INTERNATIONAL DOCTORAL SCHOOL OF THE USC

Carlos Jorge
da Rocha Balsa

PhD Thesis

Applications of Computational Mathematics to Simulation and Data Analysis: two-dimensional vortex dynamics and meteorological data reconstruction

Santiago de Compostela, 2025

Doctoral Program in Mathematics and Applications

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

APPLICATIONS OF COMPUTATIONAL MATHEMATICS TO SIMULATION AND DATA ANALYSIS: TWO-DIMENSIONAL VORTEX DYNAMICS AND METEOROLOGICAL DATA RECONSTRUCTION

Author

Carlos Jorge da Rocha Balsa

Supervisors

María Victoria Otero Espinar and **Sílvio Marques de Almeida Gama**

Tutor

María Victoria Otero Espinar

PHD PROGRAMME IN MATHEMATICS AND APPLICATIONS

SANTIAGO DE COMPOSTELA

2025

For Adelaide, André, and Aldina.

Acknowledgements

I want to extend my sincere thanks to my supervisors, Professor María Victoria Otero Espinar, from the University of Santiago de Compostela, and Professor Sílvio Marques de Almeida Gama, from the University of Porto. Your invaluable guidance, unwavering encouragement, and the knowledge you so generously shared were fundamental to the successful completion of this thesis.

I am also deeply indebted to the University of Santiago de Compostela and the other institutions in Spain and Portugal that form the University Without Borders (UNISF) project. My special thanks go to the Academic Committee of the Doctorate in Mathematics and Applications for consistently organising seminars that were crucial for the development of this work.

My sincere gratitude also goes to the Polytechnic Institute of Bragança (IPB), where I have had the privilege of teaching Mathematics since 1997. I am particularly thankful to the Research Center for Digitalization and Intelligent Robotics (CeDRI) for their continued support of my research.

I would like to extend my sincere gratitude to two researchers whose valuable insights were instrumental in some of the studies included in this thesis: Professor José Rufino of IPB and Dr. Carlos Veiga Rodrigues, Modelling and Analysis Specialist at Vestas Wind Systems A/S. Their willingness to share their knowledge and their keen interest in my work are deeply appreciated.

I am also deeply grateful to all my master's and undergraduate students who collaborated with me on these studies, namely Raphaelle Monville Letu, Murilo Montanini Breve, Baptiste André, Leonardo Araújo, and Hugo Dupuis. I hope to have successfully passed on my passion for research to them, and I am thankful for the valuable lessons I learned from their contributions.

Finally, I want to express my deepest appreciation to my family. Your dedication and the strength you provided were the indispensable foundation for this work, and it would not have been possible without you.

Carlos Balsa

Contents

List of Figures

List of Tables

Publications Partially Reproduced

- Chapter 4 is based on [16]:

  Balsa, C., Gama, S. (2022). *A control problem with passive particles driven by point vortices on the sphere*. In: Guarda, T., Portela, F. (eds) Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022. Communications in Computer and Information Science, vol. 1675. Springer, Cham. ISBN 978-3-031-20318-3. ISSN 1865-0929.

  DOI: `https://doi.org/10.1007/978-3-031-20319-0_11`

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto.

  Indexing: DBLP, EI Compendex, INSPEC, JST, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, Web of Science (CPCI-S), and zbMATH.

  SCImago Journal Rank (SJR) 2023: 0.203, Q4 for Computer Science and Mathematics.

  CiteScore Rank 2023: 1.1 for Computer Science, and Mathematics.

  Copyrights: The publication *Communications in Computer and Information Science*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6036960257850 is presented in Appendix A.

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design and development of experiments. Manuscript Writing. Funding.

- Chapter 5 is based on [19]:

  Balsa, C., Monville Letu, R., Gama, S. (2023). *Optimization of Vortex Dynamics on a Sphere*. In: Garcia, M.V., Gordón-Gallegos, C. (eds) CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI). CSEI 2022. Lecture Notes in Networks and Systems, vol. 678. Springer, Cham. ISBN 978-3-031-30591-7.

  DOI: `https://doi.org/10.1007/978-3-031-30592-4_15`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Raphaelle Monville-Letu. Université de Toulouse - INP - ENSEEIHT, France.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal.

  Indexing: DBLP, EI Compendex, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, WTI AG, and zbMATH.

  SCImago Journal Rank (SJR) 2023: 0.171, Q4 for Computer Science.

  CiteScore Rank 2023: 0.9 for Computer Science.

  Copyrights: The publication *Lecture Notes in Networks and Systems*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040670626238 is presented in Appendix A.

  Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design of experiments. Drafting of the original and revision of the final manuscript. Funding.

- Chapter 6 is based on [22]:

  Balsa, C., Otero-Espinar, M. V., Gama, S. (2024). *Exploring Controlled Passive Particle Motion Driven by Point Vortices on a Sphere*. Computation, 12(2), 23. ISSN 2079-3197.

  DOI: `https://doi.org/10.3390/computation12020023`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - M. Victoria Otero-Espinar. Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal.

  Indexing: BibCnrs, CAPlus/SciFinder, CNKI, CNPIEC, dblp Computer Science Bibliography, Dimensions, DOAJ, EBSCO, SCOPUS, Gale, Inspec, J-Gate, OpenAIRE, OSTI (U.S. Department of Energy), PATENTSCOPE, ProQuest, and Web of Science (ESCI).

  Journal Impact Factor (JIF) 2024: 2, Q2 for Mathematics.

  SCImago Journal Rank (SJR) 2024: 0.44, Q2 for Applied Mathematics.

  CiteScore Rank 2023: 3.5, Q2 for Applied Mathematics.

  Copyrights: Information about permissions can be found in the journal website: `https://www.mdpi.com/authors/rights`, where the following statement is placed: *For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.*
  The full copyright statements of the journal editor and the first page of the paper are presented in Appendix A.

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design and development of experiments. Drafting of the original and revision of the final manuscript. Funding.

- Chapter 7 is based on [23]:

  Balsa, C., Otero-Espinar, V., Gama, S. (2024). *An Approach to Environmental Cleanup Through Conceptual Autonomous Technology.* In: Gonçalves, J.A.d.C., Lima, J.L.S.d.M., Coelho, J.P., García-Peñalvo, F.J., García-Holgado, A. (eds) Proceedings of TEEM 2023. TEEM 2023. Lecture Notes in Educational Technology. Springer, Singapore. ISBN 978-981-97-1813-9. ISSN 2196-4963.

  DOI: `https://doi.org/10.1007/978-981-97-1814-6_57`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - M. Victoria Otero-Espinar. Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal.

  Indexing: ACM Digital Library, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, and SCOPUS.

  SCImago Journal Rank (SJR) 2020: 0.397, Q2 for Education, and Q3 for Computer Science Applications.

  CiteScore Rank 2023: 1.0, Q4 for Computer Science Applications and Education.

  Copyrights: The publication *Lecture Notes in Educational Technology*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040670335028 is presented in Appendix A.

  Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design and development of experiments. Manuscript Writing. Funding.

- Chapter 8 is based on [21]:

  Balsa, C., Otero-Espinar, M.V., Gama, S. (2024). *A Simple Mathematical Model to Steering Oceanic Debris to a Targeted Region.* In: Guarda, T., Portela, F., Diaz-Nafria, J.M. (eds) Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2023. Communications in Computer and Information Science, vol. 1937. Springer, Cham. ISBN 978-3-031-48929-7. ISSN 1865-0929.

  DOI: `https://doi.org/10.1007/978-3-031-48930-3_5`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - M. Victoria Otero-Espinar. Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal.

  Indexing: DBLP, EI Compendex, INSPEC, JST, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, Web of Science (CPCI-S), and zbMATH.

  SCImago Journal Rank (SJR) 2023: 0.203, Q4 for Computer Science and Mathematics.

  CiteScore Rank 2023: 1.1 for Computer Science, and Mathematics.

  Copyrights: The publication *Communications in Computer and Information Science*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040670924277 is presented in Appendix A.

  Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design and development of experiments. Manuscript Writing. Funding.

- Chapter 9 is based on [20]:

  Balsa, C., Otero-Espinar, M.V., Gama, S. (2025). *Particle Movement on a Rotating Sphere*. In: A. P. Aguiar at al. (Eds.): CONTROLO 2024, Lecture Notes in Electrical Engineering 1325, pp. 1-9, 2025. ISBN 978-3-031-81723-6. ISSN 1876-1100.

  DOI: `https://doi.org/10.1007/978-3-031-81724-3_29`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - M. Victoria Otero-Espinar. Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
  - Sílvio Gama. Centro de Matemática da Universidade do Porto, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal.

  Indexing: DBLP, EI Compendex, INSPEC, JST, SCImago, SCOPUS, WTI AG, and zbMATH.

  SCImago Journal Rank (SJR) 2024: 0.143, Q4 for Industrial and Manufacturing Engineering.

  CiteScore Rank 2023: 0.7 for Industrial and Manufacturing Engineering.

  Copyrights: The publication *Lecture Notes in Electrical Engineering*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040671111635 is presented in Appendix A.

  Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design and development of experiments. Manuscript writing.

- Chapter 10 is based on [12]:

  Balsa, C., Breve, M.M., André, B., Rodrigues, C.V., Rufino, J. (2023). *PCAnEn - Hindcasting with Analogue Ensembles of Principal Components*. In: Garcia, M.V., Gordón-Gallegos, C. (eds) CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI). CSEI 2022. Lecture Notes in Networks and Systems, vol. 678. Springer, Cham. ISBN 978-3-031-30591-7. ISSN 2367-3370.

  DOI: `https://doi.org/10.1007/978-3-031-30592-4_13`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Murilo M. Breve. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Baptiste André. Université de Toulouse - INP - ENSEEIHT, Toulouse, France.
  - Carlos V. Rodrigues. Vestas Wind Systems A/S, Design Centre Porto, Leça do Balio, Portugal.
  - José Rufino. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.

  Indexing: DBLP, EI Compendex, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, WTI AG, and zbMATH.

  SCImago Journal Rank (SJR) 2023: 0.171, Q4 for Computer Science.

  CiteScore Rank 2023: 0.9 for Computer Science.

  Copyrights: The publication *Lecture Notes in Networks and Systems*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040671349758 is presented in Appendix A.

  Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis and design of experiments. Drafting of the original and revision of the final manuscript. Funding.

and also contain an introductory section based on [25]:

Balsa, C., Rodrigues, C.V., Araújo, L., Rufino, J. (2022). *Cluster-Based Analogue Ensembles for Hindcasting with Multistations*. Computation, 10(6): 91.

DOI: `https://doi.org/10.3390/computation10060091`

Authors Affiliation:

- Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
- Carlos V. Rodrigues. Vestas Wind Systems A/S, Design Centre Porto, Leça do Balio, Portugal.
- Leonardo Araújo. Universidade Tecnológica Federal do Paraná, Campus de Ponta Grossa, Ponta Grossa, Brazil.
- José Rufino. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.

Indexing: BibCnrs, CAPlus/SciFinder, CNKI, CNPIEC, dblp Computer Science Bibliography, Dimensions, DOAJ, EBSCO, SCOPUS, Gale, Inspec, J-Gate, OpenAIRE, OSTI (U.S. Department of Energy), PATENTSCOPE, ProQuest, and Web of Science (ESCI).

Journal Impact Factor (JIF) 2024: 2, Q2 for Mathematics.

SCImago Journal Rank (SJR) 2024: 0.44, Q2 for Applied Mathematics.

CiteScore Rank 2023: 3.5, Q2 for Applied Mathematics.

The full copyright statements of the journal editor and the first page of the paper are presented in Appendix A.

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design of experiments. Drafting of the original and revision of the final manuscript. Funding.

- Chapter 11 is based on [35]:

  Breve, M., Balsa, C., Rufino, J. (2024). *Reconstruction of meteorological records with PCA-based Analog Ensembles methods*. In: Rocha, A., Adeli, H., Dzemyda, G., Moreira, F., Colla, V. (eds) Information Systems and Technologies. WorldCIST 2023. Lecture Notes in Networks and Systems, vol. 799. Springer, Cham. ISBN 978-3-031-45641-1. ISSN 2367-3370.

  DOI: `https://doi.org/10.1007/978-3-031-45642-8_8`.

  Authors Affiliation:

  - Murilo M. Breve. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - José Rufino. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.

  Copyrights and Licensing: DBLP, EI Compendex, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, WTI AG, and zbMATH.

  SCImago Journal Rank (SJR) 2023: 0.171, Q4 for Computer Science.

  CiteScore Rank 2023: 0.9 for Computer Science.

  Copyrights: The publication *Lecture Notes in Networks and Systems*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040671434833 is presented in Appendix A.

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis and design of experiments. Revision of the final manuscript. Funding.

- Chapter 12 is based on [14]:

  Balsa, C., Breve, M.M., Rodrigues, C.V., Rufino, J. (2023). *Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset*. Computation, 11(5): 98.

  DOI: `https://doi.org/10.3390/computation11050098`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Murilo M. Breve. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Carlos V. Rodrigues. Vestas Wind Systems A/S, Design Centre Porto, Leça do Balio, Portugal.
  - José Rufino. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.

  Indexing: BibCnrs, CAPlus/SciFinder, CNKI, CNPIEC, dblp Computer Science Bibliography, Dimensions, DOAJ, EBSCO, SCOPUS, Gale, Inspec, J-Gate, OpenAIRE, OSTI (U.S. Department of Energy), PATENTSCOPE, ProQuest, and Web of Science (ESCI).

  Journal Impact Factor (JIF) 2024: 2, Q2 for Mathematics.

  SCImago Journal Rank (SJR) 2024: 0.44, Q2 for Applied Mathematics.

  CiteScore Rank 2023: 3.5, Q2 for Applied Mathematics.

  Copyrights: Information about permissions can be found in the journal website: `https://www.mdpi.com/authors/rights`, where the following statement is placed:

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis, computational implementation, design of experiments. Drafting of the original and revision of the final manuscript. Funding.

- Chapter 13 is based on [15]:

  Balsa, C., Dupuis, H., Breve, MM., Guivarch, R., Rufino, J. (2024). *Optimal Latent Variables Number for the Reconstruction of Time Series with PLSR*. In: Garcia, M.V., Gordón-Gallegos, C., Salazar-Ramírez, A., Nuñez, C. (eds) Proceedings of the International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2023). CSEI 2023. Lecture Notes in Networks and Systems, vol. 775. Springer, Cham. ISBN 978-3-031-69227-7. ISSN 2367-3370.

  DOI: `https://doi.org/10.1007/978-3-031-69228-4_13`.

  Authors Affiliation:

  - Carlos Balsa. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - Hugo Dupuis. Université de Toulouse - INP - IRIT, Toulouse, France.
  - Ronan Guivarch. Université de Toulouse - INP - IRIT, Toulouse, France.
  - Murilo M. Breve. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.
  - José Rufino. Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Portugal.

Indexing: DBLP, EI Compendex, INSPEC, Norwegian Register for Scientific Journals and Series, SCImago, SCOPUS, WTI AG, and zbMATH.

SCImago Journal Rank (SJR) 2023: 0.171, Q4 for Computer Science.

CiteScore Rank 2023: 0.9 for Computer Science.

Copyrights: The publication *Lecture Notes in Networks and Systems*, which belongs to the publisher Springer Nature, authorises the publication of this book chapter as part of the thesis. The license n.º 6040680154524 is presented in Appendix A.

Specific contribution to the publication: Conceptualisation, methodology, and research. Formal analysis and design of experiments. Revision of the final manuscript.

Summary

The present thesis, entitled *Applications of Computational Mathematics to Simulation and Data Analysis: two-dimensional vortex dynamics and meteorological data reconstruction*, is the result of the work developed by Carlos Balsa during his Ph.D. studies in Mathematics and Applications granted by the University of Santiago de Compostela.

This course started in October 2022 and was carried out under the supervision of Professor María Victoria Otero Espinar, from the Department of Statistics, Mathematical Analysis and Optimisation of the University of Santiago de Compostela, and Professor Sílvio Gama, from the Department of Mathematics, Faculty of Science, University of Porto.

The research was carried out in two main domains of Applied Computational Mathematics to Simulations and Data Analysis. The first concerns the two-dimensional dynamics of vortices and is developed along Part 1, and the second concerns the reconstruction of meteorological data and is the subject of Part 2.

Part 1 - Two-Dimensional Vortex Dynamics

This part investigates the optimal control of a passive particle moving on a spherical surface under the influence of multiple point vortices. The aim is to compute a trajectory that minimises energy consumption, while fulfilling constraints on departure and arrival locations and the duration of the motion. The first part of the thesis is devoted to this research topic and is named Part 1 - *Two-Dimensional Vortex Dynamics*. This part includes chapters 4 through 9.

The research focuses on an optimal control method based on a direct approach that is an alternative to the Pontryagin's Maximum Principle method [95]. In this approach,

the problem is first discretised and then optimised. The motion of the passive particle is initially reformulated as a finite-horizon optimal control problem on a spherical surface. The temporal domain is then discretised into n subintervals all with the same amplitude, over which the control inputs are modelled as piecewise constant functions, enabling a tractable numerical approximation of the continuous-time dynamics induced by the vortex system. The discretised problem results in a nonlinear optimisation problem that is solved numerically using a shooting method that finds the optimal control variables. In each subinterval, corresponding to a constant control variable, the vortex dynamics is integrated by a fourth-order Runge-Kutta method.

Each chapter corresponds to an independent study that has already been published. Each addresses a different aspect of the control method or an application.

Chapters 4 and 5 explore modelling the system dynamics using spherical and Cartesian coordinates, respectively.

Chapter 6 mathematically formalises the optimal control problem and shows that the solution is independent of the type of coordinates used to model the system dynamics.

Chapters 7 and 8 present conceptual applications for the developed optimal control method.

Finally, Chapter 9 presents a complementary study on the effect of the sphere's rotation on the proposed optimisation method.

A detailed overview of each chapter is presented below.

Chapter 4: Modelling the Dynamics with Spherical Coordinates

This chapter presents an introductory study that models the dynamics of a passive particle advected by the velocity field induced by two-dimensional point vortices on a spherical surface. The analysis is conducted within a Lagrangian framework using spherical coordinates.

In Section 4.1, we outline the derivation of the equations in spherical coordinates that govern the dynamics of the point vortices $(R\,, \theta_i(t)\,, \phi_i(t))\,$, for $i = 1, 2, \ldots,\ N$, in the sphere with radius R, where $\theta_i(t)$ is the colatitude (or polar) angle and $\phi_i(t)$ is the longitude (or azimuthal) angle of the vortex i position at the time t.

The time evolution of the N point vortices, assuming known their positions at $t = 0$, is described by the set of ordinary differential equations ($i = 1, 2, \ldots, N$):

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})} \,,$$

$$\sin(\theta_i)\,\dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \, \frac{\sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})} \,,$$

where γ_{ij}, representing the central angle between the ith and the jth vortex points, is such that

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j) \,,$$

and k_j is the circulation of the vortex j.

A passive particle is, by definition, a point vortex with circulation set to zero. Thus, the dynamics of a system with P passive particles advected by N point vortices is given by the previous equations together with the equations for the passive particles

$$\dot{\theta}_p = -\frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,,$$

with $p = N+1, N+2, \ldots, N+P$, and the respective initial conditions.

The formulation of the control problem for the case of a single passive particle is done in Section 4.2. This problem is formulated from the equations of the passive particle ($P = 1$), adding to their right-hand sides the angular controls u_θ and u_ϕ, respectively, i.e.

$$\dot{\theta}_p = -\frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\theta \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\phi \,,$$

The functional to be minimised is

$$f \equiv \int_0^T \left((u_\theta(t))^2 + (u_\phi(t))^2 \right) \, dt \,.$$

The minimisation of f corresponds to the optimal reduction of the cost of moving the particle between the initial and final points so that the particle reaches the destination at the prescribed final time T.

Section 4.3 is dedicated to the numerical strategy, based on a direct approach, for determining the control that drives the passive particle from its initial point to its final destination, leading to the minimisation of the square of the L^2-norm of the control function.

The results of experimental tests executed with the number of vortices changing from one to three ($N = 1, 2, 3$) are presented in the following sections. In all results, quasi-optimal solutions are obtained since they allow the particle to move between the two fixed points on the sphere's surface.

Chapter 5: Modelling the Dynamics with Cartesian Coordinates

In this chapter, we explore modelling the system's dynamics composed of the passive particle and N point vortices using Cartesian coordinates.

The position of a vortex on a sphere is given by the vector $\mathbf{x}_j$ that points from the centre of the sphere to the vortex location $\mathbf{x}_j = (x_j, y_j, z_j)$ on the spherical surface $||\mathbf{x}_j|| = R$. The position of the vortex is given by

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_i}{||\mathbf{x}_i - \mathbf{x}_j||^2} \,, \qquad\qquad i = 1, 2, \ldots, N \,,$$

with the respective initial conditions, where k_j is the circulation of vortices j, and N is the total number of point vortices on the sphere.

A passive particle is, by definition, a point vortex with zero circulation ($k = 0$). The dynamics of a system with P passive particles advected by N point vortices is given by the previous equation, together with the equations for the passive particles

$$\dot{\mathbf{x}}_p = \frac{1}{2\pi R} \sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_p}{||\mathbf{x}_p - \mathbf{x}_j||^2} \ , \qquad p = N+1, N+2, \ldots, N+P \, ,$$

with the respective initial conditions. These equations are used in Section 5.1, which is devoted to the control problem underlying the determination of the optimal trajectory of the passive particle. It is also demonstrated that for the particle's motion to occur on the sphere, the control vector $\mathbf{U}$ exerted on the particle must be orthogonal to its position vector $\mathbf{x}$.

The transformation of the control problem into a nonlinear optimisation problem (NOP) in the case of a single vortex ($N = 1$) is detailed in Section 5.2. The discretisation of the objective function by the rectangular rule leads to the following approximation

$$\int_0^T \|\mathbf{U}\|^2 \, dt \approx \Delta t \sum_{j=0}^{n-1} \|\mathbf{u}_j\|^2 \ ,$$

where each $\mathbf{u}_j$ $(j = 0, 1, 2, \cdots, n-1)$ corresponds to the vector of control exercised in the subinterval $[t_{j-1}\ t_j)$. All these subintervals have amplitudes equal to $\Delta t = T/n = (t_n - t_0)/n$, where T is the time available to perform the displacement between the two fixed points.

In Section 5.3, the problem of a flow induced by multiple vortices ($N = 2, 3, 4$) is treated. In all the cases, the NOP resulting from the direct approach is solved numerically using the built-in Matlab `fmincon` function, which includes several optimisation methods. The comparison of two methods: Interior Point and the Active Set is presented. Both could find optimal trajectories regardless of the time discretisation level of the problem and the number of vortices involved. The computation times required by both methods are of the same order of magnitude. However, the Active Set requires more time when the number of control parameters to be determined is higher.

The effect of improving the number of constant controls, n, from one to four is also analysed. The results show that increasing the number of control parameters allows greater freedom in the trajectory choice, minimising the energy used for the displacement. On the

other hand, the computation time is higher.

For all the control problems analysed ($N = 2, 3, 4$), it is possible to show the existence of near or quasi-optimal trajectories for the passive particle.

Chapter 6: Exploring and Comparing Controls

The possibility of modelling the problem through two different coordinate systems raises the question of whether the solution depends on the chosen coordinates or whether it is unique regardless of the coordinate system. The study presented in this chapter focuses on the answer to this question. The formulation of the problem is deduced in spherical and Cartesian coordinates. It is then analytically shown how it is possible to convert the problem formulation from one coordinate system to another.

In Section 6.1, we undertake the derivation of equations governing the dynamics of the passive particle in Cartesian coordinates. For the case of a single vortex located at the north pole of the sphere, the dynamics of the controlled passive particle is governed by the following equation system:

$$\begin{cases} \dot{x} = -\frac{k}{2\pi}\frac{y}{x^2+y^2+(R-z)^2} + \alpha\, y \\ \\ \dot{y} = \frac{k}{2\pi}\frac{x}{x^2+y^2+(R-z)^2} - \alpha\, x + \beta\, z \\ \\ \dot{z} = -\beta\, y \end{cases}$$

complemented with the initial condition corresponding to the initial position of the particle. In this equation, the control parameters are $\alpha(\cdot)$ and $\beta(\cdot)$.

Section 6.2 explores the modelling of dynamics in spherical coordinates. For the case of a single vortex located at the north pole of the sphere, the equation system responsible for the dynamics of the passive particle is deduced as a function of Cartesian controls $\alpha(\cdot)$ and $\beta(\cdot)$:

$$\begin{cases} \dot{\theta} = \beta \sin\phi \\ \\ \dot{\phi} = \frac{k}{4\pi R^2}\frac{1}{1-\cos\theta} - \alpha + \beta\cos\phi\cot\theta \end{cases}$$

Using these two systems of equations allows us to directly compare the solutions obtained

with the two types of coordinates. In the same way, the equations governing the system are obtained for the case of $N = 1, 2$ and 3 vortices.

The formulation of control problems is detailed in Section 6.3. Three optimal control problems are derived. The first uses Cartesian coordinates with Cartesian controls, the second uses spherical coordinates with angular controls, and the third uses spherical coordinates and Cartesian controls. All of them are transformed into nonlinear optimisation problems using the same direct approach.

Computational experiments obtained from the numerical solution of the three optimisation problems for the case of $N = 1, 2, 3$ vortices are outlined in Section 6.4. The results suggest that each formulation produces a feasible solution. Each formulation allows the passive particle to attain the desired displacement within the allocated time. These three versions of the same problem affirm the effectiveness of the algorithm used to convert the optimal control problem into an optimisation problem. The numerical resolution of these formulations through a direct approach consistently yields near or quasi-optimal solutions, irrespective of the chosen coordinate system.

Chapter 7: Application to Conceptual Autonomous Technology

This chapter centres around a conceptual mathematical toy model that describes the behaviour of autonomous technology (ATs) responsible for the garbage collection done by autonomous vehicles (AVs) in oceanic settings. This simplified mathematical model combines conceptual insights into the physical world with the symbolic structures of mathematics, thus serving as a valuable pedagogic instrument to engage students and emphasise the significance of mathematics and physics in addressing environmental concerns. Atmospheric or oceanic circulation is modelled using a point vortex on a sphere, and the ATs displacements are modelled through optimal control of the trajectory of passive particles.

