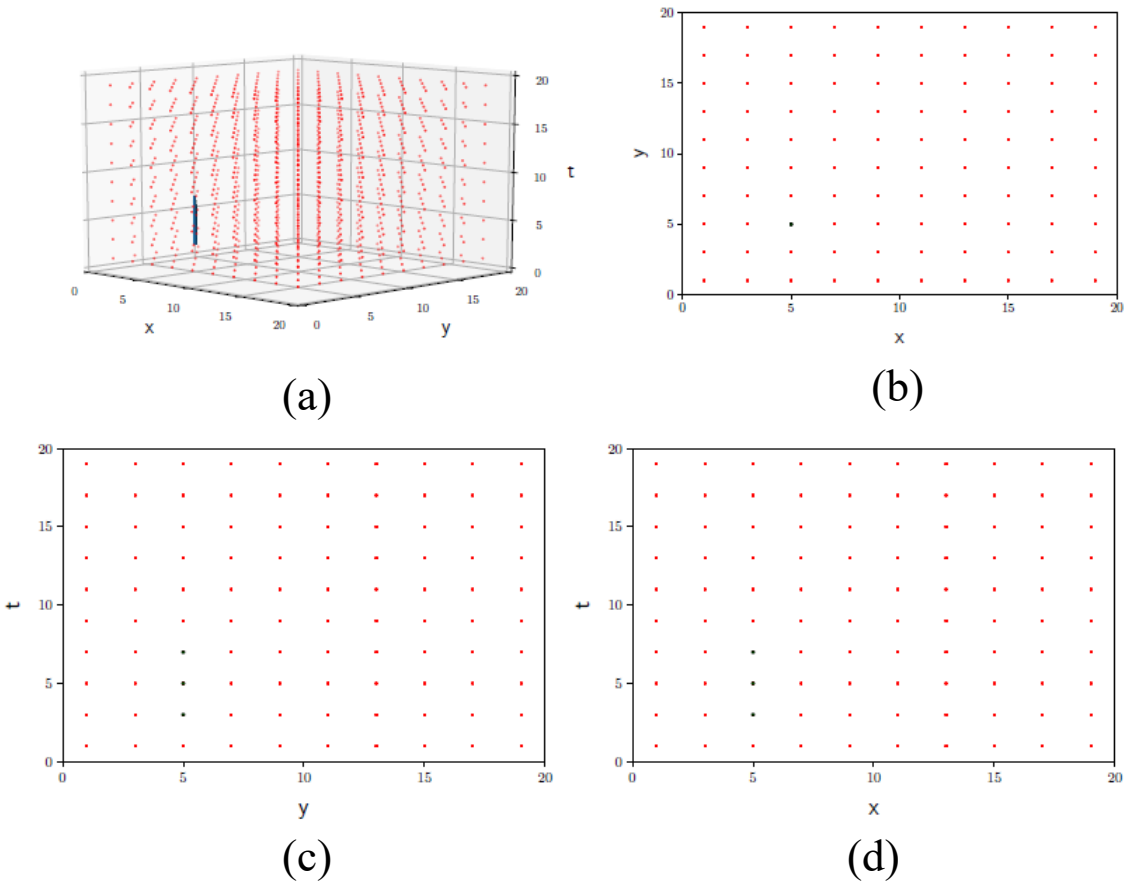

Figure O4. Various angles [**(a)** 3D view of the spatio-temporal domain, **(b)** 2D spatial view, **(c)** 2D *t*-*y* view, **(d)** 2D *t*-*x* view] of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to regular grid data. In all subfigures, the events contained within the scan statistic are coloured green, while those outside in red. In **(a)**, the elliptic scan statistic is drawn in transparent light blue.

Table O1. Input parameter values for space-time scan statistic on regular grid data.

Input parameter	Values
Centroid location	{5, 5, 5}
Length of the semi-major axis	0.1
Shape of elliptic base	3
Angle	0
Height	4

For the regular grid data, the scan statistic results show the expected result based on the geometry of the dataset and the iterative procedure for selecting the parameter combinations. For regular grid data, the location of the elliptic cylinder scanning volume within the inner $\{X = [5, 15], Y = [5, 15], Z = [5, 15]\}$ space-time cube would return identical likelihood ratio values with the selected input parameters. As iteration through the permutations starts at $\{X = 5, Y = 5, T = 5\}$, the centroid location is found there. With the length of the semi-major axis as 0.1 and shape parameter as 3, the base area of the elliptic cylinder is minimised. The height parameter is increased until the scanning volume encompasses two additional point events.

Appendix P. Spatial and temporal views of the three storm reference Model 4 and its variations

This appendix consists of the two-dimensional spatial and one-dimensional temporal views of the three storm reference Model 4 (**Table 3.7**) and its variations. It is intended as additional material for **Figure 3.25** and the results of the K-function analysis framework as applied to this model and its variations (**Section 4.2.1**).