Section 7.1 presents the simplified mathematical model for ATs, based on Cartesian Coordinates. In the case of the the dynamics of a passive particle (AV) in a spherical flow with motion induced by two vortices, $N = 2$, initially localised at $\mathbf{x}_1(0) = V_{01}$ and $\mathbf{x}_2(0) = V_{02}$, with circulations k_1 and k_2, the position of the two vortices ($\mathbf{x}_1$ and $\mathbf{x}_2$) are governed

by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{1}{2\,\pi\,R}\; k_2\, \frac{\mathbf{x}_2 \times \mathbf{x}_1}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{1}{2\,\pi\,R}\; k_1\, \frac{\mathbf{x}_1 \times \mathbf{x}_2}{\|\mathbf{x}_2 - \mathbf{x}_1\|^2} \end{cases}$$

with the respective initial conditions $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$. The controlled AV, initially at $\mathbf{x}(0) = P_0$, is governed by the equation

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_1\|^2} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_2\|^2}\right) + \alpha_i\;(y, -x, 0) + \beta_i\;(0, z, -y)\,,$$

for $i = 0, 1, \ldots, n-1$, where α_i and β_i are the controllers.

In Section 7.2, the computational simulations using the model is examined, specifically focusing on the cases of ocean currents induced by one and two point vortices. The proposed direct approach computes the 30 AVs' trajectories for transporting the collected trash. The results indicate the existence of multiple trajectories for transporting garbage particles. However, energy costs are lower for the most direct trajectories between the starting and destination points. The number of controls corresponds to increased vehicle autonomy and can help define trajectories that minimise energy costs.

Chapter 8: Application to Steering Oceanic Debris to a Targeted Region

This chapter presents a second application of the system of point vortices and passive particles to model the trash collection in the ocean by autonomous vehicles. But this time, modelling the dynamics of AVs and ocean currents is done using spherical coordinates.

As in the previous chapter, it is assumed to be an idealised scenario where garbage collection is done by a set of autonomous vehicles initially located where the garbage has a higher density. Let's assume that the autonomous vehicles are relatively small, such that they can be treated as passive particles, meaning they do not alter the flow's local structure of the velocity field.

Once the holds of these autonomous vehicles are full, they must travel to a specific area/location to deposit the collected material there. The movement of these vehicles uses the ocean current as the primary source of displacement. It has a system of piecewise

constant controls, based on a small number of pre-defined switching points, whose sum of the squares of the amplitudes (objective function) is minimal. Here, the motion of point vortices on a sphere generates the ocean current.

Section 8.1 introduces the equations that model the dynamics of AVs and point vortices on a sphere. Section 8.2 presents the numerical method to solve the optimal control problem in determining AVs' trajectories.

In the two vortices ($N = 2$) problem, the dynamics of the vortices positions $V_1(\tau) = (\theta_1, \phi_1)$ and $V_2(\tau) = (\theta_2, \phi_2)$ are given by

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi} \frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \\ \dot{\phi}_1 = -\frac{k_2}{4\pi} \frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{\sin(\theta_1)(1-\cos(\gamma_{12}))} \\ \\ \dot{\theta}_2 = -\frac{k_1}{4\pi} \frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \\ \dot{\phi}_2 = -\frac{k_1}{4\pi} \frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{\sin(\theta_2)(1-\cos(\gamma_{21}))} \end{cases}$$

with the given initial conditions $V_1(0) = (\theta_{10},\, \phi_{10})$ and $V_2(0) = (\theta_{20},\, \phi_{20})$. The value of $\cos(\gamma_{12})$ and $\cos(\gamma_{21})$ are computed in agreement with

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,.$$

The interaction between the two vortices and the control significantly influences the dynamics of the passive particle, as depicted in

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi}\left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\theta \\ \\ \dot{\phi}_p = \frac{1}{4\pi}\left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\phi \end{cases}$$

with the initial conditions $P_0 = (\theta_{p0},\, \phi_{p0})$. The control applied to the passive particle is represented by the vector $\mathbf{U} = (u_\theta\ u_\phi)$.

Section 8.3 discusses applying the mathematical model to the problem of ocean debris removal. The analysis includes cases where one and two vortices ($N = 1, 2$) induce ocean

currents. In both cases, the results indicate the existence of multiple trajectories for transporting garbage particles. However, energy costs are lower for the most direct trajectories between the starting and destination points.

Chapter 9: Controlled Particle Motion on a Rotating Sphere

This chapter presents a complementary study to the previous ones that explores the complex motion of a passive particle within a flow induced by two-point vortices located on the surface of a rotating sphere.

The problem of controlling the displacement of a passive particle on a rotating spherical surface is analysed. For this purpose, two vortex points are considered to induce movement. The optimal control problem corresponding to the displacement of a particle immersed in a flow induced by two vortex points on the surface of a rotating sphere is formulated in Cartesian coordinates.

Section 9.1 introduces the equations governing the displacement of vortices and the passive particle, and formulates the corresponding optimisation control problem.

It is considered two point vortices, denoted by $\mathbf{x}_i \equiv \mathbf{x}_i(t) = (x_i(t), y_i(t), z_i(t))\,,\ i = 1, 2\,,$ initially located at $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$. It is assumed that these vortices reside on the surface of a sphere undergoing rotation about the $z-$axis. The dynamics of these two point vortices is given by the differential equations:

$$\begin{cases} \dot{\mathbf{x}}_1 = \dfrac{k_2}{4\,\pi\,R}\,\dfrac{\mathbf{x}_2 \times \mathbf{x}_1}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}_1} + \Omega\,(-y_1, x_1, 0) \\ \\ \dot{\mathbf{x}}_2 = \dfrac{k_1}{4\,\pi\,R}\,\dfrac{\mathbf{x}_1 \times \mathbf{x}_2}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}_2} + \Omega\,(-y_2, x_2, 0) \end{cases}$$

where R is the radius of the sphere, k_i is the circulation of the vortex i (i=1,2), and Ω is the angular velocity of the sphere rotation. The position of the controlled passive particle $\mathbf{x} \equiv (x, y, z)$ is then governed by

$$\dot{\mathbf{x}} = \frac{1}{4\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}}\right) + \Omega\,(-y, x, 0) + \mathbf{U}^c\,,$$

with the initial condition $\mathbf{x}(0)$. Here, $\mathbf{U}^c \equiv (u_x, u_y, u_z)$ is the control acting in the passive particle that maintains the motion on the spherical surface.

Section 9.2 outlines the numerical methodology based on the direct approach to address the optimal control problem. Numerical results, dependent on the number of constant controls and sphere rotation speed, are discussed in Section 9.3. It is observed that changing the sphere's rotation speed changes the solution's nature, allowing for displacement with either reduced energy requirements or, conversely, increased energy demands. Typically, when the rotation speed is relatively high, the particle must complete multiple rotations around the sphere before reaching the destination.

Part 2 - Meteorological Data Reconstruction

Regarding the data analysis research domain, the present study focuses on analysing time series from meteorological stations to fill significant gaps derived from missing records using the Analog Ensemble (AnEn) method [85]. In this context, an attempt will be made to develop a methodology that takes advantage of the maximum possible information available in the predictor stations and is correct and computationally efficient. The second part of the thesis is devoted to this research topic and is denoted Part 2 - *Meteorological Data Reconstruction*. This part includes chapters 10 through 13.

The work developed in Part 2 of this thesis focuses on using the AnEn method with many predictor time series. On the one hand, many predictor series contribute more information and thus can improve the quality of reconstruction (or prediction). On the other hand, the AnEn method does not work well when the number of predictor time series is high. It is, therefore, essential to have techniques that allow the AnEn method to be compatible with many predictor series, which may come from different variables or locations.

Each chapter of Part 2 also corresponds to an independent study that has already been published. Each addresses a different aspect of developing AnEn-based methods through dimension reduction techniques. Chapter 10 is a preliminary study that explores the application of Principal Component Analysis (PCA) to reduce the dimension of the predictor dataset. In Chapter 11, the PCA technique is combined with the classical AnEn method and a K-means cluster-based variant. Chapter 12 presents different approaches, based on

PCA and Partial Least Squares (PLS), that reduce the dimension of the predictor dataset without losing its richness and diversity of information. Chapter 13 is a complementary study comparing the most popular criteria used in choosing latent variables resulting from the PLS method.

Below, the contents of each chapter are described in more detail.

Chapter 10: Analogue Ensembles of Principal Components

This chapter presents an exploratory study where the Principal Components Analysis (PCA) is combined with the Analog Ensembles (AnEn) methods into a new technique called PCAnEn. This method allows for reconstructing meteorological data from other time series, called the predictors. The interest of this combination is to enable the use of the AnEn method with a large number of predictor series, and thus not lose the information they contain.

First, PCA is applied to the input data to reduce the dataset's dimension. Then, AnEn is used to reconstruct the data from a station using the neighbouring station's data. This allows more stations/information, but without compromising processing time.

Section 10.1 is devoted to introducing the AnEn method and its different variants. The other metrics used to determine the analogues and the ways (dependent or independent) in which the predictor series can be used are presented. In previous studies, it was found that the dependent form leads to more accurate data reconstructions. A variant of the AnEn method is also introduced, which previously clusters the analogues using the K-means method. This variant significantly reduces the calculation times required to determine the analogues.

Section 10.2 is devoted to characterising the dataset used and introducing the application of the PCA technique to reduce the dimension of the predictor dataset to be used by the AnEn method.

The PCA technique identifies the dimensions along which the data are most dispersed. In this way, we can identify the dimensions that best differentiate the dataset under analysis: its principal components. This can be achieved by the singular value decomposition

of the scaled data matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, which is given by

$$\mathbf{H} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$. The diagonal matrix $\mathbf{\Sigma}$ contains the singular values σ_i of $\mathbf{H}$, for $i = 1, \ldots, n$, where $\sigma_1 > \sigma_2 > \ldots > \sigma_n$. The right singular vectors v_i are the *principal component directions* of $\mathbf{H}$.

The vector

$$\mathbf{z}_1 = \mathbf{H}\mathbf{v}_1$$

has the largest sample variance, given by σ_1^2/m, among all normalised linear combinations of the columns of $\mathbf{H}$. Vector $\mathbf{z}_1$ represents the first new variable called the first principal component (PC_1). The second principal component (PC_2) is $\mathbf{z}_2 = \mathbf{H}\mathbf{v}_2$ because $\mathbf{v}_2$ corresponds to the second largest variance (σ_2^2/m), and the remaining principal components are defined similarly. The new variables are linear combinations of the columns of $\mathbf{H}$, i.e, they are linear combinations of the normalized variables $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$,

$$\mathbf{z}_i = v_{1i}\mathbf{h}_1 + v_{2i}\mathbf{h}_2 + \ldots + v_{ni}\mathbf{h}_n\,, \quad \text{for} \quad i = 1, 2, \ldots, n\,,$$

where the coefficients v_{ji} , $j = 1, 2, \ldots, n$ (called *loadings*) are the elements of the vector $\mathbf{v}_i$. The magnitude of a coefficient is related to the relative importance of the corresponding variable in the principal component.

The substitution criterion for the original variables by a few new variables must consider the influence of the new variables on the variance of the original data. This influence is directly proportional to the magnitude of the corresponding singular values. The first few principal components (PCs), corresponding to the most significant singular values, are expected to account for a large proportion of the total variance, so they are all needed for future analyses.

Section 10.3 is dedicated to computational experiments with the new PCAnEn method to reconstruct meteorological variables of a given station. The tests were developed using MATLAB and R software. Some error metrics are also introduced to evaluate the accuracy of the reconstructions.

The results show that the PCAnEn method offers better hindcasting accuracy than the classical AnEn method. The data reconstruction of a single station using the nearest

station is successful. However, the choice of predictor stations must consider the proximity and correlation between them.

Regarding computational performance, the PCAnEn method allows for a considerable reduction in processing time compared to the classical AnEn method. It was also verified that the implementation in MATLAB is faster than the implementation in R .

Chapter 11: Cluster-Based AnEn method Combined with PCA

The present chapter explores combining the PCAnEn method with the K-means clustering method (which results in the new PCClustAnEn variant). Thus, data from multiple predictor time series were reduced to one or two principal components (PCs) that are then used (instead of the original variable records) by the original AnEn method or its ClustAnEn variant to reconstruct the missing data. The PCAnEn and PCClustAnEn techniques were also compared regarding numerical accuracy and computational efficiency.

Section 11.1 describes the dataset used and points out the correlations between meteorological variables and stations. Sections 11.2 introduce the PCA technique as stated in the previous chapter and propose a criterion based on the standard deviation to choose the principal components (PCs). Principal components (PCs) with standard deviations greater than 1 exhibit higher variance—and thus contain more information—than the original normalised variables, which each have a standard deviation of 1. These PCs are selected as the new predictor variables in place of the original dataset.

Section 11.3 describes the combination of the AnEn and K-means methods at the base of the ClustAnEn method. When applied to the chosen PCs, we designate it as the PCClustAnEn method. Section 11.4 applies this new method to reconstruct meteorological variables. Both the PCAnEn and PCClustAnEn methods were put to the test for the reconstruction in a given meteorological station of the four variables selected for this study (wind speed: $WSPD$, gust speed: GST, atmospheric temperature: $ATMP$, and pressure: $PRES$), every 6 minutes, within the 10- to 18-hour period, during the whole year of 2019 (prediction period). The remaining stations (within a 30 km radius around the predicted location) were used as predictor stations, considering the training period from 2011 to 2018. The computational tests are developed simultaneously with MATLAB and R.

The results of the experiments show that the PCA technique improved the accuracy

of prediction without compromising the computational performance, since it is possible to increase the number of predictor stations without increasing the quantity of input time series. It is also shown that the efficacy of PCA is heavily influenced by the correlation between the time series of several predictors, as higher correlation allows for a higher proportion of information (variance) contained in the first components.

Furthermore, the two different implementations of the methods studied in MATLAB and R allow for double-checking the numerical results. The scalability of both codes was also studied in a medium-scale multicore system. The performance evaluation showed the superiority of the AnEn method, where PCA is combined with clustering.

Chapter 12: AnEn method with Dimension Reduction of the Predictor Dataset

The study presented in this chapter enriches the AnEn method with techniques that take advantage of many predictor variables through dimension reduction. We include reducing the original predictor dataset to a small number of new predictor variables, without losing essential information. This approach improves the quality of the reconstructions as well as their computational efficiency. We explore the dimension reduction through two alternative methods: principal component analysis (PCA) and partial least squares (PLS).

Section 12.1 presents the various reconstruction methods employed in this study. As we have seen in previous chapters, the PCA technique identifies the dimensions along which the data are most dispersed, i.e., with the most significant variance. In this way, we can determine the dimensions that best differentiate the dataset under analysis, i.e., its principal components, which are used as the new predictor variables. We present the combination of the AnEn-based methods with PCA to take advantage of the potentialities of the AnEn method and the dimension reduction provided by the PCA technique. Furthermore, PCA is combined with multivariate linear regression, giving rise to the principal components regression (PCR) method.

PCR is used to reconstruct missing meteorological data. Considering that the original dataset is represented by the matrix $\mathbf{X}$, where each column represents a time series with the records of a meteorological variable. The multivariate regression model on the original dataset $\mathbf{X}$ can be applied only if this matrix has full column rank. The near-collinearity

of columns can occur if there are highly correlated predictor variables. The PCR method circumvents the rank deficiency by replacing the original predictor variables $\mathbf{X}$ by their principal components (PCs) in the regression model.

Once the principal components, $\mathbf{Z} = \mathbf{XV}$, are obtained from matrix $\mathbf{X}$, a few of them (p) are used in the regression model to estimate $\mathbf{y}$.

Therefore, the PCR method consists of regressing $\mathbf{y}$ (the predicted time series) not on $\mathbf{X}$ itself but on the principal components matrix $\mathbf{Z} \in \mathbb{R}^{m \times p}$, assuming that p PCs have been previously selected. This implies solving the linear system by linear least squares

$$\mathbf{Zc} \approx \mathbf{y}\,,$$

whose solution is the parameter vector

$$\mathbf{c} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}\,.$$

The PCR regression model on $\mathbf{Z}$ is then given by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\mathbf{c}\,,$$

where $\tilde{\mathbf{y}} \in \mathbb{R}^{n-m}$ is, as before, the vector of the reconstructed/predicted values of $\mathbf{y}$ during the reconstruction/prediction period, $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{V} \in \mathbb{R}^{(n-m)\times p}$ contains the values of the p selected PCs along the reconstruction/prediction period and $\mathbf{c} \in \mathbb{R}^p$ is the parameter vector of the PCR.

The partial least squares (PLS) technique extracts from the set of predictor variables (represented by the matrix $\mathbf{X}$) a set of latent variables (represented by the columns of matrix $\mathbf{T}$) which have the best predictive power. These new predictor variables are obtained by maximising the covariance between the predictors and the predicted variable.

The latent variables that model $\mathbf{X}$ and $\mathbf{y}$ and best predict $\mathbf{y}$ results from the variable decompositions

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \qquad \text{and} \qquad \mathbf{y} = \mathbf{Rq}^T + \mathbf{f},$$

where $\mathbf{T} \in \mathbb{R}^{m\times p}$ and $\mathbf{R} \in \mathbb{R}^{m\times p}$ are the matrix with p latent vectors (also known as scores) extracted from $\mathbf{X}$ and $\mathbf{y}$, respectively; $\mathbf{P} \in \mathbb{R}^{q\times p}$ and $\mathbf{q} \in \mathbb{R}^p$ represent the loading vectors; the matrix $\mathbf{E} \in \mathbb{R}^{m\times q}$ and vector $\mathbf{f} \in \mathbb{R}^m$ represent the residuals, whose norms

are minimized. Additionally, the scores matrix $\mathbf{T}$ is orthogonal, that is, $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I}$. The regression model

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{c}^T,$$

where $\mathbf{c}^T \in \mathbb{R}^p$ denotes the regression vector, makes it possible to estimate $\mathbf{y}$ based on the latent variables $\mathbf{T}$. Additionally, we present the combination of the AnEn-based methods with the latent variables (LVs) obtained with PLS, resulting in the PLSAnEn and PLSClustAnEn methods.

Section 12.2 introduces the meteorological datasets that test and validate the different methods. A correlation study between variables and stations is also presented. Sections 12.3 and 12.4 focus on selecting principal components and latent variables. Several criteria are introduced to choose the new predictor variables.

Section 12.5 presents and discusses the numerical results of the tests performed with the various reconstruction methods. We present a comparative study of the performance of all these methods in a hindcasting problem, corresponding to the reconstruction of missing data in a given meteorological station using data from a set of predictor stations with different geographical locations. It shows that the results produced by the PLS-based techniques were slightly more accurate than those obtained with the PCA-based ones, especially in reconstructing or forecasting meteorological variables with a significant amount of oscillation, such as wind speed.

Section 12.6 analyses the computational performance of the same methods. Despite the accuracy of the PLSAnEn method, it is verified that it is very demanding from a computational point of view, which benefits from a parallel implementation. PLSClustAnEn, which combines the AnEn methods with the prior clustering of analogues, could be an alternative to the PLSAnEn method, as it is much more computationally efficient. It is also verified that the regression methods PCR and PLSR are very fast and allow for accurate reconstructions, particularly for highly correlated variables.

Chapter 13: Optimal Latent Variables Number for PLSR

The Partial Least Squares Regression (PLSR) is an efficient method for filling gaps in meteorological time series. It enables the reduction of the dimension of the predictor dataset to a reduced number of latent variables, without loss of significant information.

Due to this, the number of latent variables to be used is an essential aspect of the success of the PLSR method.

Section 13.1 revisits the partial least squares decomposition and the PLS regression (PLSR) method.

Section 13.2 describes the different criteria that can be applied to estimate the optimal number of latent variables. Four criteria are based on the Cross-Validation method: Q-squared method (Q^2), the Root Mean Squared Error of Cross-Validation (RMSECV), World's R and Osten's F. The Other two are based on the Information criteria: Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC). The remaining one is the Variable Importance in Projection (VIP).

Section 13.3 compares the criteria used to determine the optimal number of latent variables in the PLSR method applied to a meteorological dataset from the US National Data Buoy Centre (NDBC). The variables atmospheric pressure, temperature, and wind speed are reconstructed over a year in a single station using the neighbouring ones. The study involved a total of 41 different time series.

The results of the experiments indicate that the cross-validation-based criteria, such as the RMSECV or Osten's F, are efficient and give an amount of latent variables relatively close to the experimental optimal number; that is why these criteria are currently the most commonly used. However, these criteria have the disadvantage of being very computationally demanding.

The information criteria, such as the AIC and the BIC, also give interesting results and are often used for prediction. Its simplicity makes its computational costs negligible. However, its efficacy depends mainly on the threshold value. Currently, no studies have obtained a theoretical way to estimate its value.

Resumo

A presente tese, titulada "*Aplicacións da Matemática Computacional á Simulación e Análise de Datos: dinámica bidimensional de vórtices e reconstrución de datos meteorolóxicos*", é o resultado do traballo desenvolvido por Carlos Balsa durante os seus estudos no Programa de Doutoramento en Matemáticas e Aplicacións pola Universidade de A Coruña; a Universidade de Santiago de Compostela; a Universidade de Vigo; Universidade de Porto; Universidade de Trás-os-Montes e Alto Douro e Universidade do Minho.

Este traballo comezou en outubro de 2022 e levouse a cabo baixo a supervisión da profesora María Victoria Otero Espinar, do Departamento de Estatística, Análise Matemática e Optimización da Universidade de Santiago de Compostela, e do profesor Sílvio Gama, do Departamento de Matemáticas, Facultade de Ciencias da Universidade de Porto.

A investigación realizouse en dous dominios principais da matemática computacional aplicada á simulación e á análise de datos: a dinámica bidimensional de vórtices e a reconstrución de datos meteorolóxicos.

Parte 1 - Dinámica de Vórtices Bidimensionais

No primeiro ámbito, o problema que se estuda é o control do desprazamento dunha partícula pasiva sobre unha esfera, na cal varios puntos de vórtice inducen o movemento. Procúrase a traxectoria óptima que minimice a enerxía empregada no desprazamento e respecte as restricións impostas, como os puntos de saída e chegada e o tempo dispoñible para realizar o desprazamento. A primeira parte da tese está dedicada a este tema de investigación e denomínase Parte 1 – *Dinámica Bidimensional de Vórtices*. Esta parte comprende dende o capítulo 4 ao 9.

A investigación céntrase nun método de control óptimo baseado nun enfoque directo, que constitúe unha alternativa ao Principio do Máximo de Pontryagin [95]. Neste enfoque, o problema discretízase en primeiro lugar e despois optimízase. Inicialmente, o desprazamento da partícula pasiva transfórmase nun problema de control. A continuación, o tempo para realizar o desprazamento discretízase nun número fixo, n, de subintervalos, nos cales as variables de control permanecen constantes. O problema discretizado convértese nun problema de optimización non lineal, que se resolve numericamente mediante un método de shooting que determina as variables de control óptimas. En cada subintervalo, correspondente a unha variable de control constante, a dinámica dos vórtices intégrase mediante o método de Runge-Kutta de orde catro.

Cada capítulo corresponde a un estudo independente que xa foi publicado: Cada un aborda un aspecto distinto do método de control ou unha aplicación concreta. Os Capítulos 4 e 5 exploran a modelización da dinámica do sistema utilizando coordenadas esféricas e cartesianas, respectivamente. O Capítulo 6 formula matematicamente o problema de control óptimo e mostra que a solución é independente do tipo de coordenadas empregadas para modelizar a dinámica do sistema. Os Capítulos 7 e 8 presentan aplicacións conceptuais do método de control óptimo desenvolvido. Por último, o Capítulo 9 presenta un estudo complementario sobre o efecto da rotación da esfera no método de optimización proposto.

A continuación, descríbese con máis detalle o contido de cada capítulo.

Capítulo 4: Modelización da Dinámica con Coordenadas Esféricas

Este primeiro capítulo presenta un estudo introdutorio no que se modeliza a dinámica dunha partícula pasiva *advected* por vórtices puntuales bidimensionais nunha esfera, empregando coordenadas esféricas.

Na Sección 8.1, esbozamos a dedución das ecuacións en coordenadas esféricas que modelizan a dinámica dos vórtices $(R\,,\theta_j(t)\,,\phi_i(t))$, para $i = 1, 2, \ldots, N$, na esfera de radio R. A evolución temporal dos N vórtices puntuais, asumindo que se coñecen as súas posicións en $t = 0$, está rexida polo seguinte conxunto de ecuacións diferenciais ordinarias (para $i = 1, 2, \ldots, N$):

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_i - \phi_j)}{1-\cos(\gamma_{ij})} \,,$$

$$\sin(\theta_i)\,\dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)}{1-\cos(\gamma_{ij})} \,,$$

onde γ_{ij} , que representa o ángulo central entre os puntos de vórtice i e j, é tal que

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j) \,,$$

e k_j é a circulación do vórtice j.

Unha partícula pasiva é, por definición, un vórtice puntual cunha circulación igual a cero. Así, a dinámica dun sistema con P partículas pasivas *advected* por N vórtices puntuais vén dada polas ecuacións anteriores xunto coas ecuacións para as partículas pasivas

$$\dot{\theta}_p = -\frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})} \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})} \,,$$

con $p = N+1, N+2, \ldots, N+P$, e as correspondentes condicións iniciais.

A formulación do problema de control para o caso dunha única partícula pasiva faise na Sección 4.2. Este problema fórmase a partir das ecuacións da partícula pasiva ($P = 1$), engadindo nos seus lados dereitos os controis angulares u_θ e u_ϕ , respectivamente, é dicir:

$$\dot{\theta}_p = -\frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})} + u_\theta \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})} + u_\phi \,,$$

O funcional a minimizar é

$$f \equiv \int_0^T \left((u_\theta(t))^2 + (u_\phi(t))^2\right)\, dt\,.$$

A minimización de f corresponde á redución óptima do custo de mover a partícula entre os puntos inicial e final, garantindo que chegue ao seu destino no tempo final prescrito T ,.

A Sección 4.3 está dedicada á estratexia numérica, baseada nun enfoque directo, para determinar o control que guía a partícula pasiva desde o seu punto inicial ata o seu destino final, minimizando a norma cadrática L^2 do control.

Os resultados das probas experimentais realizadas con un número de vórtices que varía de un a tres preséntanse nas seguintes seccións. En todos os casos, obtéñense solucións cuasi-óptimas, xa que todas permiten o movemento da partícula entre os dous puntos fixos na superficie da esfera.

Capítulo 5: Modelización da Dinámica con Coordenadas Cartesianas

Neste capítulo exploramos a modelización da dinámica do sistema composto pola partícula pasiva e N vórtices puntuais empregando coordenadas cartesianas.

A posición dun vórtice sobre a esfera vén dada polo vector $\mathbf{x}_j$ que apunta desde o centro da esfera ata a localización do vórtice $\mathbf{x}_j = (x_j, y_j, z_j)$ na superficie esférica $||\mathbf{x}_j|| = R$. A posición do vórtice vén dada por

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_i}{||\mathbf{x}_i - \mathbf{x}_j||^2}\ , \qquad\qquad i = 1, 2, \ldots, N\,,$$

coas correspondentes condicións iniciais, onde k_j é a circulación do vórtice j , e N é o número total de vórtices puntuais na esfera.

Como unha partícula pasiva é, por definición, un vórtice puntual con circulación $k = 0$. a dinámica dun sistema con P partículas pasivas *advected* por N vórtices puntuais está

dada pola ecuación anterior xunto coas ecuacións para as partículas pasivas

$$\dot{\mathbf{x}}_p = \frac{1}{2\pi R}\sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_p}{||\mathbf{x}_p - \mathbf{x}_j||^2} \ , \qquad p = N+1, N+2, \ldots, N+P \, ,$$

coas correspondentes condicións iniciais. Estas ecuacións úsanse na Sección 5.1, que está dedicada ao problema de control subxacente na determinación da traxectoria óptima da partícula pasiva. Tamén se demostra que, para que o movemento da partícula ocorra na esfera, o vector de control $\mathbf{U}$ aplicado sobre a partícula debe ser ortogonal ao seu vector de posición $\mathbf{x}$.

A transformación do problema de control nun problema de optimización non lineal (NOP) no caso dun único vórtice ($N = 1$) descríbese con detalle na Sección 5.2. A discretización da función obxectivo mediante a regra do rectángulo lévanos á seguinte aproximación:

$$\int_0^T \|\mathbf{U}\|^2 \, dt \approx \Delta t \sum_{j=0}^{n-1} \|\mathbf{u}_j\|^2 \ ,$$

onde cada $\mathbf{u}_i$ $(i = 0, 1, 2, \cdots, n-1)$ corresponde ao vector de control exercido no subintervalo $[t_{i-1}\ t_i)$. Todos estes subintervalos teñen amplitudes iguais a $\Delta t = T/n = (t_n - t_0)/n$, onde T é o tempo dispoñible para realizar o desprazamento entre os dous puntos fixos.

Na Sección 5.3, trátase o problema dun fluxo inducido por múltiples vórtices ($N = 2, 3, 4$). En todos os casos, o NOP resultante do enfoque directo resólvese numericamente empregando a función `fmincon` de Matlab, que inclúe varios métodos de optimización. Preséntase a comparación entre dous métodos: *Interior Point* e *Active Set*. Ambos foron capaces de atopar traxectorias óptimas independentemente do nivel de discretización temporal do problema e do número de vórtices implicados. Os tempos de cálculo requiridos por ambos métodos son da mesma orde de magnitude. Porén, o método *Active Set* require máis tempo cando o número de parámetros de control a determinar é maior.

Tamén se analiza o efecto de mellorar o número de controis constantes, n, $n = 1, \cdots, 4$. Os resultados mostran que aumentar o número de parámetros de control permite unha maior liberdade na elección da traxectoria, minimizando así a enerxía utilizada no desprazamento. Por outra banda, o tempo de cálculo aumenta.

Para todos os problemas de control analizados ($N = 2, 3, 4$), é posible demostrar a

existencia de traxectorias próximas ou cuasi-óptimas para a partícula pasiva.

Capítulo 6: Exploración e Comparación de Controis

A posibilidade de modelizar o problema mediante dous sistemas de coordenadas diferentes suscita a cuestión de se a solución depende do sistema de coordenadas escollido ou se, pola contra, é única independentemente do sistema. O estudo presentado neste capítulo céntrase na resposta a esta cuestión. A formulación do problema dedúcese en coordenadas esféricas e cartesianas. Posteriormente, demóstrase analiticamente como é posible converter a formulación do problema dun sistema de coordenadas a outro.

Na Sección 6.1, realizamos a dedución das ecuacións que rexen a dinámica da partícula pasiva en coordenadas cartesianas. No caso dun único vórtice situado no polo norte da esfera, a dinámica da partícula pasiva controlada está rexida polo seguinte sistema de ecuacións:

$$\begin{cases} \dot{x} = -\frac{k}{2\pi}\frac{y}{x^2+y^2+(R-z)^2} + \alpha\, y \\ \\ \dot{y} = \frac{k}{2\pi}\frac{x}{x^2+y^2+(R-z)^2} - \alpha\, x + \beta\, z \\ \\ \dot{z} = -\beta\, y \end{cases}$$

coas condicións iniciais. Nesta ecuación, $\alpha(\cdot)$ e $\beta(\cdot)$ son os parámetros de control.

Na Sección 6.2 exploramos a modelización da dinámica en coordenadas esféricas. No caso dun único vórtice situado no polo norte da esfera, o sistema de ecuacións que describe a dinámica da partícula pasiva dedúcese como función dos controis cartesianas $\alpha(\cdot)$ e $\beta(\cdot)$.

$$\begin{cases} \dot{\theta} = \beta \sin\phi \\ \\ \dot{\phi} = \frac{k}{4\pi R^2}\frac{1}{1-\cos\theta} - \alpha + \beta\cos\phi\cot\theta \end{cases}$$

O uso destes dous sistemas de ecuacións permítenos comparar directamente as solucións obtidas cos dous tipos de coordenadas. Do mesmo xeito, obtéñense as ecuacións que rexen o sistema para o caso de N vórtices, $N = 1, 2, 3$.

A formulación dos problemas de control detállase na Sección 6.3. Formúlanse tres problemas de control óptimo. O primeiro emprega coordenadas cartesianas con controis

cartesianos, o segundo emprega coordenadas esféricas con controis angulares e o terceiro emprega coordenadas esféricas e controis cartesianos. Todos eles son transformados en problemas de optimización non lineal utilizando o mesmo enfoque directo.

Os experimentos computacionais obtidos a partir da solución numérica dos tres problemas de optimización para o caso de $N = 1, 2, 3$ vórtices son descritos na Sección 6.4. Os resultados suxiren que cada formulación produce unha solución viable. Cada formulación permite que a partícula pasiva alcance o desprazamento desexado dentro do tempo asignado. Estas tres versións do mesmo problema confirmam a eficacia do algoritmo empregado para converter o problema de control óptimo nun problema de optimización. A resolución numérica destas formulacións mediante un enfoque directo dá sistematicamente solucións próximas ou case óptimas, independentemente do sistema de coordenadas elixido.

Capítulo 7: Aplicación á Tecnoloxía Autónoma Conceptual

Este capítulo céntrase nun modelo matemático conceptual que describe o comportamento da tecnoloxía autónoma (ATs), responsable, por exemplo, da recollida de lixo realizada por vehículos autónomos (AVs) en contornas oceánicas. Este modelo matemático simplificado combina ideas conceptuais sobre o mundo físico coas estruturas simbólicas das matemáticas, servindo así como un valioso instrumento pedagóxico para involucrar aos estudantes e enfatizar a importancia das matemáticas e a física na abordaxe de problemas ambientais. A circulación atmosférica ou oceánica modelízase utilizando vórtices puntuais sobre unha esfera, e os ATs modelízanse mediante o control óptimo da traxectoria de partículas pasivas.

Na Sección 7.1 preséntase o modelo matemático simplificado para os ATs, baseado en coordenadas cartesianas. No caso da dinámica dunha partícula pasiva (AV) nun fluxo esférico inducido por dous vórtices, $N = 2$, inicialmente localizados en $\mathbf{x}_1(0) = V_{01}$ e $\mathbf{x}_2(0) = V_{02}$, con circulacións k_1 e k_2, a posición dos dous vórtices ($\mathbf{x}_1$ e $\mathbf{x}_2$) está rexida por

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{1}{2\,\pi\,R}\; k_2\, \frac{\mathbf{x}_2 \times \mathbf{x}_1}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{1}{2\,\pi\,R}\; k_1\, \frac{\mathbf{x}_1 \times \mathbf{x}_2}{\|\mathbf{x}_2 - \mathbf{x}_1\|^2} \end{cases}$$

coas correspondentes condicións iniciais $\mathbf{x}_1(0)$ e $\mathbf{x}_2(0)$. O AV controlado, inicialmente en

$\mathbf{x}(0) = P_0$, está rexido pola ecuación

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1\times\mathbf{x}}{\|\mathbf{x}-\mathbf{x}_1\|^2} + k_2\,\frac{\mathbf{x}_2\times\mathbf{x}}{\|\mathbf{x}-\mathbf{x}_2\|^2}\right) + \alpha_i\;(y,-x,0) + \beta_i\;(0,z,-y)\,,$$

para $i = 0, 1, \ldots, n-1$, onde α_i e β_i son os controladores.

Na Sección 7.2, examínanse simulacións computacionais utilizando o modelo, centrándose especificamente nos casos de correntes oceánicas inducidas por un e dous vórtices puntuais. As traxectorias realizadas polos 30 AVs para transportar o lixo recollido calcúlanse mediante o enfoque directo proposto. Os resultados indican a existencia de múltiples traxectorias para o transporte de partículas de lixo. Con todo, os custos de enerxía son menores para as traxectorias máis directas entre os puntos de inicio e destino. O número de controis corresponde a unha maior autonomía dos vehículos e pode axudar a definir traxectorias que minimicen os custos enerxéticos.

Capítulo 8: Aplicación á Condución de Residuos Oceánicos cara a unha Rexión Específica

Este capítulo presenta unha segunda aplicación do sistema de vórtices puntuais e partículas pasivas para modelizar a recollida de lixo no océano por vehículos autónomos. Pero neste caso, a modelización da dinámica dos AVs e das correntes oceánicas faise utilizando coordenadas esféricas.

Como no capítulo anterior, asúmese un escenario idealizado onde a recollida de lixo é realizada por un conxunto de vehículos autónomos inicialmente situados nas áreas con maior densidade de lixo. Supoñerase que os vehículos autónomos son relativamente pequenos, de modo que poden tratarse como partículas pasivas, o que significa que non alteran a estrutura local do campo de velocidade do fluxo.

Unha vez que as bodegas destes vehículos autónomos están cheas, deben desprazarse a unha área ou localización específica para depositar alí o material recollido. O movemento destes vehículos utiliza a corrente oceánica como a principal fonte de desprazamento. Conta cun sistema de controis constantes por partes, baseado nun pequeno número de puntos de cambio predefinidos, que minimizan a suma dos cadrados das amplitudes (funcional obxectivo). Neste caso, a corrente oceánica é xerada polo movemento de vórtices puntuais liv

sobre unha esfera.

Na Sección 8.1 introdúcense as ecuacións que modelizan a dinámica dos AVs e dos vórtices puntuais sobre a esfera. A Sección 8.2 presenta o método numérico empregado para resolver o problema de control óptimo implicado na determinación das traxectorias dos AVs.

En o problema de dous vórtices ($N = 2$), a dinámica das posicións dos vórtices $V_1(\tau) = (\theta_1, \phi_1)$ e $V_2(\tau) = (\theta_2, \phi_2)$ vén dada por

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi} \frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \\ \dot{\phi}_1 = -\frac{k_2}{4\pi} \frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{\sin(\theta_1)(1-\cos(\gamma_{12}))} \\ \\ \dot{\theta}_2 = -\frac{k_1}{4\pi} \frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \\ \dot{\phi}_2 = -\frac{k_1}{4\pi} \frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{\sin(\theta_2)(1-\cos(\gamma_{21}))} \end{cases}$$

coas condicións iniciais dadas $V_1(0) = (\theta_{10},\ \phi_{10})$ e $V_2(0) = (\theta_{20},\ \phi_{20})$. Os valores de $\cos(\gamma_{12})$ e $\cos(\gamma_{21})$ calcúlanse de acordo con

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,.$$

A interacción entre os dous vórtices e o control inflúe significativamente na dinámica da partícula pasiva, como se mostra en

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi}\left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\theta \\ \\ \dot{\phi}_p = \frac{1}{4\pi}\left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\phi \end{cases}$$

coas condicións iniciais $P_0 = (\theta_{p0}, , \phi_{p0})$. O control aplicado á partícula pasiva está representado polo vector $\mathbf{U} = (u_\theta\ u_\phi)$.

Na Sección 8.3 discútese a aplicación do modelo matemático ao problema da eliminación de residuos oceánicos. A análise inclúe casos onde as correntes oceánicas están inducidas por un ou dous vórtices. En ambos casos, os resultados indican a existencia de múltiples

traxectorias para o transporte de partículas de lixo. Non obstante, os custos enerxéticos son menores para as traxectorias máis directas entre o punto de partida e o destino.

Capítulo 9: Movemento Controlado dunha Partícula nunha Esfera en Rotación

Este capítulo presenta un estudo complementario aos anteriores que explora o movemento complexo dunha partícula pasiva dentro dun fluxo inducido por dous vórtices puntuais situados sobre a superficie dunha esfera en rotación.

Analízase o problema do control do desprazamento dunha partícula pasiva sobre unha superficie esférica en rotación. Para este propósito, considérase que o movemento está inducido por dous vórtices puntuais. O problema de control óptimo correspondente ao desprazamento dunha partícula inmersa nun fluxo inducido por dous vórtices puntuais sobre a superficie dunha esfera en rotación formúlase en coordenadas cartesianas.

Na Sección 9.1 son introducidas as ecuacións que rexen o desprazamento dos vórtices e da partícula pasiva, xunto coa formulación do problema de control de optimización.

Considéranse dous vórtices puntuais, denotados por $\mathbf{x}_i \equiv \mathbf{x}_i(t) = (x_i(t), y_i(t), z_i(t))$, $i = 1, 2$, inicialmente localizados en $\mathbf{x}_1(0)$ e $\mathbf{x}_2(0)$. Asúmese que estes vórtices están sobre a superficie dunha esfera que xira arredor do eixe z. A dinámica destes dous vórtices puntuais vén dada polas ecuacións diferenciais:

$$\begin{cases} \dot{\mathbf{x}}_1 = \dfrac{k_2}{4\,\pi\,R}\,\dfrac{\mathbf{x}_2 \times \mathbf{x}_1}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}_1} + \Omega\,(-y_1, x_1, 0) \\ \\ \dot{\mathbf{x}}_2 = \dfrac{k_1}{4\,\pi\,R}\,\dfrac{\mathbf{x}_1 \times \mathbf{x}_2}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}_2} + \Omega\,(-y_2, x_2, 0) \end{cases}$$

onde R é o raio da esfera, k_i é a circulación do vórtice i (i=1,2), e Ω é a velocidade angular da rotación da esfera. A posición da partícula pasiva controlada $\mathbf{x} \equiv (x, y, z)$ está gobernada por

$$\dot{\mathbf{x}} = \frac{1}{4\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}}\right) + \Omega\,(-y, x, 0) + \mathbf{U}^c\,,$$

coas condicións iniciais $\mathbf{x}(0)$. Aquí, $\mathbf{U}^c \equiv (u_x, u_y, u_z)$ é o control que actúa na partícula pasiva e mantén o movemento sobre a superficie esférica.

Na Sección 9.2 presentamos a metodoloxía numérica baseada no enfoque directo para abordar o problema de control óptimo. Os resultados numéricos, dependentes do número de controis constantes e da velocidade de rotación da esfera, discútense na Sección 9.3. Obsérvase que o cambio na velocidade de rotación da esfera altera a natureza da solución, permitindo desprazamentos con requisitos de enerxía reducidos ou, pola contra, maiores demandas enerxéticas. Normalmente, cando a velocidade de rotación é relativamente alta, a partícula debe completar múltiples rotacións arredor da esfera antes de alcanzar o destino.

Parte 2 - Reconstrución de Datos Meteorolóxicos

No ámbito da investigación en análise de datos, este estudo ten como obxectivo analizar series temporais obtidas de estacións meteorolóxicas para cubrir lagoas importantes provocadas pola falta de rexistros, empregando para iso o método Analog Ensemble (AnEn) [85]. Neste contexto, intentarase desenvolver unha metodoloxía que aproveite ao máximo posible a información dispoñible nas estacións predictoras e sexa correcta e computacionalmente eficiente. A segunda parte da tese dedícase a este tema de investigación e denomínase Parte 2 - *Reconstrución de Datos Meteorolóxicos*. Esta parte inclúe os capítulos 10 a 13.

O traballo desenvolvido na Parte 2 desta tese céntrase no uso do método AnEn con múltiplas series temporais predictoras. Por un lado, contar con moitas series predictoras achega máis información e pode mellorar a calidade da reconstrución (ou predición). Por outro lado, o método AnEn non funciona ben cando o número de series predictoras é elevado. É, polo tanto, esencial dispoñer de técnicas que fagan compatible o método AnEn cun gran número de series predictoras, que poden proceder de distintas variables ou localizacións.

Cada capítulo da Parte 2 tamén corresponde a un estudo independente que xa foi publicado. Cada un aborda un aspecto distinto do desenvolvemento de métodos baseados en AnEn a través de técnicas de redución de dimensión. O Capítulo 10 é un estudo preliminar que explora a aplicación da Análise de Compoñentes Principais (PCA) para reducir a dimensión do conxunto de datos predictivos. No Capítulo 11, a técnica PCA combínase co método AnEn clásico e unha variante baseada en clusters K-means. O

Capítulo 12 presenta diferentes enfoques, baseados en PCA e Mínimos Cadrados Parciais (PLS), que reducen a dimensión do conxunto de datos predictivos sen perder a súa riqueza e diversidade de información. O Capítulo 13 é un estudo complementario que compara os criterios máis empregados na elección das variables latentes resultantes do método PLS.

A continuación, resumimos con máis detalle os contidos de cada un dos capítulos.

Capítulo 10: Conxuntos Análogos de Compoñentes Principais

Este capítulo presenta un estudo exploratorio onde se combina a Análise de Compoñentes Principais (PCA) co método Analog Ensemble (AnEn) para crear unha nova técnica chamada PCAnEn. Este método permite reconstruír datos meteorolóxicos a partir doutras series temporais chamadas predictoras. O interese desta combinación radica en permitir o uso do método AnEn con un gran número de series predictoras, evitando así perder a información que conteñen.

Primeiro, aplícase PCA aos datos de entrada para reducir a dimensión dos conxuntos de datos. A continuación, utilízase AnEn para reconstruír os datos dunha estación empregando información das estacións veciñas. Isto permite utilizar máis estacións/información sen comprometer o tempo de procesamento.

A Sección 10.1 está dedicada a introducir o método AnEn e as súas distintas variantes. Preséntanse as diferentes métricas utilizadas para determinar os análogos e as formas (dependentes ou independentes) en que se poden empregar as series predictoras. Estudos previos demostraron que a forma dependente conduce a reconstrucións de datos máis precisas. Tamén se presenta unha variante do método AnEn que previamente agrupa os análogos empregando o método K-means. Esta variante reduce significativamente os tempos de cálculo necesarios para determinar os análogos.

A Sección 10.2 está dedicada a caracterizar o conxunto de datos utilizado e a introducir a aplicación da técnica PCA para reducir a dimensión do conxunto de datos predictivos empregado polo método AnEn.

A técnica PCA identifica as dimensións ao longo das cales os datos están máis dispersos. Deste xeito, podemos identificar as dimensións que mellor diferencian o conxunto de datos analizado, é dicir, os seus compoñentes principais. Isto consíguese mediante a descomposición en valores singulares da matriz de datos escalada $\mathbf{H} \in \mathbb{R}^{m \times n}$, dada por

$$\mathbf{H} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

onde $\mathbf{U} \in \mathbb{R}^{m\times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ e $\mathbf{V} \in \mathbb{R}^{n\times n}$ (para detalles, véxase por exemplo [53]). A matriz diagonal $\mathbf{\Sigma}$ contén os valores singulares σ_i de $\mathbf{H}$, para $i = 1, \ldots, n$, onde $\sigma_1 > \sigma_2 > \ldots > \sigma_n$. Os vectores singulares dereitos v_i son as direccións dos *compoñentes principais* de $\mathbf{H}$.

O vector

$$\mathbf{z}_1 = \mathbf{H}\mathbf{v}_1$$

ten a maior varianza mostral, dada por σ_1^2/m, entre todas as combinacións lineais normalizadas das columnas de $\mathbf{H}$. O vector $\mathbf{z}_1$ representa a primeira nova variable e denomínase o primeiro compoñente principal (PC_1). O segundo compoñente principal (PC_2) é $\mathbf{z}_2 = \mathbf{H}\mathbf{v}_2$, xa que $\mathbf{v}_2$ corresponde á segunda maior varianza (σ_2^2/m), e as restantes compoñentes principais defínense de maneira semellante.

As novas variables son combinacións lineais das columnas de $\mathbf{H}$, é dicir, combinacións lineais das variables normalizadas $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$,

$$\mathbf{z}_i = v_{1i}\mathbf{h}_1 + v_{2i}\mathbf{h}_2 + \ldots + v_{ni}\mathbf{h}_n\,, \quad \text{para} \quad i = 1, 2, \ldots, n\,,$$

onde os coeficientes v_{ji}, $j = 1, 2, \ldots, n$, (denominados cargas ou *loadings*) son os elementos do vector $\mathbf{v}_i$.

A magnitude dun coeficiente está relacionada coa importancia relativa da variable correspondente no compoñente principal.

O criterio para substituír as variables orixinais por unhas poucas novas variables debe ter en conta a influencia destas novas variables na varianza dos datos orixinais. Esta influencia é directamente proporcional á magnitude dos valores singulares correspondentes. Espérase que as primeiras compoñentes principais (PCs), que corresponden aos valores singulares máis significativos, representen unha grande proporción da varianza total, polo que son necesarios para análises futuras.

A Sección 10.3 está dedicada a experimentos computacionais co novo método PCA-nEn para reconstruír as variables meteorolóxicas dunha estación determinada. Os ensaios desenvolvéronse empregando os programas Matlab e R. Tamén se introducen métricas de erro para avaliar a precisión das reconstrucións.

Os resultados demostran que o método PCAnEn ofrece unha maior precisión nas reconstrucións históricas (*hindcasting*) que o método AnEn clásico. A reconstrución de datos dunha única estación empregando a estación máis próxima parece óptima. Non obstante, a elección das estacións predictoras debe ter en conta tanto a proximidade como a correlación entre elas.

En termos de rendemento computacional, o método PCAnEn permite reducir considerablemente o tempo de procesamento en comparación co método AnEn clásico. Tamén se verificou que a implementación en MATLAB é máis rápida que a implementación en R.

Capítulo 11: Método AnEn baseado en Clusters combinado con PCA

O presente capítulo explora a combinación do método PCAnEn co método de agrupamento K-means (que dá lugar á nova variante PCClustAnEn). Deste xeito, os datos de múltiples series temporais predictoras reducíronse a compoñentes principais (PCs), que logo son utilizados (en lugar dos rexistros das variables orixinais) polo método AnEn orixinal ou a súa variante ClustAnEn para reconstruír os datos ausentes. As técnicas PCAnEn e PCClustAnEn tamén se compararon en termos de precisión numérica e eficiencia computacional.

Na Sección 11.1 descríbese o conxunto de datos empregado e sinala as correlacións entre as variables meteorolóxicas e as estacións. A Sección 11.2 introduce a técnica PCA como se menciona no capítulo anterior e propón un criterio baseado na desviación estándar para escoller as compoñentes principais (PCs). As PCs con valores de desviación estándar superiores a 1 teñen máis variabilidade e, en consecuencia, máis información que as variables normalizadas orixinais, cuxa desviación estándar é igual a 1. Estas PCs escóllense como novas variables predictoras en lugar do conxunto de datos orixinal.

Na Sección 11.3 descríbese a combinación dos métodos AnEn e K-means na base do método ClustAnEn. Cando se aplica as PCs escollidas, denominámolo método PCClustAnEn. A Sección 11.4 aplica este novo método para reconstruír variables meteorolóxicas. Os métodos PCAnEn e PCClustAnEn probáronse para a reconstrución nunha estación meteorolóxica das catro variables seleccionadas para este estudo (velocidade do vento: $WSPD$, velocidade de refacho: GST, temperatura atmosférica: $ATMP$ e presión: $PRES$), cada 6 minutos, no período de 10 da mañá a 6 da tarde, durante todo o ano 2019 (período

de predición). As estacións restantes (nun radio de 30 km ao redor da estación predita/reconstruída) utilizáronse como estacións predictoras, considerando o período de adestramento de 2011 a 2018. As probas computacionais desenvolvéronse simultaneamente en Matlab e R.

Os resultados dos experimentos mostran que a técnica PCA mellorou a precisión da predición sen comprometer o rendemento computacional, xa que é posible aumentar o número de estacións predictoras sen incrementar a cantidade de series temporais de entrada. Tamén se demostra que a eficacia da PCA está fortemente influenciada pola correlación entre as series temporais de varios predictores, xa que unha maior correlación permite unha maior proporción de información (varianza) contida nas primeiras compoñentes.

Ademais, as dúas implementacións distintas dos métodos estudados en MATLAB e R permiten verificar os resultados numéricos. Tamén se estudou a escalabilidade de ambos códigos nun sistema multicore de escala media. A avaliación de rendemento mostrou a superioridade do método AnEn cando se combina PCA con agrupamento.

Capítulo 12: Método AnEn con Redución de Dimensións do Conxunto de Datos Predictivo

O estudo presentado neste capítulo contribúe a enriquecer o método AnEn con técnicas que aproveitan un gran número de variables predictoras mediante a redución de dimensións. Incluímos a redución do conxunto de datos preditores orixinal a un pequeno número de novas variables predictoras, sen perda de información esencial. Este enfoque mellora a calidade das reconstrucións e a súa eficiencia computacional. Exploramos a redución de dimensións mediante dous métodos alternativos: a análise de compoñentes principais (PCA) e os mínimos cadrados parciais (PLS).

Na Sección 12.1 presentamos os diversos métodos de reconstrución empregados neste estudo. Como vimos en capítulos anteriores, a técnica PCA identifica as dimensións nas que os datos presentan maior dispersión, é dicir, aquelas que teñen maior varianza. Deste xeito, podemos identificar as dimensións que mellor diferencian o conxunto de datos analizado, é dicir, as suas compoñentes principais que, á súa vez, se empregan como as novas variables predictoras. Presentamos a combinación dos métodos baseados en AnEn con PCA para aproveitar as potencialidades do método AnEn e a redución de dimensión proporcionada

pola técnica PCA. Ademais, a PCA combínase coa regresión lineal multivariante, dando lugar ao método de regresión de compoñentes principais (PCR).