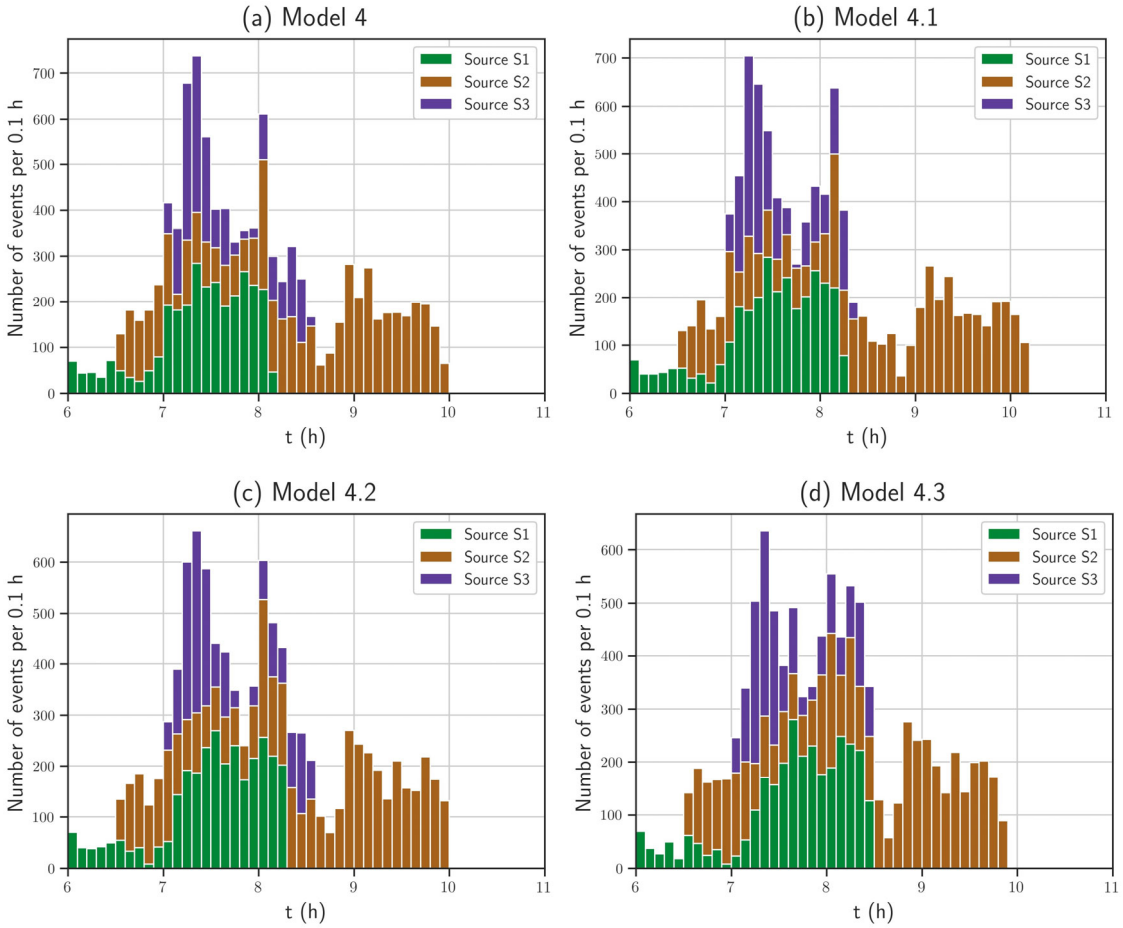

Figure P1. The point events counts per 0.1 h of temporal separation for the **(a)** reference three thunderstorm model (**Model 4**) and its variations – **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. The number of events per 0.1 h is stacked from all three sources. The colour scheme for the three storm sources (S1, S2, S3 as per **Table 3.7**) is kept the same as in the three-dimension view of **Model 4** and its variations in **Figure 3.25**.