O método PCR emprégase para reconstruír datos meteorolóxicos ausentes. Considerando que o conxunto de datos orixinal está representado pola matriz $\mathbf{X}$, onde cada columna representa unha serie temporal cos rexistros dunha variable meteorolóxica. O modelo de regresión multivariante no conxunto de datos orixinal $\mathbf{X}$ só pode aplicarse se esta matriz ten rango completo por columnas. A colinealidade entre columnas pode ocorrer se hai variables predictoras altamente correlacionadas. Neste caso, o método PCR supera a deficiencia de rango substituíndo as variables predictoras orixinais $\mathbf{X}$ polas suas compoñentes principais (PCs) no modelo de regresión. Unha vez que os compoñentes principais $\mathbf{Z} = \mathbf{XV}$ se obteñen a partir da matriz $\mathbf{X}$, utilízanse algunhas delas (p) no modelo de regresión para estimar $\mathbf{y}$.

Polo tanto, o método PCR consiste en realizar a regresión de $\mathbf{y}$ (a serie temporal predicida) non sobre $\mathbf{X}$ en si, senón sobre a matriz de compoñentes principais $\mathbf{Z} \in \mathbb{R}^{n \times p}$, asumindo que se seleccionaron previamente p PCs. Isto implica resolver por mínimos cadrados o sistema lineal

$$\mathbf{Zc} \approx \mathbf{y},$$

cuxa solución é o vector de parámetros

$$\mathbf{c} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}.$$

O modelo de regresión PCR sobre $\mathbf{Z}$ exprésase como:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\mathbf{c},$$

onde $\tilde{\mathbf{y}} \in \mathbb{R}^{n-m}$ é, como antes, o vector dos valores reconstruídos/predicidos de $\mathbf{y}$ durante o período de reconstrución/preditos, $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{V} \in \mathbb{R}^{(n-m)\times p}$ contén os valores dos p PCs seleccionados ao longo do período de reconstrución/predición, e $\mathbf{c} \in \mathbb{R}^p$ é o vector de parámetros do PCR.

A técnica de mínimos cadrados parciais (PLS) extrae do conxunto de variables predictoras (representado pola matriz $\mathbf{X}$) un conxunto de variables latentes (representadas polas columnas da matriz $\mathbf{T}$) que teñen o mellor poder de predición. Estas novas variables predictoras obtéñense maximizando a covarianza entre os predictores e a variable predita.

As variables latentes que modelan $\mathbf{X}$ e $\mathbf{y}$ e predicen mellor $\mathbf{y}$ resultan das descomposicións:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \text{e} \qquad \mathbf{y} = \mathbf{R}\mathbf{q}^T + \mathbf{f},$$

onde $\mathbf{T} \in \mathbb{R}^{m\times p}$ e $\mathbf{R} \in \mathbb{R}^{m\times p}$ son as matrices con p vectores latentes (tamén coñecidos como puntuacións) extraídos de $\mathbf{X}$ e $\mathbf{y}$, respectivamente; $\mathbf{P} \in \mathbb{R}^{q\times p}$ e $\mathbf{q} \in \mathbb{R}^p$ representan os vectores de carga; a matriz $\mathbf{E} \in \mathbb{R}^{m\times q}$ e o vector $\mathbf{f} \in \mathbb{R}^m$ representan os residuos, cuxas normas se minimizan. Ademais, a matriz de puntuacións $\mathbf{T}$ é ortogonal, é dicir, $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I}$. O modelo de regresión

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{c}^T,$$

onde $\mathbf{c}^T \in \mathbb{R}^p$ denota o vector de regresión, permite estimar $\mathbf{y}$ en función das variables latentes $\mathbf{T}$. Ademais, presentamos a combinación dos métodos baseados en AnEn coas variables latentes (LVs) obtidas con PLS, dando lugar aos métodos PLSAnEn e PLSClustAnEn.

Na Sección 12.2 introducimos os conxuntos de datos meteorolóxicos utilizados para probar e validar os diferentes métodos. Tamén se presenta un estudo de correlación entre variables e estacións. As Seccións 12.3 e 12.4 céntranse na selección de compoñentes principais e variables latentes, respectivamente, presentando diversos criterios para escoller as novas variables predictoras.

Na Sección 12.5, preséntanse e discútense os resultados numéricos das probas realizadas cos distintos métodos de reconstrución. Realizamos un estudo comparativo do rendemento de todos estes métodos nun problema de predición retrospectiva, correspondente á reconstrución de datos ausentes nunha estación meteorolóxica mediante datos procedentes dun conxunto de estacións predictoras con diferentes localizacións xeográficas. Os resultados mostran que as técnicas baseadas en PLS foron lixeiramente máis precisas que as obtidas con PCA, especialmente na reconstrución ou predición de variables meteorolóxicas cunha cantidade significativa de oscilacións, como a velocidade do vento.

A Sección 12.6 proporciona unha análise do rendemento computacional dos mesmos métodos. A pesar da precisión do método PLSAnEn, comprobouse que é moi esixente desde o punto de vista computacional, que se beneficia dunha implementación paralela.

O método PLSClustAnEn, que combina os métodos AnEn co agrupamento previo dos análogos, podería ser unha alternativa ao método PLSAnEn, xa que é moito máis eficiente computacionalmente. Tamén se comproba que os métodos de regresión PCR e PLSR son moi rápidos e permiten reconstrucións moi precisas, particularmente para variables altamente correlacionadas.

Capítulo 13: Número Óptimo de Variables Latentes para PLSR

A regresión de mínimos cadrados parciais (PLSR) é un método eficiente para cubrir as lagoas nas series temporais meteorolóxicas. Permite reducir a dimensión do conxunto de datos predictivos a un número reducido de variables latentes, sen perder información significativa. Debido a isto, a configuración do número de variables latentes a utilizar é un aspecto esencial para o éxito do método PLSR.

A Sección 13.1 revisita os fundamentos da descomposición de mínimos cadrados parciais e do método de regresión PLSR.

A Sección 13.2 describimos os diferentes criterios que se poden aplicar para estimar o número óptimo de variables latentes. Catro destes criterios baséanse no método de validación cruzada: o método Q-cuadrado (Q^2), o erro cuadrático medio de validación cruzada (RMSECV), o criterio de World's R e o F de Osten. Outros dous baséanse nos criterios de información: o criterio de información de Akaike (AIC) e o criterio de información bayesiano (BIC). O último criterio é a importancia da variable na proxección (VIP).

Dedicamos a Sección 13.3 a comparar os diferentes criterios utilizados para determinar o número óptimo de variables latentes empregadas no método PLSR aplicado a un conxunto de datos meteorolóxicos do Centro Nacional de Boias de Datos dos EUA (NDBC). As variables presión atmosférica, temperatura e velocidade do vento reconstrúense ao longo dun ano completo nunha única estación empregando as estacións veciñas. O estudo implicou un total de 41 series temporais diferentes.

Os resultados dos experimentos indican que os criterios baseados na validación cruzada, como o RMSECV ou o F de Osten, son eficientes e proporcionan unha cantidade de variables latentes relativamente próxima ao número óptimo experimental; por esta razón, estes criterios son actualmente os máis utilizados. Non obstante, teñen a desvantaxe de ser computacionalmente esixentes.

Os criterios de información, como o AIC e o BIC, tamén proporcionan resultados interesantes e adoitan utilizarse para a predición. A súa simplicidade fai que os seus custos computacionais sexan desprezables. Non obstante, a súa eficacia depende fundamentalmente do valor do limiar. Actualmente, non existen estudos que proporcionen un método teórico para estimar este valor.

Resumen

La presente tesis, titulada "*Aplicaciones de la Matemática Computacional a la Simulación y el Análisis de Datos: dinámica bidimensional de vórtices y reconstrucción de datos meteorológicos*", es el resultado del trabajo desarrollado por Carlos Balsa durante sus estudios en el Programa de Doctorado en Matemáticas y Aplicaciones por las Universidade de A Coruña; Universidade de Santiago de Compostela; Universidade de Vigo; Universidade de Porto; Universidade de Trás-os-Montes e Alto Douro e Universidade do Minho.

Este trabajo comenzó en octubre de 2022 y se llevó a cabo bajo la supervisión de la profesora María Victoria Otero Espinar, del Departamento de Estadística, Análisis Matemático y Optimización de la Universidad de Santiago de Compostela, y el profesor Sílvio Gama, del Departamento de Matemáticas, Facultad de Ciencias, Universidad de Oporto.

La investigación se realizó en dos dominios principales de la Matemática Computacional aplicada a la simulación y análisis de datos: la dinámica bidimensionales de vórtices y la reconstrucción de datos meteorológicos.

Parte 1 - Dinámica Bidimensional de Vórtices

En el primer ámbito, el problema estudiado es el control del desplazamiento de una partícula pasiva sobre una esfera en la que varios puntos de vórtice inducen el movimiento. Se busca la trayectoria óptima que minimice la energía empleada en el desplazamiento y que respete las restricciones impuestas, como los puntos de llegada y partida y el tiempo disponible para el desplazamiento. La primera parte de la tesis está dedicada a este tema de investigación y se denomina Parte 1 - *Dinámica Bidimensional de Vórtices*. Esta parte

abarca los capítulos del 4 al 9.

La investigación se centra en un método de control óptimo basado en un enfoque directo que constituye una alternativa al Principio del Máximo de Pontryagin [95]. En este enfoque, el problema se discretiza y luego se optimiza. Inicialmente, el desplazamiento de la partícula pasiva se transforma en un problema de control. Después, el tiempo para realizar el desplazamiento se discretiza en un número fijo, n, de subintervalos, donde las variables de control son constantes. El problema discretizado se convierte en un problema de optimización no lineal que se resuelve numéricamente mediante un método de disparo que encuentra las variables de control óptimas. En cada subintervalo, correspondiente a una variable de control constante, la dinámica de vórtices se integra mediante un método de Runge-Kutta de cuarto orden.

Cada capítulo corresponde a un estudio independiente que ya ha sido publicado. Cada uno aborda un aspecto diferente del método de control o una aplicación. Los capítulos 4 y 5 exploran la modelización de la dinámica del sistema utilizando coordenadas esféricas y cartesianas, respectivamente. El capítulo 6 formaliza matemáticamente el problema de control óptimo y muestra que la solución es independiente del tipo de coordenadas utilizadas para modelizar la dinámica del sistema. Los capítulos 7 y 8 presentan aplicaciones conceptuales para el método de control óptimo desarrollado. Finalmente, el capítulo 9 presenta un estudio complementario sobre el efecto de la rotación de la esfera en el método de optimización propuesto.

A continuación, se resume con más detalle el contenido de cada uno de los capítulos.

Capítulo 4: Modelización de la Dinámica con Coordenadas Esféricas

Este primer capítulo presenta un estudio introductorio en el que la dinámica de la partícula pasiva advectada por vórtices puntuales bidimensionales en una esfera se modela mediante coordenadas esféricas.

En la Sección 8.1, esbozamos la deducción de las ecuaciones en coordenadas esféricas que modelizan la dinámica de los vórtices $(R\,, \theta_j(t)\,, \phi_i(t))\,$, para $i = 1, 2, \ldots,\ N$, en la esfera de radio R. La evolución temporal de los N vórtices puntuales, suponiendo conocidas sus posiciones en $t = 0\,$, está gobernada por el siguiente conjunto de ecuaciones diferenciales

ordinarias ($i = 1, 2, \ldots, N$):

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})} \,,$$

$$\sin(\theta_i)\,\dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})} \,,$$

donde γ_{ij} , que representa el ángulo central entre los puntos de vórtice i y j, es tal que

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j) \,,$$

y k_j es la circulación del vórtice j.

Una partícula pasiva es, por definición, un vórtice puntual con circulación igual a cero. Así, la dinámica de un sistema con P partículas pasivas advectadas por N vórtices puntuales está dada por las ecuaciones anteriores junto con las ecuaciones para las partículas pasivas

$$\dot{\theta}_p = -\frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,,$$

con $p = N+1, N+2, \ldots, N+P$, y las correspondientes condiciones iniciales.

La formulación del problema de control para el caso de una única partícula pasiva se realiza en la Sección 4.2. Este problema se forma a partir de las ecuaciones de la partícula pasiva ($P = 1$), añadiendo en sus lados derechos los controles angulares u_θ y u_ϕ , respectivamente, es decir:

$$\dot{\theta}_p = -\frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\theta \,,$$

$$\dot{\phi}_p = \frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\phi \,,$$

El funcional a minimizar es

$$f \equiv \int_0^T \left((u_\theta(t))^2 + (u_\phi(t))^2 \right) \, dt \,.$$

La minimización de f corresponde a la reducción óptima del coste de mover la partícula entre los puntos inicial y final, garantizando que llegue a su destino en el tiempo final prescrito T.

La Sección 4.3 está dedicada a la estrategia numérica, basada en un enfoque directo, para determinar el control que guía la partícula pasiva desde su punto inicial hasta su destino final, minimizando la norma cuadrática L^2 del control.

Los resultados de las pruebas experimentales realizadas con un número de vórtices que varía de uno a tres se presentan en las siguientes secciones. En todos los casos, se obtienen soluciones cuasi-óptimas, ya que todas permiten el movimiento de la partícula entre los dos puntos fijos en la superficie de la esfera.

Capítulo 5: Modelización de la Dinámica con Coordenadas Cartesianas

En este capítulo analizamos la modelización de la dinámica del sistema compuesto por la partícula pasiva y N vórtices puntuales utilizando coordenadas cartesianas.

La posición de un vórtice sobre una esfera está dada por el vector $\mathbf{x}_j$ que apunta desde el centro de la esfera hasta la ubicación del vórtice $\mathbf{x}_j = (x_j, y_j, z_j)$ en la superficie esférica $||\mathbf{x}_j|| = R$. La posición del vórtice está dada por

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_i}{||\mathbf{x}_i - \mathbf{x}_j||^2} \,, \qquad\qquad i = 1, 2, \ldots, N \,,$$

con las correspondentes condiciones iniciales, donde k_j es la circulación de los vórtices j,

y N es el número total de vórtices puntuales en la esfera.

Como una partícula pasiva es, por definición, un vórtice puntual con circulación $k = 0$. la dinámica de un sistema con P partículas pasivas advectadas por N vórtices puntuales está dada por la ecuación anterior junto con las ecuaciones para las partículas pasivas

$$\dot{\mathbf{x}}_p = \frac{1}{2\pi R} \sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_p}{||\mathbf{x}_p - \mathbf{x}_j||^2} \ , \qquad p = N+1, N+2, \ldots, N+P \, ,$$

con las correspondentes condiciones iniciales. Estas ecuaciones se utilizan en la Sección 5.1, que está dedicada al problema de control subyacente en la determinación de la trayectoria óptima de la partícula pasiva. También se demuestra que, para que el movimiento de la partícula ocurra en la esfera, el vector de control $\mathbf{U}$ aplicado sobre la partícula debe ser ortogonal a su vector de posición $\mathbf{x}$.

La transformación del problema de control en un problema de optimización no lineal (NOP) en el caso de un único vórtice ($N = 1$) se detalla en la Sección 5.2. La discretización de la función objetivo mediante la regla del rectángulo lleva a la siguiente aproximación

$$\int_0^T \|\mathbf{U}\|^2 \, dt \approx \Delta t \sum_{j=0}^{n-1} \|\mathbf{u}_j\|^2 \ ,$$

donde cada $\mathbf{u}_i$ $(i = 0, 1, 2, \cdots, n-1)$ corresponde al vector de control ejercido en el subintervalo $[t_{i-1} \ t_i)$. Todos estos subintervalos tienen amplitudes iguales a $\Delta t = T/n = (t_n - t_0)/n$, donde T es el tiempo disponible para realizar el desplazamiento entre los dos puntos fijos.

En la Sección 5.3, se trata el problema de un flujo inducido por múltiples vórtices ($N = 2, 3, 4$). En todos los casos, el NOP resultante del enfoque directo se resuelve numéricamente utilizando la función `fmincon` de Matlab, que incluye varios métodos de optimización. Se presenta la comparación de dos métodos: *Interior Point* y *Active Set*. Ambos fueron capaces de encontrar trayectorias óptimas independientemente del nivel de discretización temporal del problema y del número de vórtices involucrados. Los tiempos de cálculo requeridos por ambos métodos son del mismo orden de magnitud. Sin embargo, el método *Active Set* requiere más tiempo cuando el número de parámetros de control a determinar es mayor.

También se analiza el efecto de mejorar el número de controles constantes, n, de $N = 1, \cdots, 4$. Los resultados muestran que aumentar el número de parámetros de control permite una mayor libertad en la elección de la trayectoria, lo que minimiza la energía utilizada para el desplazamiento. Por otro lado, el tiempo de cálculo es mayor.

Para todos los problemas de control analizados, es posible demostrar la existencia de trayectorias cercanas o cuasi-óptimas para la partícula pasiva.

Capítulo 6: Exploración y Comparación de Controles

La posibilidad de modelar el problema mediante dos sistemas de coordenadas diferentes plantea la cuestión de si la solución depende del sistema de coordenadas elegido o si, por el contrario, es única independientemente del sistema. El estudio presentado en este capítulo se centra en la respuesta a esta pregunta. La formulación del problema se deduce en coordenadas esféricas y cartesianas. Luego, se demuestra analíticamente cómo es posible convertir la formulación del problema de un sistema de coordenadas a otro.

En la Sección 6.1, se realiza la derivación de las ecuaciones que gobiernan la dinámica de la partícula pasiva en coordenadas cartesianas. Para el caso de un único vórtice ubicado en el polo norte de la esfera, la dinámica de la partícula pasiva controlada está gobernada por el siguiente sistema de ecuaciones:

$$\begin{cases} \dot{x} = -\frac{k}{2\pi}\frac{y}{x^2+y^2+(R-z)^2} + \alpha\, y \\ \dot{y} = \frac{k}{2\pi}\frac{x}{x^2+y^2+(R-z)^2} - \alpha\, x + \beta\, z \\ \dot{z} = -\beta\, y \end{cases}$$

con las correspondentes condiciones iniciales. En esta ecuación, $\alpha(\cdot)$ y $\beta(\cdot)$ son los parámetros de control.

En la Sección 6.2 se explora la modelización de la dinámica en coordenadas esféricas. Para el caso de un único vórtice ubicado en el polo norte de la esfera, el sistema de ecuaciones responsable de la dinámica de la partícula pasiva se deduce en función de los

controles cartésianos $\alpha(\cdot)$ y $\beta(\cdot)$:

$$\begin{cases} \dot{\theta} = \beta \sin \phi \\ \\ \dot{\phi} = \frac{k}{4\pi R^2} \frac{1}{1-\cos\theta} - \alpha + \beta \cos\phi \cot\theta \end{cases}$$

El uso de estos dos sistemas de ecuaciones permite comparar directamente las soluciones obtenidas con los dos tipos de coordenadas. Del mismo modo, se obtienen las ecuaciones que rigen el sistema para el caso de $N = 1, 2, 3$ vórtices.

La formulación de los problemas de control se detalla en la Sección 6.3. Se derivan tres problemas de control óptimo. El primero utiliza coordenadas cartesianas con controles cartesianos, el segundo utiliza coordenadas esféricas con controles angulares y el tercero emplea coordenadas esféricas con controles cartesianos. Todos ellos se transforman en problemas de optimización no lineal mediante el mismo enfoque directo.

Los experimentos computacionales obtenidos a partir de la solución numérica de los tres problemas de optimización para el caso de $N = 1, 2, 3$ vórtices se describen en la Sección 6.4. Los resultados sugieren que cada formulación produce una solución factible. Cada formulación permite que la partícula pasiva logre el desplazamiento deseado dentro del tiempo asignado. Estas tres versiones del mismo problema confirman la efectividad del algoritmo utilizado para convertir el problema de control óptimo en un problema de optimización. La resolución numérica de estas formulaciones mediante un enfoque directo produce de manera consistente soluciones cercanas o cuasi-óptimas, independientemente del sistema de coordenadas elegido.

Capítulo 7: Aplicación a Tecnología Autónoma Conceptual

Este capítulo se centra en un modelo matemático conceptual que describe el comportamiento de la tecnología autónoma (ATs), utilizada, por ejemplo, en la recolección de basura realizada por vehículos autónomos (AVs) en entornos oceánicos. Este modelo matemático simplificado combina conocimientos conceptuales sobre el mundo físico con estructuras matemáticas simbólicas, lo que lo convierte en un valioso instrumento pedagógico para involucrar a los estudiantes y enfatizar la importancia de las matemáticas y la física en la resolución de problemas ambientales. La circulación atmosférica o oceánica se modela mediante vórtices puntuales en una esfera, y los ATs se representan mediante el control

óptimo de la trayectoria de partículas pasivas.

En la Sección 7.1 se presenta el modelo matemático simplificado de los ATs, basado en coordenadas cartesianas. En el caso de la dinámica de una partícula pasiva (AV) en un flujo esférico con movimiento inducido por dos vórtices, $N = 2$, inicialmente localizados en $\mathbf{x}_1(0) = V_{01}$ y $\mathbf{x}_2(0) = V_{02}$, con circulaciones k_1 y k_2, la posición de los dos vórtices ($\mathbf{x}_1$ y $\mathbf{x}_2$) está gobernada por

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{1}{2\,\pi\,R}\; k_2\; \frac{\mathbf{x}_2 \times \mathbf{x}_1}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{1}{2\,\pi\,R}\; k_1\; \frac{\mathbf{x}_1 \times \mathbf{x}_2}{\|\mathbf{x}_2 - \mathbf{x}_1\|^2} \end{cases}$$

con las respectivas condiciones iniciales $\mathbf{x}_1(0)$ y $\mathbf{x}_2(0)$. El AV controlado, inicialmente en $\mathbf{x}(0) = P_0$, está gobernado por la ecuación

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_1\|^2} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_2\|^2}\right) + \alpha_i\;(y, -x, 0) + \beta_i\;(0, z, -y)\,,$$

para $i = 0, 1, \ldots, n-1$, donde α_i y β_i son los controladores.

En la Sección 7.2, se analizam simulaciones computacionales utilizando el modelo, centrándose específicamente en los casos de corrientes oceánicas inducidas por uno y dos vórtices puntuales. Las trayectorias realizadas por los 30 AVs para transportar la basura recolectada se calculan mediante el enfoque directo propuesto. Los resultados indican la existencia de múltiples trayectorias para el transporte de partículas de basura. Sin embargo, los costos energéticos son menores para las trayectorias más directas entre los puntos de inicio y destino. El número de controles corresponde a una mayor autonomía de los vehículos y puede ayudar a definir trayectorias que minimicen los costos energéticos.

Capítulo 8: Aplicación al Guiado de Residuos Oceánicos hacia una Región Objetivo

Este capítulo presenta una segunda aplicación del sistema de vórtices puntuales y partículas pasivas para modelizar la recolección de basura en el océano mediante vehículos autónomos. Pero esta vez, la modelización de la dinámica de los AVs y las corrientes oceánicas se realiza utilizando coordenadas esféricas.

Al igual que en el capítulo anterior, se asume un escenario idealizado donde la recogida

de basura es realizada por un conjunto de vehículos autónomos inicialmente ubicados en las zonas con mayor densidad de residuos. Se supone que los vehículos autónomos son relativamente pequeños, de modo que pueden tratarse como partículas pasivas, lo que significa que no alteran la estructura local del campo de velocidad del flujo.

Una vez que los compartimentos de estos vehículos autónomos están llenos, deben trasladarse a una zona específica para depositar el material recolectado. El movimiento de estos vehículos utiliza la corriente oceánica como fuente principal de desplazamiento. Dispone de un sistema de controles constantes por tramos, basado en un pequeño número de puntos de cambio predefinidos, que minimizan la suma de los cuadrados de las amplitudes (funcional objetivo). Aquí, el movimiento de los vórtices puntuales en una esfera genera la corriente oceánica.

En la Sección 8.1 se introducen las ecuaciones que modelizan la dinámica de los AVs y los vórtices puntuales en una esfera. En la Sección 8.2 se presenta el método numérico utilizado para resolver el problema de control óptimo involucrado en la determinación de las trayectorias de los AVs.

En el problema de dos vórtices ($N = 2$), la dinámica de las posiciones de los vórtices $V_1(\tau) = (\theta_1, \phi_1)$ y $V_2(\tau) = (\theta_2, \phi_2)$ está dada por

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi} \frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \\ \dot{\phi}_1 = -\frac{k_2}{4\pi} \frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{\sin(\theta_1)(1-\cos(\gamma_{12}))} \\ \\ \dot{\theta}_2 = -\frac{k_1}{4\pi} \frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \\ \dot{\phi}_2 = -\frac{k_1}{4\pi} \frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{\sin(\theta_2)(1-\cos(\gamma_{21}))} \end{cases}$$

con las condiciones iniciales dadas $V_1(0) = (\theta_{10},\, \phi_{10})$ y $V_2(0) = (\theta_{20},\, \phi_{20})$. Los valores de $\cos(\gamma_{12})$ y $\cos(\gamma_{21})$ se calculan de acuerdo con

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,.$$

La interacción entre los dos vórtices y el control influye significativamente en la dinámica

de la partícula pasiva, como se muestra en

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi}\left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\theta \\ \\ \dot{\phi}_p = \frac{1}{4\pi}\left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\phi \end{cases}$$

con las condiciones iniciales $P_0 = (\theta_{p0}, , \phi_{p0})$. El control aplicado a la partícula pasiva está representado por el vector $\mathbf{U} = (u_\theta \ u_\phi)$.

La Sección 8.3 analiza la aplicación del modelo matemático al problema de eliminación de residuos oceánicos. El análisis incluye casos en los que las corrientes oceánicas son inducidas por uno o dos vórtices. En ambos casos, los resultados indican la existencia de múltiples trayectorias para transportar partículas de basura. Sin embargo, los costos energéticos son menores en las trayectorias más directas entre el punto de inicio y el destino.

Capítulo 9: Movimiento Controlado de una Partícula en una Esfera en Rotación

Este capítulo presenta un estudio complementario a los anteriores que explora el movimiento complejo de una partícula pasiva dentro de un flujo inducido por dos vórtices puntuales situados en la superficie de una esfera en rotación.

Se analiza el problema del control del desplazamiento de una partícula pasiva en una superficie esférica en rotación. Para ello, se considera que el movimiento es inducido por dos puntos de vórtice. El problema de control óptimo correspondiente al desplazamiento de una partícula inmersa en un flujo inducido por dos puntos de vórtice en la superficie de una esfera en rotación se formula en coordenadas cartesianas.

En la Sección 9.1 se introducen las ecuaciones que gobiernan el desplazamiento de los vórtices y la partícula pasiva, junto con la formulación del problema de optimización del control.

Se consideran dos vórtices puntuales, denotados por $\mathbf{x}_i \equiv \mathbf{x}_i(t) = (x_i(t), y_i(t), z_i(t))$, $i = 1, 2$, inicialmente ubicados en $\mathbf{x}_1(0)$ y $\mathbf{x}_2(0)$. Se asume que estos vórtices se encuentran en la superficie de una esfera que gira alrededor del eje z. La dinámica de estos dos vórtices puntuales está dada por las ecuaciones diferenciales:

$$\begin{cases} \dot{\mathbf{x}}_1 = \dfrac{k_2}{4\,\pi\,R}\,\dfrac{\mathbf{x}_2 \times \mathbf{x}_1}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}_1} + \Omega\,(-y_1, x_1, 0) \\ \\ \dot{\mathbf{x}}_2 = \dfrac{k_1}{4\,\pi\,R}\,\dfrac{\mathbf{x}_1 \times \mathbf{x}_2}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}_2} + \Omega\,(-y_2, x_2, 0) \end{cases}$$

donde R es el radio de la esfera, k_i es la circulación de los dos vórtices i $(i = 1, 2)$, y Ω es la velocidad angular de rotación de la esfera. La posición de la partícula pasiva controlada $\mathbf{x} \equiv (x, y, z)$ está entonces gobernada por

$$\dot{\mathbf{x}} = \frac{1}{4\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}}\right) + \Omega\,(-y, x, 0) + \mathbf{U}^c\,,$$

con la condición inicial $\mathbf{x}(0)$. Aquí, $\mathbf{U}^c \equiv (u_x, u_y, u_z)$ es el control que actúa sobre la partícula pasiva y mantiene su movimiento en la superficie esférica.

La Sección 9.2 describe la metodología numérica basada en el enfoque directo para abordar el problema de control óptimo. Los resultados numéricos, dependientes del número de controles constantes y la velocidad de rotación de la esfera, se analizan en la Sección 9.3. Se observa que cambiar la velocidad de rotación de la esfera altera la naturaleza de la solución, permitiendo desplazamientos con menores requisitos energéticos o, por el contrario, mayores demandas de energía. Generalmente, cuando la velocidad de rotación es relativamente alta, la partícula debe completar múltiples giros alrededor de la esfera antes de alcanzar su destino.

Parte 2 - Reconstrucción de Datos Meteorológicos

En el ámbito de la investigación en análisis de datos, el presente estudio se centra en el análisis de series temporales de estaciones meteorológicas con el objectivo de rellenar lagunas significativas provocadas pela falta de registros utilizando el método del Ensamble de Análogos (AnEn) [85]. En este contexto, se intentará desarrollar una metodología que aproveche al máximo la información disponible en las estaciones predictoras, garantizando corrección y eficiencia computacional. La segunda parte de la tesis está dedicada a este tema de investigación y se denomina Parte 2 - *Reconstrucción de Datos Meteorológicos*. Esta parte abarca los capítulos del 10 al 13.

El trabajo desarrollado en la Parte 2 de esta tesis se centra en el uso del método AnEn con múltiples series temporales predictoras. Por un lado, una mayor cantidad de series predictoras aporta más información y, por lo tanto, puede mejorar la calidad de la reconstrucción (o predicción). Por otro lado, el método AnEn no funciona bien cuando el número de series temporales predictoras es alto. Por ello, es muy importante contar con técnicas que permitan que el método AnEn sea compatible con múltiples series predictoras, que pueden provenir de diferentes variables o ubicaciones.

Cada capítulo de la Parte 2 también corresponde a un estudio independiente que ya ha sido publicado. Cada uno aborda un aspecto diferente en el desarrollo de métodos basados en AnEn mediante técnicas de reducción de dimensión. El capítulo 10 es un estudio preliminar que explora la aplicación del Análisis de Componentes Principales (PCA) para reducir la dimensión del conjunto de datos predictivo. En el capítulo 11, la técnica PCA se combina con el método AnEn clásico y una variante basada en agrupamiento K-means. El capítulo 12 presenta diferentes enfoques, basados en PCA y Mínimos Cuadrados Parciales (PLS), que reducen la dimensión del conjunto de datos predictivo sin perder su riqueza y diversidad de información. El capítulo 13 es un estudio complementario sobre la comparación de los criterios más utilizados en la elección de variables latentes resultantes del método PLS.

A continuación, se resume con más detalle el contenido de cada uno de los capítulos.

Capítulo 10: Ensambles Análogos de Componentes Principales

Este capítulo presenta un estudio exploratorio donde el Análisis de Componentes Principales (PCA) se combina con los métodos de Ensambles Análogos (AnEn) en una nueva técnica llamada PCAnEn. Este método permite la reconstrucción de datos meteorológicos a partir de otras series temporales, denominadas predictoras. El interés de esta combinación es posibilitar el uso del método AnEn con un gran número de series predictoras, evitando la pérdida de información contenida en ellas.

Primero, se aplica PCA a los datos de entrada para reducir la dimensión de los conjuntos de datos. Luego, se utiliza AnEn para reconstruir los datos de una estación a partir de los datos de estaciones vecinas. Esto permite el uso de más estaciones/información sin comprometer el tiempo de procesamiento.

La Sección 10.1 está dedicada a introducir el método AnEn y sus diferentes variantes. Se presentan las distintas métricas utilizadas para determinar los análogos y las formas (dependiente o independiente) en que se pueden utilizar las series predictoras. En estudios anteriores se encontró que la forma dependiente conduce a reconstrucciones de datos más precisas. También se presenta una variante del método AnEn que agrupa previamente los análogos utilizando el método K-means, lo que reduce significativamente los tiempos de cálculo requeridos para determinar los análogos.

La Sección 10.2 está dedicada a caracterizar el conjunto de datos utilizado e introducir la aplicación de la técnica PCA para reducir la dimensión del conjunto de datos predictivo a ser utilizado por el método AnEn.

La técnica PCA identifica las dimensiones a lo largo de las cuales los datos están más dispersos. De esta manera, se pueden identificar las dimensiones que mejor diferencian el conjunto de datos analizado, es decir, sus componentes principales. Esto se logra mediante la descomposición en valores singulares de la matriz de datos escalada $\mathbf{H} \in \mathbb{R}^{m \times n}$, dada por

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

donde $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ y $\mathbf{V} \in \mathbb{R}^{n \times n}$. La matriz diagonal $\mathbf{\Sigma}$ contiene los valores singulares σ_i de $\mathbf{H}$ para $i = 1, \ldots, n$, donde $\sigma_1 > \sigma_2 > \ldots > \sigma_n$. Los vectores singulares derechos v_i representan las *direcciones de las componentes principales* de $\mathbf{H}$.

El vector

$$\mathbf{z}_1 = \mathbf{H}\mathbf{v}_1$$

tiene la mayor varianza muestral, dada por σ_1^2/m, entre todas las combinaciones lineales normalizadas de las columnas de $\mathbf{H}$. El vector $\mathbf{z}_1$ representa la primera nueva variable y se denomina el primer componente principal (PC_1). La segunda componente principal (PC_2) es $\mathbf{z}_2 = \mathbf{H}\mathbf{v}_2$, ya que $\mathbf{v}_2$ corresponde a la segunda mayor varianza (σ_2^2/m), y las componentes principales restantes se definen de manera similar. Las nuevas variables son combinaciones lineales de las columnas de $\mathbf{H}$, es decir, combinaciones lineales de las variables normalizadas $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$,

$$\mathbf{z}_i = v_{1i}\mathbf{h}_1 + v_{2i}\mathbf{h}_2 + \ldots + v_{ni}\mathbf{h}_n \,, \quad \text{para} \quad i = 1, 2, \ldots, n \,,$$

donde los coeficientes v_{ji} , $j = 1, 2, \ldots, n$, (denominados *cargas*) son los elementos del

vector $\mathbf{v}_i$. La magnitud de un coeficiente está relacionada con la importancia relativa de la variable correspondiente en la componente principal.

El criterio de sustitución de las variables originales por algunas nuevas debe considerar la influencia de estas nuevas variables en la varianza de los datos originales. Esta influencia es directamente proporcional a la magnitud de los valores singulares correspondientes. Se espera que los primeros componentes principales (PCs), correspondientes a los valores singulares más significativos, representen una gran proporción de la varianza total, por lo que serán suficientes para futuros análisis.

La Sección 10.3 está dedicada a experimentos computacionales con el nuevo método PCAnEn para la reconstrucción de variables meteorológicas de una estación dada. Las pruebas se desarrollaron utilizando los programas Matlab y R. También se introducen algunas métricas de error para evaluar la precisión de las reconstrucciones.

Los resultados muestran que el método PCAnEn ofrece una mejor precisión de reconstrucción que el método AnEn clásico. La reconstrucción de datos de una sola estación mediante la estación más cercana parece ser óptima. Sin embargo, la elección de estaciones predictoras debe considerar la proximidad y correlación entre ellas.

En términos de rendimiento computacional, el método PCAnEn permite reducir significativamente el tiempo de procesamiento en comparación con el método AnEn clásico. También se verificó que la implementación en MATLAB es más rápida que la implementación en R.

Capítulo 11: Método AnEn Basado en Clústeres Combinado con PCA

El presente capítulo explora la combinación del método PCAnEn con el método de agrupamiento K-means, lo que da lugar a la nueva variante PCClustAnEn. Así, los datos de múltiples series temporales predictoras se redujeron a una o dos componentes principales (PCs), que luego se utilizan (en lugar de los registros originales de las variables) en el método AnEn original o su variante ClustAnEn para reconstruir los datos faltantes. Las técnicas PCAnEn y PCClustAnEn también se compararon en términos de precisión numérica y eficiencia computacional.

En la Sección 11.1 se describe el conjunto de datos utilizado y señala las correlaciones entre variables meteorológicas y estaciones. En la Sección 11.2 se introduce la técnica PCA, como se explicó en el capítulo anterior, y se propone un criterio basado en la desviación estándar para seleccionar las componentes principales (PCs). Las PCs con valores de desviación estándar superiores a 1 tienen más varianza y, en consecuencia, más información que las variables normalizadas originales, cuya desviación estándar es igual a 1. Estas PCs se seleccionan como nuevas variables predictoras en lugar del conjunto de datos original.

En la Sección 11.3 se describe la combinación de los métodos AnEn y K-means en la base del método ClustAnEn. Cuando se aplica a las PCs elegidos, se designa como el método PCClustAnEn. La Sección 11.4 aplica este nuevo método a la reconstrucción de variables meteorológicas. Se probaron los métodos PCAnEn y PCClustAnEn en la reconstrucción de cuatro variables meteorológicas en una estación dada (velocidad del viento: $WSPD$, velocidad de ráfaga: GST, temperatura atmosférica: $ATMP$ y presión: $PRES$), cada 6 minutos, desde las 10 de la mañana hasta las 6 de la tarde, durante todo el año 2019 (período de predicción). Las estaciones restantes (ubicadas en un radio de 30 km alrededor de la estación prevista) se utilizaron como estaciones predictoras, considerando el período de entrenamiento de 2011 a 2018. Las pruebas computacionales se desarrollaron simultáneamente en Matlab y R.

Los resultados de los experimentos muestran que la técnica PCA mejoró la precisión de la predicción sin comprometer el rendimiento computacional, ya que es posible aumentar el número de estaciones predictoras sin incrementar la cantidad de series temporales de entrada. También se demuestra que la eficacia de PCA está fuertemente influenciada por la correlación entre las series temporales de múltiples predictores, ya que una mayor correlación permite que una gran proporción de la información (varianza) esté contenida en los primeros componentes.

Además, las dos implementaciones de los métodos estudiados en MATLAB y R permiten verificar los resultados numéricos. La escalabilidad de ambos códigos también se estudió en un sistema multicore de escala media. La evaluación de rendimiento mostró la superioridad del método AnEn cuando PCA se combina con el agrupamiento.

Capítulo 12: Método AnEn con Reducción de Dimensión del Conjunto de Datos Predictivos

El estudio presentado en este capítulo contribuye a enriquecer el método AnEn con técnicas que aprovechan un gran número de variables predictoras mediante la reducción de dimensión. Se incluye la reducción en el conjunto de datos predictivos original a un número reducido de nuevas variables predictoras, sin pérdida de información esencial. Este enfoque mejora la calidad de las reconstrucciones, así como su eficiencia computacional. Se explora la reducción de dimensión a través de dos métodos alternativos: el análisis de componentes principales (PCA) y mínimos cuadrados parciales (PLS).

En la Sección 12.1 se presentan los diversos métodos de reconstrucción empleados en este estudio. Como hemos visto en capítulos anteriores, la técnica PCA identifica las dimensiones en las que los datos presentan mayor dispersión, es decir, aquellas con mayor varianza. De esta manera, se pueden identificar las dimensiones que mejor diferencian el conjunto de datos analizado, es decir, sus componentes principales, que a su vez se utilizan como las nuevas variables predictoras. Se presenta la combinación de los métodos basados en AnEn con PCA para aprovechar las potencialidades del método AnEn y la reducción de dimensión proporcionada por la técnica PCA. Además, PCA se combina con la regresión lineal multivariada, dando lugar al método de regresión de componentes principales (PCR).

El método PCR se utiliza para reconstruir datos meteorológicos faltantes. Considerando que el conjunto de datos original está representado por la matriz $\mathbf{X}$, donde cada columna representa una serie temporal con los registros de una variable meteorológica, el modelo de regresión multivariado sobre el conjunto de datos $\mathbf{X}$ puede aplicarse solo si esta matriz tiene rango completo en sus columnas. La colinealidad cercana entre columnas puede ocurrir si hay variables predictoras altamente correlacionadas. En este caso, el método PCR sortea la deficiencia de rango reemplazando las variables predictoras originales $\mathbf{X}$ por sus componentes principales (PCs) en el modelo de regresión. Una vez obtenidos las componentes principales, $\mathbf{Z} = \mathbf{XV}$, a partir de la matriz $\mathbf{X}$, se utilizan algunos de ellos (p) en el modelo de regresión para estimar $\mathbf{y}$.

Por lo tanto, el método PCR consiste en realizar la regresión de $\mathbf{y}$ (la serie temporal predicha) no sobre $\mathbf{X}$ en sí misma, sino sobre la matriz de componentes principales $\mathbf{Z} \in \mathbb{R}^{m \times p}$, asumiendo que p PCs han sido previamente seleccionados. Esto implica resolver

mediante mínimos cuadrados el sistema lineal

$$\mathbf{Zc} \approx \mathbf{y},$$

cuya solución es el vector de parámetros

$$\mathbf{c} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}.$$

El modelo de regresión PCR sobre $\mathbf{Z}$ queda entonces definido por:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\mathbf{c},$$

donde $\tilde{\mathbf{y}} \in \mathbb{R}^{n-m}$ es el vector de valores reconstruidos/predichos de $\mathbf{y}$ durante el periodo de reconstrucción/predicción, $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{V} \in \mathbb{R}^{(n-m)\times p}$ contiene los valores de los p PCs seleccionados a lo largo del periodo de reconstrucción/predicción y $\mathbf{c} \in \mathbb{R}^p$ es el vector de parámetros del modelo PCR.

La técnica de mínimos cuadrados parciales (PLS) extrae del conjunto de variables predictoras (representado por la matriz $\mathbf{X}$) un conjunto de variables latentes (representado por las columnas de la matriz $\mathbf{T}$), que poseen el mejor poder predictivo. Estas nuevas variables predictoras se obtienen maximizando la covarianza entre los predictores y la variable predicha.

Las variables latentes que modelan $\mathbf{X}$ y $\mathbf{y}$ y que mejor predicen $\mathbf{y}$ resultan de las siguientes descomposiciones:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \qquad \text{y} \qquad \mathbf{y} = \mathbf{Rq}^T + \mathbf{f},$$

donde $\mathbf{T} \in \mathbb{R}^{m\times p}$ y $\mathbf{R} \in \mathbb{R}^{m\times p}$ son las matrices con p vectores latentes (también llamados scores) extraídos de $\mathbf{X}$ y $\mathbf{y}$, respectivamente; $\mathbf{P} \in \mathbb{R}^{q\times p}$ y $\mathbf{q} \in \mathbb{R}^p$ representan los vectores de cargas; la matriz $\mathbf{E} \in \mathbb{R}^{m\times q}$ y el vector $\mathbf{f} \in \mathbb{R}^m$ representan los residuos, cuyas normas se minimizan. Además, la matriz de scores $\mathbf{T}$ es ortogonal, es decir, $\mathbf{T}^T\mathbf{T} = \mathbf{TT}^T = \mathbf{I}$. El modelo de regresión

$$\hat{\mathbf{y}} = \mathbf{Tc}^T,$$

donde $\mathbf{c}^T \in \mathbb{R}^p$ denota el vector de regresión, permite estimar $\mathbf{y}$ basado en las variables

latentes **T**. Se presenta la combinación de los métodos basados en AnEn con las variables latentes (LVs) obtenidas con PLS, dando lugar a los métodos PLSAnEn y PLSClustAnEn.

En la Sección 12.2 se introducen los conjuntos de datos meteorológicos utilizados para validar los diferentes métodos, además de presentar un estudio de correlación entre variables y estaciones. Las Secciones 12.3 y 12.4 abordan la selección de componentes principales y variables latentes, respectivamente, mostrando varios criterios para elegir las nuevas variables predictoras.

En la Sección 12.5, se presentan y analizan los resultados numéricos de las pruebas realizadas con los diversos métodos de reconstrucción. Se lleva a cabo un estudio comparativo del rendimiento de estos métodos en un problema de hindcasting, correspondiente a la reconstrucción de datos faltantes en una estación meteorológica mediante información de un conjunto de estaciones predictoras con diferentes ubicaciones geográficas. Se observa que los resultados obtenidos con las técnicas basadas en PLS fueron ligeramente más precisos que los obtenidos con PCA, especialmente en la reconstrucción o predicción de variables meteorológicas con una cantidad significativa de oscilaciones, como la velocidad del viento.

La Sección 12.6 proporciona un análisis del rendimiento computacional de los métodos evaluados. A pesar de la precisión del método PLSAnEn, se verifica que es muy exigente computacionalmente, beneficiándose de una implementación paralela. El método PLSClustAnEn, que combina los métodos AnEn con la agrupación previa de análogos, podría ser una alternativa eficiente al método PLSAnEn. También se comprueba que los métodos de regresión PCR y PLSR son muy rápidos y permiten reconstrucciones muy precisas, especialmente en el caso de variables altamente correlacionadas.

Capítulo 13: Número Óptimo de Variables Latentes para PLSR

La regresión de mínimos cuadrados parciales (PLSR) es un método eficiente para el llenado de vacíos en series temporales meteorológicas. Permite reducir la dimensión del conjunto de datos predictivos a un número reducido de variables latentes, sin pérdida de información significativa. Debido a esto, la elección del número de variables latentes a utilizar es un aspecto esencial para el éxito del método PLSR.

En la Sección 13.1 se revisa los fundamentos de la descomposición de mínimos cuadrados parciales y del método de regresión PLS (PLSR).

En la Sección 13.2 se describen los diferentes criterios que pueden aplicarse para estimar el número óptimo de variables latentes. Cuatro de estos criterios se basan en el método de validación cruzada: el método Q-cuadrado (Q^2), el error cuadrático medio de validación cruzada (RMSECV), el criterio de World's R y el de Osten's F. Otros dos se basan en criterios de información: el criterio de información de Akaike (AIC) y el criterio de información Bayesiano (BIC). El último criterio es la importancia de la variable en la proyección (VIP).

La Sección 13.3 está dedicada a la comparación de los diferentes criterios utilizados para determinar el número óptimo de variables latentes en el método PLSR aplicado a un conjunto de datos meteorológicos del Centro Nacional de Boyas de Datos de USA (NDBC). Las variables presión atmosférica, temperatura y velocidad del viento se reconstruyen durante un año completo en una única estación utilizando las estaciones vecinas. El estudio involucró un total de 41 series temporales diferentes.

Los resultados de los experimentos indican que los criterios basados en la validación cruzada, como el RMSECV o Osten's F, son eficientes y proporcionan un número de variables latentes relativamente cercano al número óptimo experimental; por eso, estos criterios son actualmente los más utilizados. Sin embargo, tienen la desventaja de ser muy exigentes desde el punto de vista computacional.

Los criterios de información, como AIC y BIC, también ofrecen resultados interesantes y se utilizan con frecuencia para la predicción. Su simplicidad hace que sus costos computacionales sean casi despreciables. Sin embargo, su eficacia depende principalmente del valor umbral. Actualmente, ningún estudio ha obtenido una forma teórica de estimar su valor.

Chapter 1

Introduction

The content of this thesis, entitled *Applications of Computational Mathematics to Simulation and Data Analysis: two-dimensional vortex dynamics and meteorological data reconstruction*, is the result of the work developed by Carlos Balsa during the Ph.D. studies in Mathematics and Applications granted by the Universidade de Santiago de Compostela.

This course started in October 2022 and was carried out under the supervision of Professor María Victoria Otero Espinar, from the Department of Statistics, Mathematical Analysis and Optimisation of the University of Santiago de Compostela, and Professor Sílvio Gama, from the Department of Mathematics, Faculty of Science, University of Porto.

The research was carried out in two main domains of applied computational mathematics to simulations and data analysis: simulation of dynamic processes and data analysis.

In the first domain, the problem studied is the displacement of a passive particle over a sphere in which several vortex points induce the motion. The optimal trajectory is sought that minimises the energy spent on the displacement and respects the restrictions imposed, such as the arrival and departure points and the time available for the displacement. The first part of the thesis is devoted to this research topic and is denoted Part 1 - *Two-Dimensional Vortex Dynamics*. This part includes chapters 4 through 9.

Regarding the data analysis research domain, the study focuses on reconstructing time series from meteorological stations, aiming to fill significant gaps derived from missing records. In this context, an attempt will be made to develop a methodology that takes advantage of the maximum possible information available in the predictor stations and

is accurate and computationally efficient. The second part of the thesis is devoted to this research topic and is denoted Part 2 - *Meteorological Data Reconstruction*. This part includes chapters 10 through 13.

Part 1 - Two-Dimensional Vortex Dynamics

Point vortices are finite-dimensional approximations to the two-dimensional vortex dynamics of incompressible ideal fluids (zero viscosity). This research topic, initiated by Helmholtz [57] and continued a few years later by Kelvin [115] and Kirchhoff [68], continues to lead to a great deal of work using theories of dynamical systems, differential geometry, numerical analysis, optimal control, and so on.

A point vortex can be seen as a mathematical model used to describe the dynamics of vortex-dominated flows. These models are based on a low-dimensional flow feature description [117]. The idea is to describe fluids through a quantity that measures the local rotation of the flow (the vorticity) by modelling the objects that generate that rotation. Vortex dynamics, based on point vortex models, have been employed in many science and engineering areas like geophysics, turbulence, superfluids, or hydrodynamic [40, 65, 41, 52, 6]. The solutions of these vortex dynamics models are usually obtained with low computational costs. In contrast, solving partial differential equations in realistic problems requires high computational time and ample memory storage. This makes the point vortex models attractive, especially in flow control problems [97].

Consequently, in most control problems, concerning realistic flows, the solution is achieved using simplified models such as point vortex [117]. There is a special interest in using control methods applied to vortex dynamics in geophysical fluid dynamics, aeronautics and hydrodynamics [97].

In the context of hydrodynamics, the fish-like locomotion is an application of point vortex methods that have received some attention recently. This is primarily due to the development of autonomous underwater vehicles for data collection concerning the multiple oceanic phenomena [93]. In some approaches, the robotic fish locomotion is modelled through a point mass in vector fields defined by the vortex dynamics. In [72], the displacement is achieved by constantly generating periodic and predefined vortices. In [93], the control action is exercised by generating one vortex.

There are two main classes of control methods: direct and indirect. Generally speaking, the direct approach consists of first discretising the problem and then optimising it, while the indirect approach optimises and then discretises. Direct methods discretise the problem relatively to the time to get a nonlinear linear optimization problem (NOP) that can be solved by an optimisation method like Interior Point [123], for instance. These methods are generally handy for singular or constrained trajectory arcs, but their accuracy can be affected by the discretisation [28, 29]. Indirect methods use mainly Pontryagin's maximum principle to derive optimal conditions, being necessary to maximise the Hamiltonian, which can be achieved by collocation or shooting methods [95]. Indirect methods are fast and accurate, but they are also sensitive to the initial guess of the adjoint problem.

Point vortices have been studied on many types of surfaces, such as a plane [11], a sphere [120], or a hyperbolic sphere [60, 87]. In previous work [26, 17], the dynamics of a passive particle in the infinite plane advected by a two-dimensional point vortex inviscid flow was addressed. In [18], we solve the same problem in a viscous fluid.

A passive particle is small enough not to perturb the velocity field, but also large enough not to perform a Brownian motion. Particles of this type are the tracers used for flow visualisation in fluid mechanics experiments [11]. We also consider that the passive particle has the same density as the fluid in which it is embedded. The main goal is to drive a passive particle from an initial starting point to a final terminal point, both given *a priori*, in a given finite time. The flow originated from the displacement of a certain number, say N, of point vortices. The vortex dynamics is governed by N point vortices, and the control is due to the possibility of impulsion in any direction in the two-dimensional plane. Of course, we wanted to minimise the total energy spent on the impulsive needs along the total displacement. This issue can also be seen as part of the general open control problems proposed by Protas [97].

In [26, 17, 18], we solve the problem in the infinite plane by a direct approach. The displacement of the passive particle is transformed into a control problem. The time disposable to perform the displacement is divided into a fixed number, n, of subintervals, where the control variables are constant. The discretised problem is solved numerically through a single shooting method. The fourth-order Runge-Kutta method integrates the vortex dynamics in each subinterval.

This thesis concerns the problem of controlling, by a direct approach, the displacement of a passive particle driven by a certain number of point vortices on the surface of a sphere.