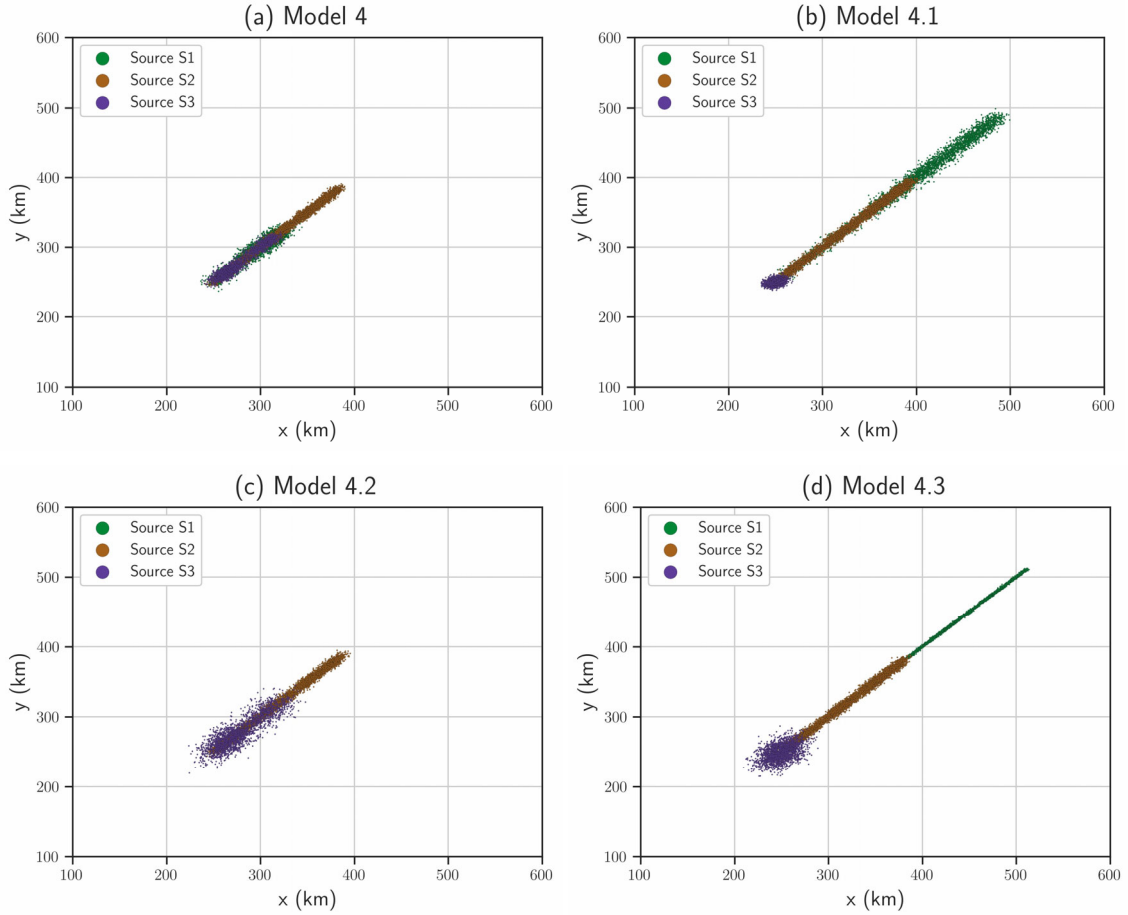

Figure P2. The two-dimension spatial view of the point events as scatter point for the **(a)** three thunderstorm reference model (**Model 4**) and its variations: **(b) Model 4.1**, **(c) Model 4.2** and **(d) Model 4.3**. The colour scheme for the three storm sources (S1, S2, S3 as per **Table 3.7**) is kept the same as in the three-dimension view of **Model 4** and its variations in **Figure 3.25**.

Appendix Q. Inclined elliptic space-time scan statistic spatio-temporal clustering results for Models 4-4.3

This appendix consists of the three-dimensional views of the identified individual cluster of point events as returned by applying the inclined elliptic space-time scan statistic to the three storm reference Model 4 (**Table 3.7**) and its variations. It is intended as additional material for **Figure 5.7**, where the emphasis was put on explaining the methodology in more detail by focusing on one model – Model 4.