The main goal of this topic of the PhD thesis is to develop a control strategy based on a direct approach, an alternative to the classical Pontryagin's maximum principle method, that enables the drive a passive particle from an initial starting point to a final point, both on sphere surface, and that minimises the total amount of energy in a given finite time.

This thesis begins by following the classic methodology in mathematics research. In general, the research starts by conducting a comprehensive study of the topics of point vortex and reviewing some classical and recent bibliographical references.

The main result consists of an optimal control method that allows controlling the displacement on a sphere of a passive particle inserted in a flow induced by several point vortices.

Part I of this thesis, entitled *Two-Dimensional Vortex Dynamics*, includes chapters 4 through 9. Each chapter is based on a published study.

Chapter 4 is based on the preliminary study [16] that explores the control of the motion of a passive particle advected by $N = 1, 2, 3$ point vortices in a sphere, whose mathematical modelling uses spherical coordinates.

Chapter 5 is based on the study [19] that explores the control of the motion of a passive particle advected by $N = 1, 2, 3, 4$ point vortices in a sphere, whose mathematical modelling uses Cartesian coordinates.

Chapter 6 is based on the study [22]. In this chapter, the modelling of particle and point vortices dynamics, whether in Cartesian or spherical coordinates, gives rise to alternative formulations of the identical problem. Thanks to these two versions of the same problem, we can assert that the algorithm employed to transform the optimal control problem into an optimisation problem is effective. The numerical resolution of these formulations through a direct approach consistently produces optimal solutions, regardless of the selected coordinate system.

Chapter 7 is based on the study [23]. This chapter proposes an application for the method of controlling the trajectory of a passive particle on a sphere. A mathematical toy model is proposed for garbage collection encompassing the coordination of a vast array of autonomous devices, where each device needs only a few trajectory adjustments. In this simplified conceptual model, each autonomous device is represented by a passive particle. At the same time, the environmental currents are depicted by the dynamic behaviour of

point vortices on a spherical surface. The dynamic of the system is modelled through Cartesian coordinates.

Chapter 8 is based on the study [21]. This chapter continues to exploit the potential of the simplified mathematical model presented in Chapter 7, to depict the process of collecting ocean debris. But this time, spherical coordinates are used to model the system's dynamics.

Chapter 9 is based on the complementary study [20] that introduces the sphere's rotation. It explores the control of the motion of a passive particle within a flow induced by two-point vortices ($N = 2$) located on the surface of a rotating sphere. It is verified that the quasi-optimal solution, obtained by the proposed control method, also depends on the rotation speed of the sphere.

Finally, the Chapter 14 presents some final remarks and possibilities for future work.

Part 2 - Meteorological Data Reconstruction

A recurrent problem in data analysis is the presence of gaps (sequences of missing records) in time series, which makes any analysis difficult or even impossible. The problem is usually solved by interpolation if the length of the missing gap is reduced. Still, this technique can no longer be applied if the gap is larger than the time scale that characterises the predictability of the time series. In meteorology, one of the techniques used for fault reconstruction is based on the analog ensemble method (AnEn).

Preliminary weather predictions by correlation with similar states in the past (analogues) were initially established by Lorenz [73], who suggested that two atmospheric states that are initially very close will remain somewhat similar. This was introduced as an alternative to classical weather forecasting based on systems of equations underlying deterministic Numerical Weather Prediction (NWP) models. However, Lorenz's proposal was discarded for many years because of limited historical data on past weather conditions (especially over vast geographical areas) and insufficient computing capacity to implement his approach. Two decades later, Van den Dool [51] revisited analogue-based short-range weather forecasting and found it feasible and effective when applied to limited geographical areas.

Monache [86] showed the applicability of an analogue scheme (named AN) for post-

processing numerical weather forecasts to reduce systematic and random errors. The basic idea is that if previous forecasts (analogues) exist similar to the current NWP forecast (predictor), it is possible to produce an AN forecast using the observations corresponding to these previous forecasts. The analogue prediction is then compared with the NWP prediction to infer the prediction error and thus improve the NWP forecast.

Later, Monache refined the use of analogues to estimate the probability distribution of the future state of the atmosphere [85]. Instead of improving a single deterministic NWP prediction, the goal was to derive a Probability Density Function (PDF) of the range of possible future states (forecasts), a more realistic approach due to imperfect initial conditions and model limitations that generate prediction errors. In the same paper, *analog ensemble* (AnEn) was coined to refer to the observations corresponding to past analogue predictions and the method used to select them.

Since its introduction, the AnEn method has gained traction in many contexts, e.g. renewable energy management, especially wind and solar energy forecasting [5, 4]. It has also been combined with other approaches such as artificial neural networks, e.g. for predicting the power generated by photovoltaic power plants [38].

Studies aiming at the development of AnEn techniques favour applications in *downscaling* and *forecasting* rather than *hindcasting*, yet all topics share affinities. In the area of downscaling, we highlight the study conducted by Rozoff and Alessandrini comparing AnEn with convolutional neural networks (CNN) for reconstructing high-resolution 10 m winds over complex terrain, where AnEn was found to produce lower errors than the CNN [105]. In forecasting, AnEn is mainly used to post-process forecast results. In recent years, post-processing with AnEn has been applied to predict various meteorological variables and energy. For example, the AnEn technique has been used to improve the accuracy of ozone and particulate matter forecasts [111]. Solar forecasting, which involves the prediction of irradiance and solar power forecasting, has also attracted considerable attention in the last decade [128].

Among all post-processing forecast applications, there is a consensus that AnEn benefits from increasingly large training datasets. However, the large amount of data that must be processed to determine the analogues sometimes makes the computational cost prohibitive [121]. Therefore, there is a great interest in improving the computational efficiency of the AnEn method. One of the significant outcomes of this interest has been the development of the Parallel Analog Ensemble (PAnEn) library [59]. This library, which

provides an efficient parallel implementation of the AnEn method and user-friendly interfaces in R and C++, has been successfully applied to a huge dataset related to photovoltaic power generation [58].

The AnEn method also plays a vital role in weather hindcasting. Classically, weather hindcasting involves applying a forecast model to a past starting point to validate the model by comparing its forecasts with available observations (*reanalysis*). If some observations are unavailable, their time series can be additionally complemented (*reconstruction*) by using the corresponding forecasts as a substitute for the missing observations. However, the combination of hindcasting with the AnEn method also makes it possible to reconstruct meteorological data. For example, a variable at a meteorological station can be reconstructed based on data of correlated variables from the same station and/or other nearby predictor stations.

The last option, known as hindcasting with multistations, was explored in previous work [44]. There, cosine similarity, normalisation, and K-means clustering were used as similarity metrics for analogue selection and as an alternative to the classical Monache metric. The coupling of the AnEn method with K-means clustering proved to be promising. At the same time, the results pointed to the need to consider other clustering methods and conduct a parametric study on the essential parameters of the methods under consideration. This study was initiated in [7], in the context of reconstructing a single meteorological variable. The study restricted the choice of analogues to the classical Monache metric, the K-means and the C-means clustering methods and confirmed that K-means provides the best accuracy. Heuristics were also identified to determine the number of clusters, the number of analogues, and the analogue window size, which minimised the prediction errors.

The combination of K-means clustering with the AnEn method was further investigated in [26]. The mathematical formulation of the resulting approach was introduced to emphasise the essential options and parameters of the new process. A parametric study was conducted regarding hindcasting meteorological variables from the same dataset used in [7]. The computational performance of the K-means-based and classical AnEn methods was compared for a limited parallel execution scenario.

More recently, the computational efficiency of the AnEn method was improved, not with the focus on the parallelisation of the original algorithm, but by reducing the number of operations required to compute the analogues, which is the most demanding task [24, 25].

The various computational loops needed to determine the analog ensembles were replaced by a single step that employs clustering through the K-means method. This approach was developed in the field of hindcasting. Since this field has many similarities with downscaling and forecasting, the proposed cluster-based variant of the AnEn method could be easily adapted to these fields. Furthermore, the algorithm of the new approach was presented in detail. The effects of its most essential parameters on the reliability and computational efficiency of the method were examined.

The work developed in this thesis focuses on using the AnEn method with many predictor time series. On the one hand, many predictor series contribute more information and thus can improve the reconstruction quality (or prediction). On the other hand, the AnEn method does not work well when the number of predictor time series is high. It is, therefore, essential to have techniques that allow the AnEn method to be compatible with many predictor time series, which may come from different variables or locations.

The main goal of this part of the PhD thesis is to develop a mathematical methodology that enables the application of the AnEn method with a large number of predictor time series.

This thesis topic begins by conducting a comprehensive study of dimensionality reduction and reviewing some classical and recent bibliographical references.

The main result of this thesis is the development of missing data reconstruction methods that allow for solving hindcasting and forecasting problems with many predictors. These developed methods combine the robustness of the AnEn method (with or without clustering) with the dimension reduction methods such as Principal Components Analysis (PCA) and Partial Least Squares (PLS).

Part II of this thesis, entitled *Meteorological Data Reconstruction*, includes chapters 10 through 13. Each chapter is based on a published study.

Chapter 10 is based on the preliminary study [12] that explores the application of Principal Components Analysis (PCA) to reduce the dimension of the predictor dataset used by the AnEn method. The proposed combination is greatly influenced by the choice of the predictor time series according to its correlation with the predicted one. In this condition, PCA associated with AnEn decreased the errors in the prediction/reconstruction of meteorological data and improved the computational time.

Chapter 11 is based on the study [35]. In this study, the PCA technique is combined with the classical AnEn method and a K-means cluster-based variant, within the context of reconstructing missing meteorological data at a particular station using information from neighbouring stations. This combination reduces the dimension of the number of predictor time series, while ensuring better accuracy and higher computational performance than the AnEn method.

Chapter 12 is based on the study [14]. This study presents different methods that reduce the dimension of the predictor dataset without losing the richness and diversity of information contained in the original data with large dimensions. This is accomplished with PCA or Partial Least Squares (PLS) methods. Combining the AnEn with PCA or PLS techniques results in efficient hybrid methods for reconstructing or forecasting unstable meteorological variables, such as wind speed. This hybrid method is computationally demanding, but its performance can be improved by introducing cluster-based variants of the AnEn method. The multivariate linear regression methods applied to the new variables resulting from the PCA or PLS techniques proved efficient, especially for predicting meteorological variables without local oscillations, such as the pressure.

Chapter 13 is based on the complementary study [15]. It is about comparing eight different criteria, used in the choice of latent variables, resulting from applying the PLS method to reduce the dimension of the predictor data set. The test results indicate that the criteria based on cross-validation are the most efficient, being, however, more computationally demanding.

Finally, the Chapter 14 presents some final remarks and possibilities for future work.

Chapter 2

Objectives

The objectives of this thesis are outlined for each research domain in the two parts of the manuscript.

Part 1 - Two-Dimensional Vortex Dynamics

The research topic that is studied in Part I is the displacement of a passive particle over a sphere in which the motion is induced by several vortex points. The aim is to determine the optimal trajectory that minimises the energy required for the displacement while satisfying various constraints, including prescribed departure and arrival points as well as a fixed travel time. The objectives described in this section are developed in the chapters 4 through 9 of Part I.

Objectives of Part I

The main goal of the Ph.D. thesis is to develop a control strategy based on a direct approach, an alternative to the classical Pontryagin's Maximum Principle method, that enables the drive of a passive particle from an initial starting point to a final point, both on sphere surface, that minimises the total amount of energy in a given finite time. For the realization of this objective, we work on specific initiatives, such as the following:

- Introduce the mathematical models for the dynamics of point vortices and passive particles on the surface of the sphere using
  - Cartesian coordinates,
  - Spherical coordinates.
- Formulate the optimal control problem for the two different types of coordinates.
- Formulate the discretized optimal control problem.
- Numerically solve the discretized optimal control problem.
- Analyse the effect of the type of coordinates in the resulting trajectories.
- Analyse the effect of the discretization parameters in the resulting trajectories.
- Analyse the performance of the method in function of the physical parameters:
  - Number of point vortices,
  - Rotation of the sphere.
- Apply the proposed optimal control method to a specific scenario.

Part 2 - Meteorological Data Reconstruction

The research topic that is studied in Part II focus on the analysis of time-series from meteorological stations, with the aim of filling large gaps, derived from missing records. In this context, the goal is to develop a methodology that leverages the maximum amount of information available from a large set of predictor variables. The objectives outlined in this section are detailed in Chapters 10 through 13 of Part II.

Objectives of Part II

The main goal of the Ph.D. thesis is to develop or apply a mathematical methodology that enables the application of the AnEn method with a large number of predictor time series. With the aim of achieving this objectives, our actions are focused on the following points:

- Exploit the dimensionality reduction through the *Principal Component Analysis* (PCA).
- Exploit the dimensionality reduction through the *Partial Least Squares* (PLS).
- Integrate the new variables resulting from dimension reduction with PCA in AnEn-based methods.
- Integrate the new variables resulting from dimension reduction with PLS in AnEn-based methods.
- Implementation of multivariate regression methods based on the PCA.
- Implementation of multivariate regression methods based on the PLS.
- To indicate the advantages and disadvantages of reconstruction methods applied with dimension reduction of the predictor data set.
- To indicate the most suitable values of the internal parameters for the proposed methods.
- The proposed reconstruction methodology will be applied to datasets from real meteorological records.

Chapter 3

Methodology

The methodology of this thesis is presented separately for each research domain in the two parts of the manuscript.

Part I - Two-Dimensional Vortex Dynamics

Part I explores the dynamics of a passive particle displaced over a sphere, with its motion driven by multiple vortex points. This study seeks to determine the optimal trajectory that minimises energy expenditure under prescribed constraints like departure and arrival locations and the time available for the motion. The methodology discussed in this section aligns with the content of chapters 4 through 9.

Methodology of Part I

Adopting a classic mathematical research methodology, this thesis first performs a comprehensive study on the topic of point vortices. This involves a thorough review of relevant classical and recent bibliographical references. The methodology applied to the development of the proposed objectives is described herein.

- Revue of the main contribution to the study of point vortex in the sphere using Cartesian or spherical coordinates and application to modeling systems made of N

point vortex and P passive particles.

- Apply control theory to identify the different components of the control problem, namely the restrictions due to motion on the sphere surface.
- Apply a direct approach to obtain a nonlinear optimization problem (NOP).
- Development of a computational algorithm for the resolution of the NOP with the Matlab Optimization Toolbox.
- Comparison of the computational results obtained using the two different types of coordinates.
- Development of a parametric study of the number of control variables.
- Extend the solution method to an arbitrary number of point vortex $N = 1, 2, 3, 4$ and resolution with the proposed algorithm.
- Introduction in the modeling equations of the sphere rotation parameter and resolution with the proposed algorithm considering different values of this parameter.
- Apply the proposed optimal control method to conceptual models for ocean waste removal.

Part II - Meteorological Data Reconstruction

Part II addresses the analysis of meteorological time series to mitigate the issue of significant data gaps. In this context, the objective is to develop mathematical methods designed to maximize the utilization of information from a large set of predictor variables. The methodology outlined here is detailed across Chapters 10 through 13.

Methodology of Part II

This part of the thesis presents a comprehensive description of dataset dimensionality reduction, alongside a review of seminal and contemporary bibliographical references concerning the Analog Ensemble method. The specific methodologies employed to develop the proposed objectives are as follows:

- Revue of the main contribution to the study of dimensionality reduction through PCA.
- Revue of the main contribution to the study of dimensionality reduction through PLS.
- Combinations of principal components (PCs) with AnEn and AnEn-based methods such as Cluster AnEn.
- Combinations of latent variables (LVs) with AnEn and AnEn-based methods such as Cluster AnEn.
- Implementation of the principal components regression (PCR).
- Implementation of the partial least squares regression (PLSR).
- Computational implementation of the proposed methodology using R and Matlab software and execution of
  - Methods validation using real meteorological data,
  - Comparative tests of the reconstruction quality of the proposed methods,
  - Computational performance tests of the proposed methods.
- Performing parametric tests to identify the optimal number of PCs or LVs.

Part I

Two-Dimensional Vortex Dynamics

Chapter 4

Modelling the Dynamics with Spherical Coordinates

The results from this chapter have already been published as:
Balsa, C., Gama, S. (2022). *A control problem with passive particles driven by point vortices on the sphere*. In: Guarda, T., Portela, F. (eds) Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022. Communications in Computer and Information Science, vol. 1675. Springer, Cham. ISBN 978-3-031-20318-3. ISSN 1865-0929. https://doi.org/10.1007/978-3-031-20319-0_11.

Point vortices are finite-dimensional approximations of the two-dimensional vortex dynamics of incompressible ideal fluids (zero viscosity) [8, 9, 10].

This research topic, initiated by Helmholtz [57] and continued a few years later by Kelvin [115] and Kirchhoff [68], continues to lead to a great deal of work using theories of dynamical systems, differential geometry, numerical analysis, optimal control, and so on.

Thus, point vortices have been studied on many types of surfaces, such as a plane [11], a sphere [120], or a hyperbolic sphere [60, 87].

The present study concerns the motion of point vortices on a sphere. Point vortices on the sphere are relevant because they represent a simplified approximation to the behaviour of specific geophysical flows for which the curvature of the Earth is essential and persist over

long periods of time [120]. Indeed, many questions related to the fundamental dynamics of atmospheric flows are answered by vortex point models [94]. Conceptual models of point vortices are also used to identify and evaluate physical phenomena affecting the structure and interaction of atmospheric and oceanic vortices [32, 66, 130, 52, 84]. More recently, point vortices are also used to model pesticide dispersion in agricultural problems [131, 67].

This work focuses on optimising the displacement of a passive particle interacting with multiple vortices on the surface of a sphere. More precisely, we are interested in optimally controlling the displacement of the passive particle between two fixed points by minimising the energy spent on the displacement, considering that the time to perform the displacement is fixed. This problem can also be viewed as a simplified version of displacing a glider between two points, using atmospheric circulation to minimize energy expenditure.

To solve this problem, the displacement of the passive particle is converted into a control problem, which is solved using a direct numerical approach. Previous work has used a similar approach to solve a vortex problem in the infinite plane [17, 18]. The time T available to perform the displacement is discretised into n subintervals, where the controls are constant. The resulting nonlinear optimization problem (NOP) is solved numerically using the solver `fmincon` of the MATLAB Optimisation Toolbox [77]. This solver allows constrained NOPs to be solved using different optimisation methods. In previous work, we found that the most suitable methods are the interior point method and the active set method [17, 18].

This first chapter, which is based on [16], presents an introductory study in which the dynamics of the passive particle advected by two-dimensional point vortices in a sphere, S^2, is modelled using spherical coordinates. In Section 4.1, we sketch the deduction of the equations that govern the dynamics of the vortices in the sphere. The formulation of the control problem is done in Section 4.2. The following section is dedicated to the numerical strategy for determining the control that drives the passive particle from its initial point to its final destination, leading to the minimisation of the square of the L^2-norm of the control function. We considered the number of vortices from one to three.

4.1 Point Vortices on a Sphere

The two-dimensional incompressible Euler equation, written in spherical coordinates (R, θ, ϕ) ,

- R is the (constant) radius of the sphere, S^2 ;
- $\theta \in \mathbb{R} \pmod{\pi} = [0, \pi]$ is the colatitude (or polar) angle of the vortex position, i.e. the angle between the radius passing through the North Pole, here located in $(0, 0, R)$ and the radius passing through the vortex, and varies from 0 rad (North Pole) to π rad (South Pole); and
- $\phi \in \mathbb{R} \pmod{2\pi} = [0, 2\pi]$ is the longitude (or azimuthal) angle, i.e. the angle that the meridian passing through $(R, 0, 0)$ makes with the meridian passing by the vortex position, and varies from 0 rad to 2π rad;

reads [66]:

$$\omega \, \mathbf{e}_r = \boldsymbol{\nabla} \times \upsilon \, , \tag{4.1}$$

$$\upsilon = (\boldsymbol{\nabla}\psi) \times \mathbf{e}_r \, , \tag{4.2}$$

where the differential operators act in spherical coordinates, ψ is a stream-function,

$$\upsilon \equiv (0 \, , v_\theta \, , v_\phi) = \left(0 \, , \frac{1}{R^2 \sin(\theta)} \frac{\partial \psi}{\partial \phi} \, , - \frac{\partial \psi}{\partial \theta} \right) \, , \tag{4.3}$$

is the velocity field, and

$$\mathbf{e}_r = (\sin(\theta) \, \cos(\phi), \sin(\theta) \, \sin(\phi) \sin(\theta) \, , \cos(\theta)) \, ,$$

is the unit vector in the radial direction. Inserting (4.2) into (4.1), we obtain

$$\nabla^2 \psi = -\omega \, , \tag{4.4}$$

being

$$\nabla^2 \equiv \frac{1}{R^2 \sin(\theta)} \frac{\partial}{\partial \theta} \left(\sin(\theta) \frac{\partial}{\partial \theta} \right) + \frac{1}{R^2 \sin^2(\theta)} \frac{\partial^2}{\partial^2 \phi} \, ,$$

the so-called Laplace-Beltrami operator on a sphere, S^2 . Replacing in Eq. (4.4) the vorticity, $\omega(\cdot)$, by the Dirac δ function $\delta(\theta - \theta') \, \delta(\phi - \phi')$, the corresponding solution is the Green's function:

$$G(\theta, \phi, \theta', \phi') = - \frac{1}{4\pi R^2} \, \ln(1 - \cos(\gamma)) \, ,$$

where γ is the central angle that the points (R, θ, ϕ) and (R, θ', ϕ') form with the origin of the sphere, and it is such that

$$\cos(\gamma) = \cos(\theta)\cos(\theta') + \sin(\theta)\sin(\theta')\cos(\phi - \phi')\,.$$

Knowing the Green's function of the Laplace-Beltrami operator on the sphere, S^2, the solution of Eq. (4.4) is the convolution $\psi = G * \omega$, i.e.

$$\psi(\theta, \phi) = \frac{1}{4\pi R^2} \iint \omega(\theta', \phi')\, \ln(1 - \cos(\gamma))\, \sin(\theta')\, d\theta'\, d\phi'\,, \tag{4.5}$$

and, thanks to (4.3), we reconstruct the velocity $\mathbf{v} = (0\,, v_\theta\,, v_\phi)$ from the stream-funtion. Assume now that the vorticity is the linear combination of N Dirac δ functions,

$$\omega(t, \theta, \phi) = \sum_{j=1}^{N} k_j\, \delta(\theta - \theta_j(t))\, \delta(\phi - \phi_j(t))\,,$$

where the "singleton" $k_j\, \delta(\theta - \theta_j(t))\, \delta(\phi - \phi_j(t))$ is interpreted as the vorticity with intensity, or circulation, $k_j \in \mathbb{R}$, at the point vortex located at $(R\,, \theta_j(t)\,, \phi_i(t))$, and N is the total number of point vortices corresponding to the total circulation $\Gamma = \sum_{j=1}^{N} k_j$. Their time evolution, assuming known their positions at $t = 0$, is governed by the set of ordinary differential equations ($i = 1, 2, \ldots, N$):

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j\, \frac{\sin(\theta_j)\sin(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})}\,, \tag{4.6}$$

$$\sin(\theta_i)\, \dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j\, \frac{\sin(\theta_i)\, \cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})}\,, \tag{4.7}$$

where γ_{ij}, representing the central angle between the ith and the jth vortex points, is such that

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,, \tag{4.8}$$

or, in Bogomolov notation [32]:

$$\dot{\theta}_i = -\frac{1}{4\pi R^2}\sum_{\substack{j=1\\ j\neq i}}^{N} k_j \,\frac{\alpha_{ij}}{1-\cos(\gamma_{ij})}\,, \quad \sin(\theta_i)\,\dot{\phi}_i = \frac{1}{4\pi R^2}\sum_{\substack{j=1\\ j\neq i}}^{N} k_j \,\frac{\beta_{ij}}{1-\cos(\gamma_{ij})}\,, \tag{4.9}$$

with $\alpha_{ij} = \sin(\theta_j)\sin(\phi_i-\phi_j)$ and $\beta_{ij} = \sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i-\phi_j)$. Moreover, Eqs (4.6)-(4.7) has the following Hamiltonian structure:

$$H = \frac{1}{4\pi R^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N} k_i\,k_j \ln\left(1-\cos(\gamma_{ij})\right)\,,$$

and they can be rewritten as

$$\frac{d\,\cos\left(\theta_i\right)}{dt} = \left\{\cos\left(\theta_i\right), H\right\}\,,$$

$$\frac{d\,\phi_i}{dt} = \left\{\phi, H\right\}\,,$$

in which the Poisson bracket between two functions f and g is defined by

$$\left\{f,g\right\} = \sum_{J=1}^{N}\frac{1}{k_j}\left(\frac{\partial f}{\partial\phi_j}\frac{\partial g}{\partial\cos\left(\theta_j\right)} - \frac{\partial g}{\partial\phi_j}\frac{\partial f}{\partial\cos\left(\theta_j\right)}\right)\,.$$

The scalar moment quantities Q, P and S defined by

$$Q = \textstyle\sum_{J=1}^{N} k_j\,\sin\left(\theta_j\right)\cos\left(\phi_j\right)\,,$$

$$P = \textstyle\sum_{J=1}^{N} k_j\,\sin\left(\theta_j\right)\sin\left(\phi_j\right)\,,$$

$$S = \textstyle\sum_{J=1}^{N} k_j\,\sin\left(\theta_j\right)\,,$$

are such that $\{H,Q\} = \{H,P\} = \{H,S\} = 0$, and, therefore, are invariant quantities (which are particularly useful for, among others, controlling the accuracy of numerical integration).

A passive particle is, by definition, a point vortex with circulation set to zero. Thus, the dynamics of a system with P passive particles advected by N point vortices is given

by Eqs. (4.6)-(4.7) together with the equations for the passive particles

$$\dot{\theta}_p = -\frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,, \tag{4.10}$$

$$\dot{\phi}_p = \frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} \,, \tag{4.11}$$

with $p = N+1, N+2, \ldots, N+P$, and the respective initial conditions.

By a reparametrization of time ($\tau := t/R^2$) in equations (4.6)-(4.7) and (4.10)-(4.11), we can henceforth fix $R = 1$.

4.2 The Control Problem

The control problem that we want to address here is formulated from the Eqs. (4.10)-(4.11), adding to their right-hand sides the angular controls u_θ and u_ϕ, respectively, i.e.

$$\dot{\theta}_p = -\frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\theta \,, \tag{4.12}$$

$$\dot{\phi}_p = \frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\phi \,, \tag{4.13}$$

The functional to be minimised is

$$f \equiv f[u_\theta(\cdot), u_\phi(\cdot)] = \int_0^T \left((u_\theta(t))^2 + (u_\phi(t))^2 \right) \, dt \,. \tag{4.14}$$

The minimisation of (4.14) corresponds to the optimal reduction of the cost of moving the particle between the initial and final points so that the particle reaches the destination at the prescribed final time T as in [17, 75, 74, 18].

As in [17], we shall consider an approximation of this functional by sampling the integrand at regular intervals and estimating based on it. This time, we used the rectangular rule [46], which considers a sequence of n constant segments to approximate the integrand

of the cost function (4.14), $g(t) = (u_\theta(t))^2 + (u_\phi(t))^2$, i.e.

$$f_n = \Delta t \sum_{i=0}^{n-1} g(t_i), \tag{4.15}$$

with $\Delta t = T/n$, and $t_i = i\,\Delta t\,,\ \ i = 0, 1, \cdots, n$.

We consider the admissible angular control strategy in the set

$$\mathcal{U} = \{u = (u_\theta\,, u_\phi) : [0, +\infty[\longrightarrow \mathbf{U}\,;\ u \ \text{ measurable}\}\,,$$

where $\mathbf{U} = [0,1]^2 \subset \mathbb{R}^2$. Rescaling the circulations, we can assume that $\|u\| \leq u_{\max} = 1$.

The state space is $M =]0, \pi] \times [0, 2\pi]$. Introducing some basic notation and following [33], let $x = (\theta_p, \phi_p) \in M$, consider (4.12)-(4.13) simply as $\dot{x} = h(x, u)$, and denote the unique solution of this ODE by $x_u(\cdot, x_0)$, such that $x_u(0, x_0) = x_0$.

With $T > 0$ and $x_0 \in M$, let $\mathcal{U}_{T,\,x_0} \subset \mathcal{U}$ to be the set of controls $u \in \mathcal{U}$, such that the trajectory $x_u(\cdot, x_0)$ is well defined in $[0, T]$. Call $\mathcal{A}(T, x_0) = \{x_u(T, x_0) : u \in \mathcal{U}_{T,\,x_0}\}$ the atteignable set from x_0 in time T, and $\mathcal{A}(x_0) = \cup_{T \geq 0} \mathcal{A}(T, x_0)$ the attainable set from x_0.

The controlled system is called controllable from x_0 if $\mathcal{A}(x_0) = M$, and it is controllable, if $\mathcal{A}(x_0) = M\,, \forall x_0 \in M$.

4.3 Numerical Solution of the Control Problem

In this section, the control problem in the sphere is solved numerically by a direct approach, converting the displacement of the passive particle into a control problem.

The time to perform the displacement is divided into a fixed number n of subintervals, where the control variables are constant. The control function $\mathbf{U}(\cdot)$ is replaced by n control variables $u_0\,, u_1\,, \cdots, u_{n-1}$ and the discretised problem is solved numerically by an optimisation procedure (for details see [17]). The vortex dynamics is integrated in each subinterval using the fourth-order numerical Runge-Kutta scheme [46].

Numerical calculations were performed in Matlab using the nonlinear programming solver `fmincon`, which provides some algorithms for constrained optimisation, such as the

interior point or the active set (see [77]).

In this work, we consider (4.15) with $n = 2$, i.e. $f_2 = \Delta t\,(g(t_0) + g(t_1))$.

4.4 A Particle Advected by a Single Vortex

In the one vortex ($N = 1$) and one particle ($P = 1$) optimal control problem, where the vortex is localised in the North Pole of the sphere, $V(0) = (0,\, 0,\, 1)$, and has circulation k, the dynamics of the passive particle is given by

$$\begin{cases} \dot{\theta}_p = u_\theta\,, \\ \dot{\phi}_p = \frac{k}{4\pi}\,\frac{1}{1-\cos(\theta_p)} + u_\phi\,, \end{cases} \tag{4.16}$$

with the given initial condition $P_0 = (\theta_{p0},\, \phi_{p0})$. The variables u_θ and u_ϕ represent the angular controls applied on the passive particle, i.e., $u = (u_\theta,\, u_\phi)$.

The control (4.16) system can be written in the following form [33]:

$$\dot{x}(t) = F_0(x(t)) + \sum_{i=1}^{2} u_i(t) F_i(x(t))\,, \tag{4.17}$$

where $x = (\theta_p, \phi_p)$, $u = (u_1, u_2) \equiv (u_\theta\,, u_\phi)$, the drift $F_0(\cdot)$ is given by

$$F_0(x) = \frac{k}{4\pi}\,\frac{1}{1-\cos(\theta)}\,\frac{\partial}{\partial \phi_p}\,,$$

and the control fields are

$$F_1 = \frac{\partial}{\partial \theta_p}\,, \quad F_2 = \frac{\partial}{\partial \phi_p}\,.$$

As demonstrated in [33], the control system (4.17) is controllable.

Figure 4.1 shows the optimal trajectory of a passive particle whose motion is induced by a single vortex (V) with circulation $k = 2$ at the north pole, $V(0) = (0,\, \phi)$. The starting position of the passive particle is $P_0 = (2\pi/3,\, 0)$ and the target is $P_f = (2\pi/3,\, \pi/3)$. The time to perform the displacement is $T = 10$. The number of control variables is $n = 2$ and the resulting optimal controls are $u_0 = (-0.0968,\, 0.1924)$ and $u_1 = (-0.1117,\, 0.2219)$, which are responsible for a value of the objective function equal to $f_2 = 1.1596$.

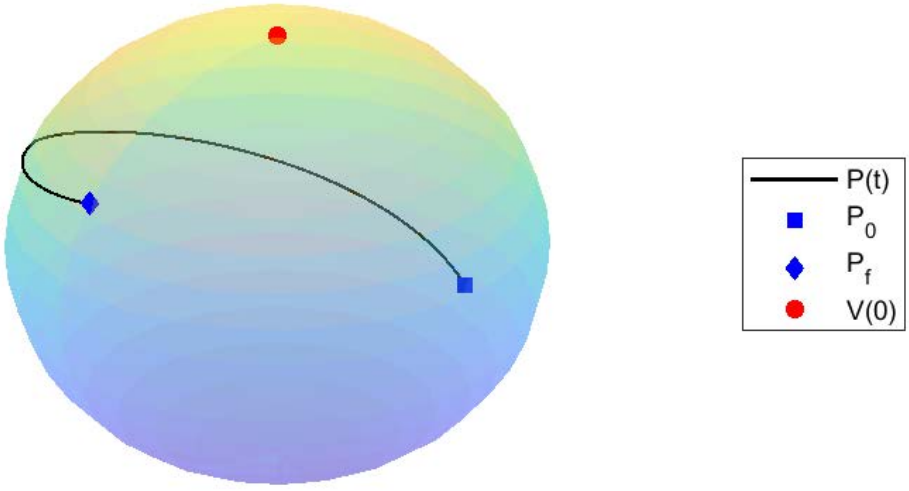


Figure 4.1: Trajectory of the passive particle advected by a single vortex localised in the North Pole: $n = 2$, $P_0 = (2\pi/3, 0)$ and $P_f = (\pi/3, \pi)$.

4.5 A Particle Advected by Two Vortices

In the two point vortices ($N = 2$) and one particle ($P = 1$) problem, the dynamics of the vortices positions $V_1(\tau) = (\theta_1, \phi_1)$ and $V_2(\tau) = (\theta_2, \phi_2)$ are given by

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi} \frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \dot{\phi}_1 = -\frac{k_2}{4\pi} \frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \dot{\theta}_2 = -\frac{k_1}{4\pi} \frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \dot{\phi}_2 = -\frac{k_1}{4\pi} \frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{1-\cos(\gamma_{21})} \end{cases} \tag{4.18}$$

with the given initial conditions $V_1(0) = (\theta_{10},\, \phi_{10})$ and $V_2(0) = (\theta_{20},\, \phi_{20})$. The value of $\cos(\gamma_{12})$ and $\cos(\gamma_{21})$ are computed in agreement with Eq. (4.8).

The dynamics of the passive particle induced by these two vortices and by the control

is given by

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi}\left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\theta \\ \dot{\phi}_p = \frac{1}{4\pi}\left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\phi \end{cases} \tag{4.19}$$

with the given initial conditions $P_0 = (\theta_{p0},\, \phi_{p0})$. As in the previous problem, $u = (u_\theta,\, u_\phi)$ represents the control applied to the passive particle.

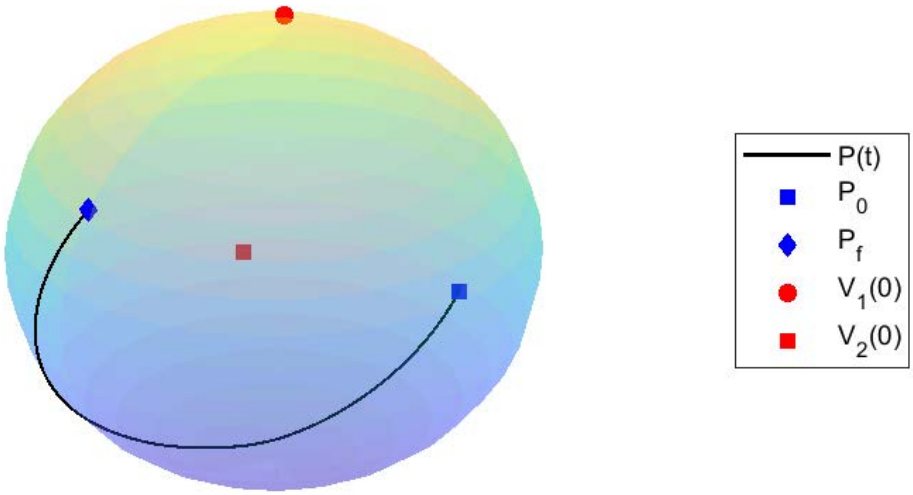


Figure 4.2: Trajectory of a passive particle advected by two point vortices initially located at $V_1(0) = (\pi/10, \pi/4)$ and $V_2(0) = (2\pi/3, \pi/3)$: $n = 2$, $P_0 = (2\pi/3, 0)$ and $P_f = (\pi/3, \pi)$.

Figure 4.2 shows the optimal trajectory of a passive particle whose motion is induced by two vortices (V_1 and V_2). The two vortices have circulation $k_1 = k_2 = 2$ and are initially located at $V_1(0) = (\pi/10, \pi/4)$ and $V_2(0) = (2\pi/3, \pi/3)$, respectively.

The initial position of the passive particle is $P_0 = (2\pi/3,\, 0)$ and the target position is $P_f = (2\pi/3,\, \pi/3)$. The time to perform the shift is $T = 10$. The number of control variables is $n = 2$ and the resulting optimal controls are $u_0 = (0,\, 0)$ and $u_1 = (-0.0699,\, 0.0274)$, which are responsible for a value of the objective function equal to $f_2 = 0.1876$.

4.6 A Particle Advected by Three Vortices

In the problem of one passive particle ($P = 1$) advected by three point vortices ($N = 3$), the dynamics of the vortices positions $V_1(\tau) = (\theta_1, \phi_1)$, $V_2(\tau) = (\theta_2, \phi_2)$ and $V_3(\tau) = (\theta_3, \phi_3)$ are given by Eqs. (4.20), where the Bogomolov notation introduced in (4.9) is used,

$$\begin{cases} \dot{\theta}_1 = -\frac{1}{4\pi}\left(k_2\frac{\alpha_{12}}{1-\cos(\gamma_{12})} + k_3\frac{\alpha_{13}}{1-\cos(\gamma_{13})}\right) \\ \dot{\phi}_1 = \frac{1}{4\pi}\left(k_2\frac{\beta_{12}}{\sin(\theta_1)(1-\cos(\gamma_{12}))} + k_3\frac{\beta_{13}}{\sin(\theta_1)(1-\cos(\gamma_{13}))}\right) \\ \dot{\theta}_2 = -\frac{1}{4\pi}\left(k_1\frac{\alpha_{21}}{1-\cos(\gamma_{21})} + k_3\frac{\alpha_{23}}{1-\cos(\gamma_{23})}\right) \\ \dot{\phi}_2 = \frac{1}{4\pi}\left(k_1\frac{\beta_{21}}{\sin(\theta_2)(1-\cos(\gamma_{21}))} + k_3\frac{\beta_{23}}{\sin(\theta_2)(1-\cos(\gamma_{23}))}\right) \\ \dot{\theta}_3 = -\frac{1}{4\pi}\left(k_1\frac{\alpha_{31}}{1-\cos(\gamma_{31})} + k_2\frac{\alpha_{32}}{1-\cos(\gamma_{32})}\right) \\ \dot{\phi}_3 = \frac{1}{4\pi}\left(k_1\frac{\beta_{31}}{\sin(\theta_3)(1-\cos(\gamma_{31}))} + k_2\frac{\beta_{32}}{\sin(\theta_3)(1-\cos(\gamma_{32}))}\right) \end{cases} \tag{4.20}$$

with the initial conditions $V_1(0) = (\theta_{10},\, \phi_{10})$, $V_2(0) = (\theta_{20},\, \phi_{20})$, and $V_3(0) = (\theta_{30},\, \phi_{30})$ given.

The dynamics of the passive particle induced by these two vortices and by the control is given by:

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi}\left(k_1\frac{\alpha_{p1}}{1-\cos(\gamma_{p1})} + k_2\frac{\alpha_{p2}}{1-\cos(\gamma_{p2})} + k_3\frac{\alpha_{p3}}{1-\cos(\gamma_{p3})}\right) + u_\theta \\ \dot{\phi}_p = \frac{1}{4\pi}\left(k_1\frac{\beta_{p1}}{\sin(\theta_p)(1-\cos(\gamma_{p1}))} + k_2\frac{\beta_{p2}}{\sin(\theta_p)(1-\cos(\gamma_{p2}))} + k_3\frac{\beta_{p3}}{\sin(\theta_p)(1-\cos(\gamma_{p3}))}\right) + u_\phi \end{cases} \tag{4.21}$$

with the given initial conditions $P_0 = (\theta_{p0},\, \phi_{p0})$. As in the previous problems, the variables u_θ and u_ϕ represent the angular controls applied to the passive particle.

Figure 4.3 shows the optimal trajectory of a passive particle whose motion is induced by three vortices with circulation $k_1 = k_2 = k_3 = 2$ and which is initially located at $V_1(0) = (\pi/10, \pi/4)$, $V_2(0) = (2\pi/3, \pi/3)$, and $V_3(0) = (\pi/2, \pi)$.

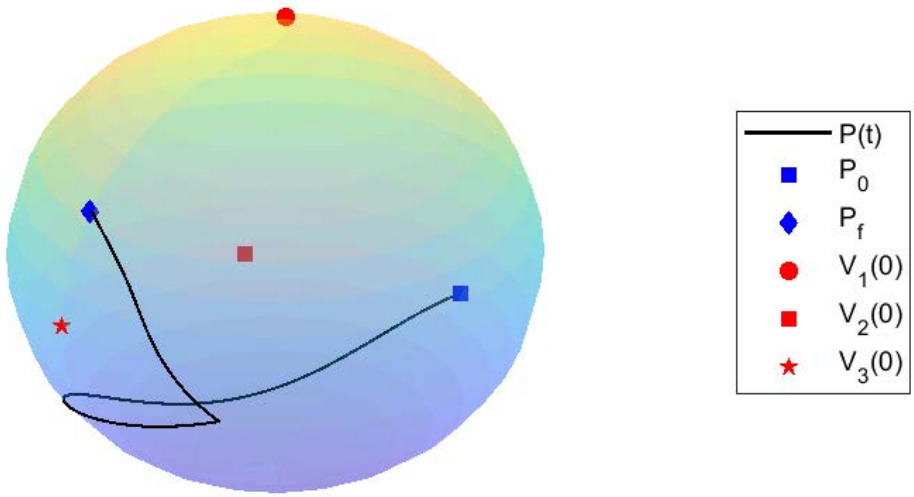


Figure 4.3: Trajectory of a passive particle advected by three point vortices initially located at $V_1(0) = (\pi/10, \pi/4)$, $V_2(0) = (2\pi/3, \pi/3)$, and $V_3(0) = (\pi/2, \pi)$: $n = 2$, $P_0 = (2\pi/3, 0)$ and $P_f = (\pi/3, \pi)$.

The initial position of the passive particle is $P_0 = (2\pi/3,\, 0)$ and the target location is $P_f = (2\pi/3,\, \pi/3)$. The time to perform the shift is $T = 10$. The number of control variables is $n = 2$ and the resulting optimal controls are $u_0 = (-0.0254,\, 0.4252)$ and $u_1 = (-0.0023,\, -0.4456)$, which are responsible for a value of the objective function equal to $f_2 = 2.1788$.

Table 4.1: Summary with the results obtained for $N = 1, 2, 3$ point vortices.

N	u_0		u_1		f_2	CPUt (s)
1	-0.0968	0.1924	-0.1117	0.2219	1.1596	29
2	0.0000	0.0000	-0.0699	0.0274	0.1876	54
3	-0.0254	0.4252	-0.0023	-0.4456	2.1788	3

Table 4.1 summarises the results of the numerical resolution of the three problems presented here. In addition, the CPU time (CPUt), in seconds, is also included for the computations with the MATLAB nonlinear programming solver fmincon, on an Intel Core i7 machine with 16GB RAM.

It can be seen in Table 4.1 that there is no regularity in the variation of the objective function (f_2) and CPU time (CPUt) values. This shows that each problem has its char-

acteristics and that the increase in the number of vortices (N) does not necessarily mean the use of more energy and the need for a longer computation time for its resolution.

4.7 Concluding Remarks

In this chapter, we have considered the motion of passive particles in a (non-rotating) sphere. These particles were driven by point vortices and controls designed to minimise the square of the L^2 norm of the control used during the prescribed time interval for performing the displacement.

It is shown that the system is controllable when the motion is generated by a single vortex ($N = 1$). The problem is also solved by a direct approach, where the control problem is transformed into a nonlinear optimisation problem that is solved numerically. In the case of one ($N = 1$), two ($N = 2$), or three ($N = 3$) point vortices, the numerical results show the existence of near/quasi-optimal control.

In connection with this analysis, several interesting problems arise. Spherical coordinates were used to model the dynamics of the system. What impact does the use of Cartesian coordinates have? On the other hand, real applications of point vortices on a sphere consider the rotating sphere to mimic the rotation of the Earth. How does the sphere's rotation affect the control of the passive particles? These issues will be addressed in forthcoming chapters.

Chapter 5

Modelling the Dynamics with Cartesian Coordinates

The results from this chapter have already been published as:
Balsa, C., Monville Letu, R., Gama, S. (2023). *Optimization of Vortex Dynamics on a Sphere*. In: Garcia, M.V., Gordón-Gallegos, C. (eds) CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI). CSEI 2022. Lecture Notes in Networks and Systems, vol. 678. Springer, Cham. ISBN 978-3-031-30591-7. https://doi.org/10.1007/978-3-031-30592-4_15.

In Chapter 4 we have considered the motion of passive particles driven by point vortices in a (non-rotating) sphere. The system's dynamics was achieved using spherical coordinates. A control was designed to minimise the square of the L^2 norm of the control used during the prescribed time interval for performing the displacement. The resulting control problem is solved numerically by a direct approach, where the control problem is transformed into a nonlinear optimisation problem that is solved numerically. In the case of one ($N = 1$), two ($N = 2$), or three ($N = 3$) point vortices, the numerical results show the existence of near/quasi-optimal control.

In this chapter, which is based on [19], we explore using Cartesian coordinates to model the system's dynamics composed of the passive particle and N point vortices. As in the previous chapter, the resulting nonlinear optimization problem (NOP) is solved numeri-

cally using the built-in Matlab `fmincon` that allows solving constrained NOPs by different methods. We compare the results obtained with two of these optimization methods: Interior Point [123] and the Active Set [108]. Additionally, we analyse the effect of improving the number of constant controls from one to four ($n = 1, 2, 3, 4$).

The chapter is organised as follows. Section 5.1 introduces the control problem underlying the determination of the optimal trajectory. The transformation of the control problem into an NOP problem in the case of a single vortex ($N = 1$) is analysed in section 5.2. In section 5.3, the problem of a flow induced by multiple vortices ($N = 2, 3, 4$) is treated. Finally, in section 5.4 some concluding considerations are given.

5.1 Statement of the Control Problem

We begin by introducing the equations of the vortex on a nonrotating sphere whose centre is the origin and whose radius is R. The representation of the vortex and the passive particle can be in different coordinate systems, e.g., spherical coordinates or Cartesian coordinates (see, e.g., [88]). In this chapter, we use only Cartesian coordinates.

The position of a vortex on a sphere is given by the vector $\mathbf{x}_j$ that points from the centre of the sphere to the vortex location $\mathbf{x}_j = (x_j, y_j, z_j)$ on the spherical surface $||\mathbf{x}_j|| = R$. The dynamics of the vortex is is determined by

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_i}{||\mathbf{x}_i - \mathbf{x}_j||^2}\,, \qquad i = 1, 2, \ldots, N\,, \tag{5.1}$$

with the respective initial conditions, where k_j is the circulation of vortices j, and N is the total number of point vortices on the sphere. We can observe that the chord distance between vortex i and j is given by

$$||\mathbf{x}_i - \mathbf{x}_j||^2 = 2\left(R^2 - \mathbf{x}_i \cdot \mathbf{x}_j\right)\,. \tag{5.2}$$

A passive particle is, by definition, a point vortex with circulation $k = 0$. Thus, the dynamics of a system with P passive particles advected by N point vortices is given

by Eq. (5.1) together with the equations for the passive particles

$$\dot{\mathbf{x}}_p = \frac{1}{2\pi R}\sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_p}{||\mathbf{x}_p - \mathbf{x}_j||^2} \ , \qquad p = N+1, N+2, \ldots, N+P \, , \tag{5.3}$$

with the respective initial conditions.

Considering a single controlled passive particle ($P = 1$) moving in a spherical fluid induced by N, the corresponding equation is

$$\dot{\mathbf{x}} = \frac{1}{2\pi R}\sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_j||^2} + \mathbf{U}\,(t) \ , \tag{5.4}$$

with the respective initial conditions. In the right and side (RHS) of Eq. (5.4), $\mathbf{U}(t)$ is the control vector function.

Since the motion of the vortices and particles takes place on a sphere with radius R, this means that the control $\mathbf{U}$ must be such that the position vector of the passive particle $\mathbf{x}$ has a norm equal to R, that is, $\mathbf{x} \cdot \mathbf{x} = R^2$. Applying $\mathbf{x}\cdot$ to both sides of the Eq. (5.4), we obtain:

$$\mathbf{x} \cdot \dot{\mathbf{x}} = \mathbf{x} \cdot \mathbf{U} \, ,$$

or, equivalently,

$$\frac{1}{2}\frac{d(\mathbf{x}\cdot\mathbf{x})}{dt} = \mathbf{x}\cdot\mathbf{U} \quad \Longleftrightarrow \quad \frac{1}{2}\frac{d(R^2)}{dt} = \mathbf{x}\cdot\mathbf{U} \, ,$$

Thus, $\mathbf{x} \cdot \mathbf{U} = 0$. In other words, for the particle's motion to occur on the sphere, the control exerted in the particle must be orthogonal to the particle's position vector, i.e. $\mathbf{U}(t) \perp \mathbf{x}(t)$, $\forall t \geq 0$. If we let $\mathbf{x}(t) = (x(t), y(t), z(t))$, we then get

$$\mathbf{U}(t) = \alpha(t)\ (y(t), -x(t), 0) \ + \ \beta(t)\ (0, z(t), -y(t)) \, ,$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are two scalar control functions. These controls allow the particle to move in any direction on the sphere's surface. The sum of the square of their magnitude, $\alpha(\cdot)^2 + \beta(\cdot)^2$, by definition, corresponds to the energy expended for the displacement. Since the control $\mathbf{U}(\cdot)$ depends on $\alpha(\cdot)$ and $\beta(\cdot)$ from now on we assume $\mathbf{U}(\cdot) = (\alpha(\cdot), \beta(\cdot))$,

and Eq. (5.4) is rewritten as

$$\dot{\mathbf{x}} = \frac{1}{2\pi R} \sum_{j=1}^{N} k_j \, \frac{\mathbf{x}_j \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_j||^2} + \alpha \, (y, -x, 0) + \beta \, (0, z, -y) \,, \tag{5.5}$$

This chapter considers that a given single passive particle ($P = 1$) is advected by $N = 1, 2, 3$ or 4 point vortices. The control problem is to move this particle between two given points ($\mathbf{x}_0$ and $\mathbf{x}_f$), on the surface of the sphere, in a given fixed time (T), while consuming as little energy as possible.

5.2 Conversion to an Optimization Problem

The control problem presented in the previous section is solved by a direct approach based on numerical optimisation of the corresponding discretised problem. In previous work, the same methodology was used to control the displacement of a passive particle in the plane [17]. The control function $\mathbf{U}(\cdot)$ is replaced by n control vectors $\mathbf{u}_0$, $\mathbf{u}_1, \cdots, \mathbf{u}_{n-1}$. Numerical calculations were performed in MATLAB using the nonlinear programming solver `fmincon`. This solver provides optimisation algorithms, such as the Interior Point or the Active Set (see [77]).

We start by solving this problem for a single passive particle in a spherical vortex flow, and then in Section 5.3, we treat the cases with up to four vortices.

The equation that describes the motion of a passive particle induced by a single vortex on the sphere is

$$\dot{\mathbf{x}} = \frac{k}{2\,\pi\,R} \, \frac{\mathbf{x}_1 \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_1||^2} \,, \tag{5.6}$$

with the given initial condition $\mathbf{x}(0) = \mathbf{x}_0$ and the control problem, previously introduced, is then defined as follows:

$$
\begin{aligned}
(\mathcal{P}) \quad & \text{Minimize :} \\
& \qquad \int_0^T \|\mathbf{U}\,(t)\|^2\, dt \\
& \text{subject to :} \\
& \qquad \dot{\mathbf{x}} = \tfrac{k}{2\,\pi\,R}\ \tfrac{\mathbf{x}_1 \times \mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{U}(t) \\
& \qquad \mathbf{x}(0) = \mathbf{x}_0 \\
& \qquad \mathbf{x}(T) = \mathbf{x}_f \\
& \qquad \|\mathbf{U}\| \leq u_{\max}
\end{aligned}
$$

with $\mathbf{U}(t) \in \mathbb{R}^3$, and $\mathbf{x}_0\,, \mathbf{x}_f \in \mathbb{R}^3$, $T > 0$ and $u_{\max} > 0$ given. In the control problem $(\mathcal{P})$, $\mathbf{x}(0)$ is the initial position (when $t = 0$) and $\mathbf{x}(T)$ is the final position of the passive particle (when $t = T$).