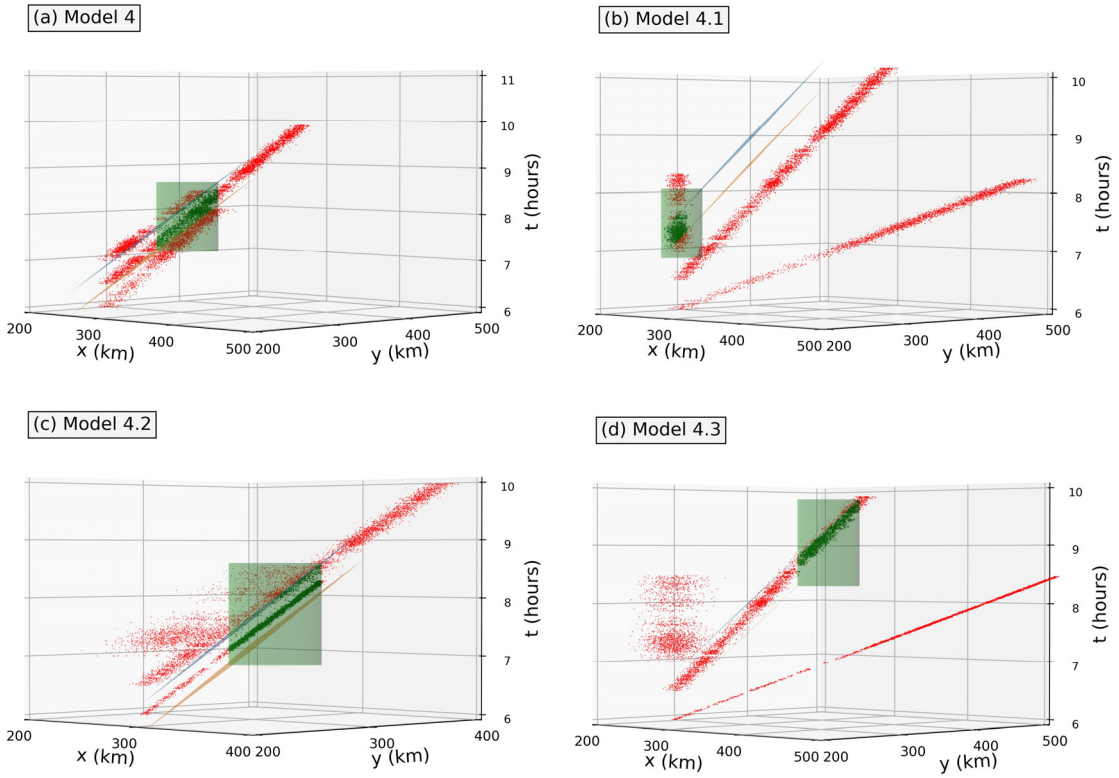

Figure Q1. 3D view of the spatio-temporal domain showing the point events contained within the elliptic space-time scan statistic as applied to **(a) Model 4 (b) Model 4.1**, **(c) Model 4.2**, and **(d) Model 4.2**. In all subfigures, the events contained within the elliptic space-time volume with an inclination height as identified by the scan statistic are coloured green, while those outside in red. The elliptic scanning volume is drawn in transparent green. The upper and lower planes, controlling the inclination height of the space-time scanning volume are coloured in light blue and orange respectively.

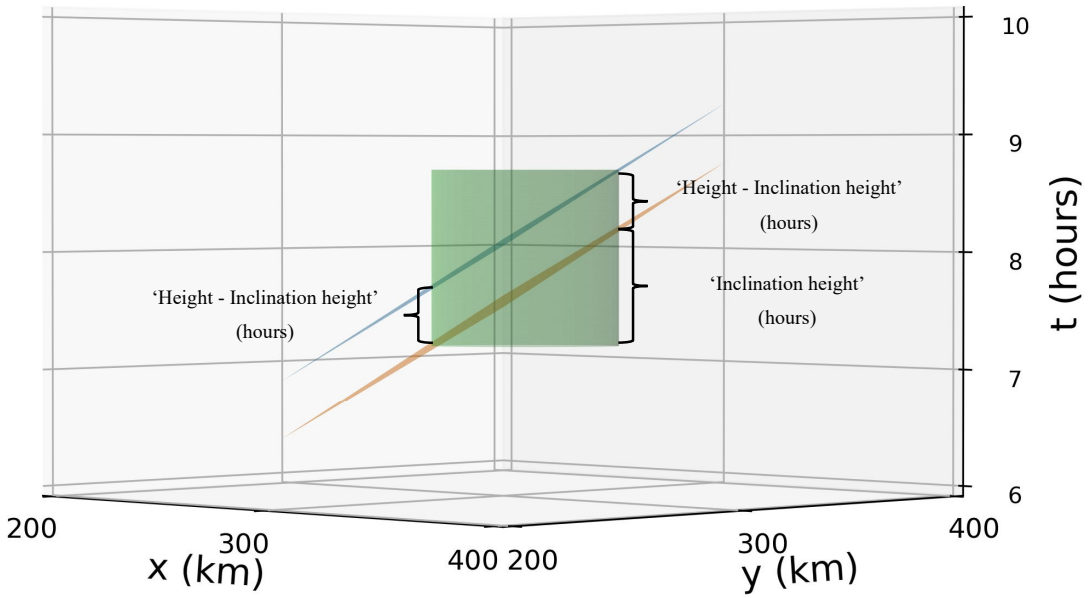

Figure Q2. 3D views looking perpendicular to the movement direction of the spatio-temporal domain showing the elliptic space-time volume with an inclination height coloured in transparent green. The upper and lower planes, controlling the inclination height of the space-time scanning volume are coloured in light blue and orange respectively. The 'Inclination height' and 'Height - Inclination height' input parameters (**Table 5.1**) are marked.