For example, in this optimisation problem $(\mathcal{P})$, the objective function (cost function) represents the energy expended by the controller $\mathbf{U}(\cdot)$ to drive the passive particle from the starting point $\mathbf{x}_0$ to the end point $\mathbf{x}_f$. The first restriction corresponds to the state equation that determines the position $\mathbf{x}(t)$ of the particle as a function of time. The vectorial control function $\mathbf{U}(t)$ is introduced into this equation to move the particle from $\mathbf{x}_0$ to $\mathbf{x}_f$ in a fixed time value $T > 0$. The points $\mathbf{x}_0$ and $\mathbf{x}_f$ are predefined, as is the time T available to reach the destination $\mathbf{x}_f$. In addition, the fourth restriction requires that the norm of the control vector is not greater than a given value $u_{\max}$.

To solve this problem, we proceed to discretise the control function $\mathbf{U}(\cdot)$ into n (discrete) vector variables defined as $(t_0 = 0,\ t_n = T)$:

$$
\begin{aligned}
\mathbf{U}(t) &= \mathbf{u}_0 \qquad \text{if} \quad t_0 \leq t < t_1\,, \\
\mathbf{U}(t) &= \mathbf{u}_1 \qquad \text{if} \quad t_1 \leq t < t_2\,, \\
\mathbf{U}(t) &= \mathbf{u}_2 \qquad \text{if} \quad t_2 \leq t < t_3\,, \\
&\qquad\ \vdots \\
\mathbf{U}(t) &= \mathbf{u}_{n-1} \qquad \text{if} \quad t_{n-1} \leq t \leq t_n\,.
\end{aligned}
$$

Thus, each $\mathbf{u}_i$ $(i = 0, 1, 2, \cdots, n-1)$ corresponds to the vector of control exercised in the subinterval $[t_{i-1}\ t_i)$. All these subintervals have amplitudes equal to $\Delta t = (t_n - t_0)\,/n$. The discretisation of the objective function by the rectangular rule leads to the approximation

$$\int_0^T \|\mathbf{U}\|^2 \, dt \approx \Delta t \sum_{j=0}^{n-1} \|\mathbf{u}_j\|^2 \equiv f_n \, . \tag{5.7}$$

The control problem $(\mathcal{P})$ is then replaced by its discretized version:

$(\mathcal{DP}_n)$ Minimize :

$$f_n = \Delta t \sum_{j=0}^{n-1} \|\mathbf{u}_j\|^2$$

subject to :

$$\begin{aligned}
&\dot{\mathbf{x}} = \tfrac{k}{2\,\pi\,R}\,\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_0, \;\; \mathbf{x}(0) = \mathbf{x}_0, \;\; \|\mathbf{u}_0\| \le u_{\max}, \;\; t_0 \le t < t_1 \\
&\dot{\mathbf{x}} = \tfrac{k}{2\,\pi\,R}\,\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_1, \;\; \mathbf{x}(t_1) = \mathbf{x}_{t_1}, \;\; \|\mathbf{u}_2\| \le u_{\max}, \;\; t_1 \le t < t_2 \\
&\vdots \\
&\dot{\mathbf{x}} = \tfrac{k}{2\,\pi\,R}\,\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_{n-1}, \;\; \mathbf{x}(t_{n-1}) = \mathbf{x}_{t_{n-1}}, \;\; \|\mathbf{u}_{n-1}\| \le u_{\max}, \;\; t_{n-1} \le t < t_n \\
&\mathbf{x}(t_n) = \mathbf{x}_f
\end{aligned}$$

First, we solve the discretised problem $(\mathcal{DP}_n)$ where n is the number of control vectors (in our study from 1 to 4), considering a single vortex with circulation $k = 2$ at the north pole of the sphere ($\mathbf{x}_1 = (0,\ 0,\ 1)$). A set of vectors $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1} \in \mathbb{R}^3$ is sought that drives the passive particle from $\mathbf{x}_0$ to $\mathbf{x}_f$ in exactly $T = 10$ (natural) time units and minimizes the objective function f_n defined by Eq. (5.7). This optimisation problem is solved numerically using the Interior Point and the Active Set optimisation algorithms, which are included in the `fmincon` Matlab solver. In the numerical calculations, $R = 1$ is assumed.

The constrained optimisation problem $(\mathcal{DP}_n)$ provides a set of n control vectors $\mathbf{u}_0$, $\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$, subject to the constraint that the final position of the particle must be equal to $\mathbf{x}_f$. In the implementation of this problem, we assume that the target position is reached when $\|\mathbf{x}(T) - \mathbf{x}_f\| \le 10^{-4}$. To verify this condition, it is necessary to solve the ordinary differential equations (ODEs) sequentially

$$\dot{\mathbf{x}} = \frac{k}{2\,\pi\,R}\,\frac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_i, \qquad i = 0, 1, \ldots, n-1 \, . \tag{5.8}$$

The final position of the previous IVP gives the initial condition of each initial value problem (IVP). For each IVP, the ordinary differential equation (ODE) is solved numerically using Matlab's built-in function `ode45`, the 4th and 5th order Runge-Kutta methods.

In Eq (5.8), $\mathbf{u}_i \in \mathbb{R}^3$ is the control vector used in each of the subintervals corresponding to $i = 0, 1, \ldots, n-1$. As mentioned, these vectors cannot be arbitrary because the particle can only move on the sphere's surface. Thus, for any particle position $\mathbf{x} = (x, y, x)$, the control vector $\mathbf{u}_i = (u_{ix}, u_{iy}, u_{iz})$ must be tangent to the sphere (orthogonal to the position vector $\mathbf{x}$), i.e. $\mathbf{x} \cdot \mathbf{u}_i = 0$ for $i = 0, 1, \ldots, n-1$. This condition implies that the control vector has the following structure

$$\mathbf{u}_i = (u_{ix}, u_{iy}, u_{iz}) = (\alpha_i\, y, -\alpha_i\, x + \beta_i\, z, -\beta_i\, y)\,, \tag{5.9}$$

where α_i and β_i are two real constants. For this reason, finding the optimal controls $\mathbf{u}_0$, $\mathbf{u}_1$, ..., $\mathbf{u}_{n-1}$ is equivalent to finding the optimal control parameters α_i and β_i for each of the subintervals $i = 0, 1, \ldots, n-1$.

As a consequence of Eq. (5.9), Eq. (5.8) can be written row wise

$$\begin{cases} \dot{x} = \frac{k}{2\,\pi\,R}\,\frac{-y}{x^2+y^2+(z-R)^2} + \alpha_i\, y \\ \dot{y} = \frac{k}{2\,\pi\,R}\,\frac{x}{x^2+y^2+(z-R)^2} - \alpha_i\, x + \beta_i\, z \quad , \qquad i = 0, 1, \ldots, n-1 \\ \dot{z} = -\beta_i\, y \end{cases} \tag{5.10}$$

The results of solving the problem $(\mathcal{DP}_n)$ obtained with the Interior Point algorithm, are shown in Table 5.1, and the results obtained with the Active Set algorithm are demonstrated in Table 5.2. In each of these tables, the controls obtained by the optimisation algorithms for the displacement of the passive particle from the original position $\mathbf{x}_0 = (0.8860, 0.0, -0.5)$ to the target position $\mathbf{x}_f = (-0.8860, 0.0, 0.5)$ are shown. The corresponding values of the objective function f_n and the computation times (CPUt), in seconds, are also included.

Table 5.1: Optimal controls obtained with the Interior Point optimisation algorithm.

	$\mathbf{u}_0$		$\mathbf{u}_1$		$\mathbf{u}_2$		$\mathbf{u}_3$			
n	α_0	β_0	α_1	β_2	α_3	β_3	α_4	β_4	f_n	CPUt
				$N=1$						
1	-0.096	-0.156							1.83	0.71
2	-0.094	-0.317	0.000	0.000					0.83	3.95
3	-0.126	-0.474	0.000	0.000	0.020	-0.065			0.93	1.07
4	-0.177	-0.568	0.000	0.000	0.000	0.000	-0.013	-0.207	1.00	4.73
				$N=2$						
1	-0.213	0.044							2.18	1.47
2	0.000	0.000	-0.386	-0.072					0.98	3.12
3	0.000	0.000	0.000	0.000	-0.540	-0.209			0.97	4.13
4	0.000	0.000	0.000	0.0000	0.000	0.000	-0.685	-0.349	0.96	4.73
				$N=3$						
1	-0.130	-0.153							2.00	2.50
2	-0.307	-0.108	0.000	0.000					0.81	5.80
3	-0.424	-0.112	0.000	0.000	-0.017	-0.005			0.76	5.16
4	-0.537	-0.133	0.000	0.000	0.000	0.000	-0.442	-0.011	0.73	7.03
				$N=4$						
1	-0.144	-0.102							1.76	1.68
2	-0.230	-0.095	-0.070	-0.008					0.80	6.22
3	-0.310	-0.126	0.000	0.000	-0.114	-0.025			0.75	5.16
4	-0.403	-0.151	0.000	0.000	0.000	0.000	-0.146	-0.039	0.73	9.98

The results obtained with the Interior Point method (Table 5.1) show no regular variation of f_n as a function of the number of controls n. The same is true for the CPU computation times (CPUt).

Comparing the results obtained with the Interior Point and Active Set methods, we can see that the results are similar, except for $n = 4$, where the Active Set is more efficient in finding the controls that minimise the objective function and reduce the computation time.

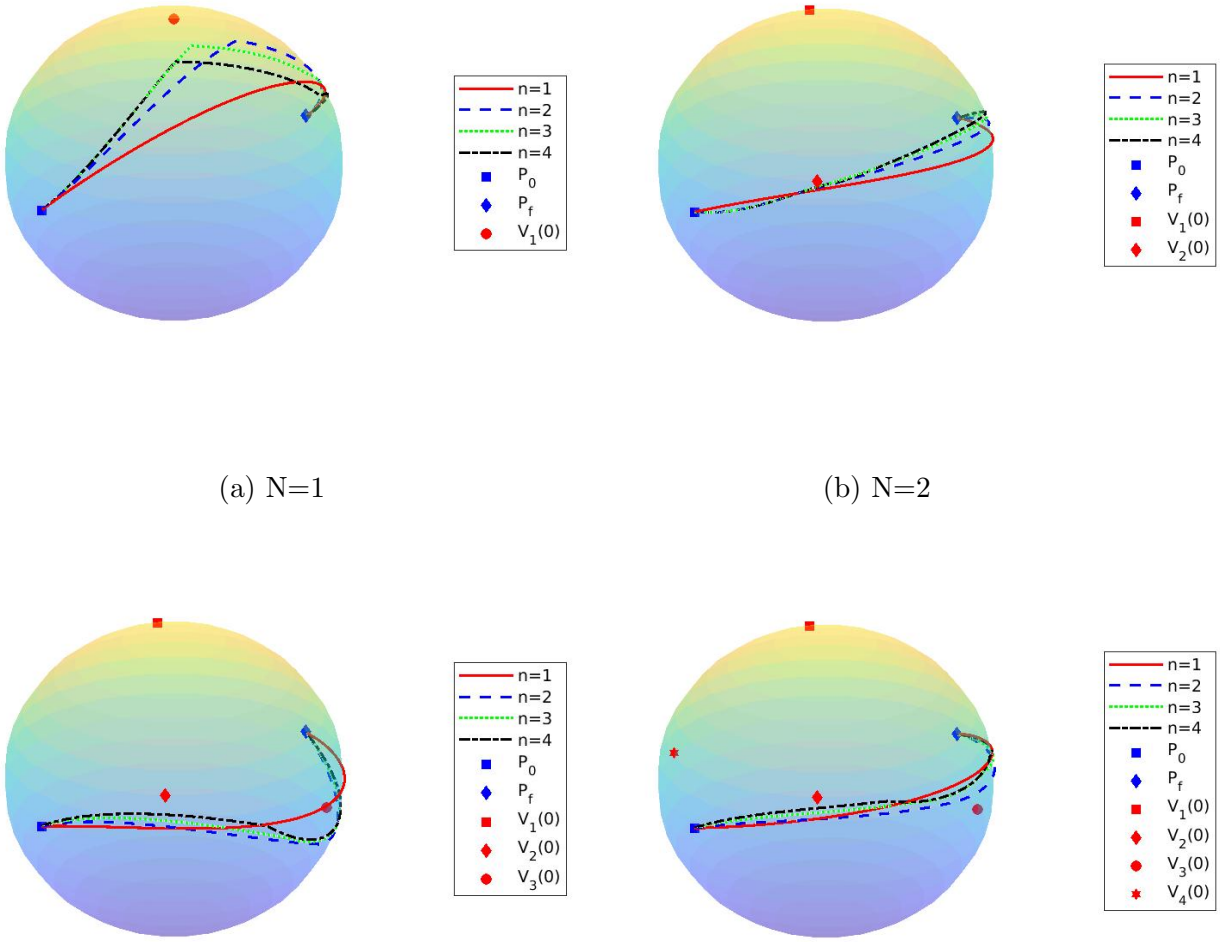
Figure 5.1 shows the trajectories obtained with the controls shown in Table 5.1, and Figure 5.2 shows the trajectories obtained with the controls shown in Table 5.2. It can be seen that the trajectories generated by the two methods (Figure 5.1(a) and Figure 5.2(a)) are very close for $n = 1, 2$, and 3 and differ a little for $n = 4$.

Table 5.2: Optimal controls obtained with the Active Set optimisation algorithm.

	$\mathbf{u}_0$		$\mathbf{u}_1$		$\mathbf{u}_2$		$\mathbf{u}_3$			
n	α_0	β_0	α_1	β_2	α_3	β_3	α_4	β_4	f_n	CPUt
				$N=1$						
1	-0.096	-0.156							1.83	0.49
2	-0.101	-0.312	0.000	0.000					0.82	0.77
3	-0.117	-0.487	0.000	0.000	0.013	-0.047			0.92	1.09
4	-0.139	-0.652	0.000	0.000	0.016	-0.063	0.000	0.000	0.10	1.62
				$N=2$						
1	-0.213	0.044							2.18	0.63
2	0.000	0.000	-0.384	-0.073					0.98	3.40
3	0.000	0.000	0.000	0.000	-0.537	-0.209			0.96	3.58
4	0.000	0.000	0.000	0.000	0.000	0.000	-0.681	-0.348	0.96	8.68
				$N=3$						
1	-0.130	-0.153							2.00	0.93
2	-0.305	-0.108	0.000	0.000					0.81	2.86
3	-0.423	-0.110	0.000	0.000	-0.014	-0.004			0.75	5.20
4	-0.545	-0.121	0.000	0.000	0.000	0.000	-0.019	-0.007	0.72	9.02
				$N=4$						
1	-0.144	-0.102							1.76	1.50
2	-0.230	-0.095	-0.070	-0.008					0.80	5.32
3	-0.313	-0.126	0.000	0.000	-0.108	-0.022			0.76	7.52
4	-0.404	-0.152	0.000	0.000	0.000	0.000	-0.140	-0.035	0.72	12.52

5.3 Flow Created by Several Vortices

In this section, we treat the displacement problem of a passive particle entrained by several vortices (N). As before, we want to find the vectors $\mathbf{u}_0$, $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ which pull the particle from $\mathbf{x}_0 = (0.8860, 0.0, -0.5)$ to $\mathbf{x}_f = (-0.8860, 0.0, 0.5)$. We also consider the time $T = 10$, and the same circulation for all vortices $k_i = 2$, for $i = 1, 2, 3, 4$. As before, to solve the three optimisation problems corresponding to $N = 2, 3, 4$, we used the Interior Point and Active Set methods described in the `fmincon` Matlab solver (see [77]).



(a) N=1 (b) N=2

(c) N=3 (d) N=4

Figure 5.1: Optimal trajectories determined with the Interior Point optimisation method.

5.3.1 Flow Created by Two Vortices (N=2).

The position of the two vortices ($\mathbf{x}_1$ and $\mathbf{x}_2$) are governed by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{k}{2\,\pi\,R}\; k_2\, \frac{\mathbf{x}_2 \times \mathbf{x}_1}{||\mathbf{x}_1 - \mathbf{x}_2||^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{k}{2\,\pi\,R}\; k_1\, \frac{\mathbf{x}_1 \times \mathbf{x}_2}{||\mathbf{x}_2 - \mathbf{x}_1||^2} \end{cases} \tag{5.11}$$

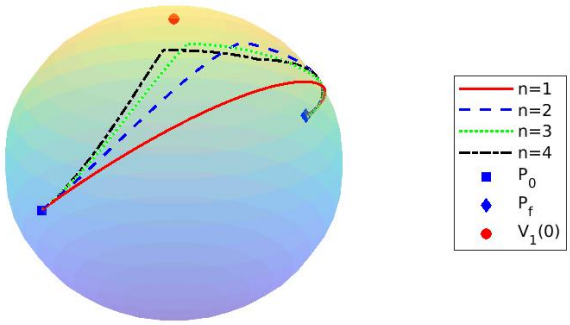
with the respective initial conditions $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$. In this case, the initial positions of the vortices are $\mathbf{x}_1(0) = (0.2185, 0.2815, 0.9511)$, and $\mathbf{x}_2(0) = (0.4330, 0.74, -0.5)$, and the circulations are $k_1 = k_2 = 2$. The passive particle, initially at $\mathbf{x}(0) = \mathbf{x}_0$, is governed by the equation

$$\dot{\mathbf{x}} = \frac{k}{2\,\pi\,R}\left(k_1 \frac{\mathbf{x}_1 \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_1||^2} + k_2 \frac{\mathbf{x}_2 \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_2||^2}\right) + \alpha\;(y, -x, 0) + \beta\;(0, z, -y)\,, \tag{5.12}$$

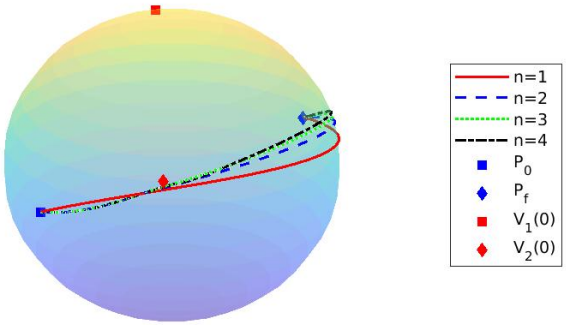
where α and β are the controllers.

This control problem is solved like the problem with a single vortex ($N = 1$) presented in Section 5.2. The problem is discretised in time and transformed into an optimal control problem similar to $(\mathcal{DP}_n)$. Since the control vectors $\mathbf{u}_0$, $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ have the same form defined by (5.9), the problem reduces to determining the control parameters α_i and β_i for each of the subintervals $i = 0, 1, 2, 3$.

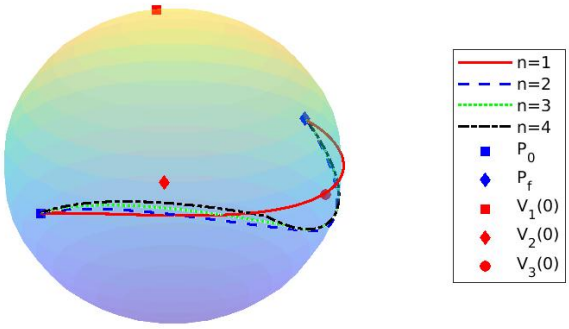
The results obtained with the Interior Point method are given in Table 5.1, and the results obtained with the Active Set method are given in Table 5.2. The corresponding trajectories are shown in Figures 5.1(b) and 5.2(b). It can be seen that the results obtained with the two methods are very similar. There is a tendency for the objective function to decrease as the number of subintervals n increases. The computation times also increase with n, which is more pronounced for the Active Set method. Since the control parameters are very similar, the trajectories obtained with the two methods are practically the same.



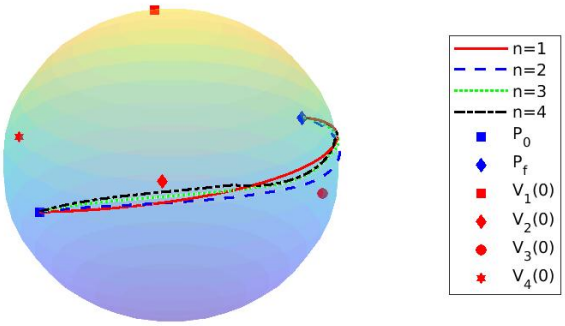
(a) N=1



(b) N=2



(c) N=3



(d) N=4

Figure 5.2: Optimal trajectories determined with the Active Set optimisation method.

5.3.2 Flow Created by Three Vortices (N=3).

The position of the three point vortices ($\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$) are governed by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{k}{2\,\pi\,R}\left(k_2 \frac{\mathbf{x}_2 \times \mathbf{x}_1}{||\mathbf{x}_1 - \mathbf{x}_2||^2} + k_3 \frac{\mathbf{x}_3 \times \mathbf{x}_1}{||\mathbf{x}_1 - \mathbf{x}_3||^2}\right) \\ \dot{\mathbf{x}}_2 = \frac{k}{2\,\pi\,R}\left(k_1 \frac{\mathbf{x}_1 \times \mathbf{x}_2}{||\mathbf{x}_2 - \mathbf{x}_1||^2} + k_3 \frac{\mathbf{x}_3 \times \mathbf{x}_2}{||\mathbf{x}_2 - \mathbf{x}_3||^2}\right) \\ \dot{\mathbf{x}}_3 = \frac{k}{2\,\pi\,R}\left(k_1 \frac{\mathbf{x}_1 \times \mathbf{x}_3}{||\mathbf{x}_3 - \mathbf{x}_1||^2} + k_2 \frac{\mathbf{x}_2 \times \mathbf{x}_3}{||\mathbf{x}_3 - \mathbf{x}_2||^2}\right) \end{cases} \tag{5.13}$$

with the respective initial conditions $\mathbf{x}_1(0)$, $\mathbf{x}_2(0)$ and $\mathbf{x}_3(0)$. The initial position of the vortices are $\mathbf{x}_1(0) = (0.2185, 0.2815, 0.9511)$, $\mathbf{x}_2(0) = (0.4330, 0.74 - 0.5)$, and $\mathbf{x}_3(0) = (-1.0, 0.0, 0.0)$, and the circulations are $k_1 = k_2 = k_3 = 2$.

The following equation gives the passive particle position

$$\dot{\mathbf{x}} = \frac{k}{2\,\pi\,R} \sum_{i=1}^{3} k_i \frac{\mathbf{x}_i \times \mathbf{x}}{||\mathbf{x} - \mathbf{x}_i||^2} + \alpha\,(y, -x, 0) + \beta\,(0, z, -y)\,, \tag{5.14}$$

with the given initial position $\mathbf{x}(0) = x_0$. This control problem is solved by a direct approach similar to the previously described problem with one ($N = 1$) and two ($N = 2$) vortices.

The results obtained with the Interior Point method are given in Tables 5.1, and the results obtained with the Active Set method are given in Table 5.2. The corresponding trajectories are shown in Figures 5.1(c) and 5.2(c). As in the previous problem ($N = 2$), it can be seen that the results of the two methods are very similar. The objective function also decreases slightly as n increases. The computation times also increase with n and are higher than the computation times of the problems with $N = 1$ and $N = 2$. The trajectories obtained with the Interior Point and Active Set methods, shown in Figures 5.1(c) and 5.2(c), are practically identical.

5.3.3 Flow Created by Four Vortices (N=4).

The position of the four point vortices ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$) are governed by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{k}{2\,\pi\,R}\left(k_2\,\frac{\mathbf{x}_2\times\mathbf{x}_1}{||\mathbf{x}_1-\mathbf{x}_2||^2} + k_3\,\frac{\mathbf{x}_3\times\mathbf{x}_1}{||\mathbf{x}_1-\mathbf{x}_3||^2} + k_4\,\frac{\mathbf{x}_4\times\mathbf{x}_1}{||\mathbf{x}_1-\mathbf{x}_4||^2}\right) \\ \dot{\mathbf{x}}_2 = \frac{k}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1\times\mathbf{x}_2}{||\mathbf{x}_2-\mathbf{x}_1||^2} + k_3\,\frac{\mathbf{x}_3\times\mathbf{x}_2}{||\mathbf{x}_2-\mathbf{x}_3||^2} + k_4\,\frac{\mathbf{x}_4\times\mathbf{x}_2}{||\mathbf{x}_2-\mathbf{x}_4||^2}\right) \\ \dot{\mathbf{x}}_3 = \frac{k}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1\times\mathbf{x}_3}{||\mathbf{x}_3-\mathbf{x}_1||^2} + k_2\,\frac{\mathbf{x}_2\times\mathbf{x}_2}{||\mathbf{x}_3-\mathbf{x}_2||^2} + k_4\,\frac{\mathbf{x}_4\times\mathbf{x}_3}{||\mathbf{x}_3-\mathbf{x}_4||^2}\right) \\ \dot{\mathbf{x}}_4 = \frac{k}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1\times\mathbf{x}_4}{||\mathbf{x}_4-\mathbf{x}_1||^2} + k_2\,\frac{\mathbf{x}_2\times\mathbf{x}_4}{||\mathbf{x}_4-\mathbf{x}_2||^2} + k_3\,\frac{\mathbf{x}_3\times\mathbf{x}_4}{||\mathbf{x}_4-\mathbf{x}_3||^2}\right) \end{cases} \tag{5.15}$$

with the respective initial conditions $\mathbf{x}_1(0)$, $\mathbf{x}_2(0)$, $\mathbf{x}_3(0)$ and $\mathbf{x}_4(0)$. In this case, the initial vortex positions are $\mathbf{x}_1(0) = (0.2185, 0.2815, 0.9511)$, $\mathbf{x}_2(0) = (0.4330, 0.74 - 0.5)$, $\mathbf{x}_3(0) = (-1.0, 0.0, 0.0)$, and $\mathbf{x}_4(0) = (1.0, 0.0, 0.0)$, and the circulations are $k_1 = k_2 = k_3 = k_4 = 2$.

The dynamic of the passive particle induced by the four vortices is described by

$$\dot{\mathbf{x}} = \frac{k}{2\,\pi\,R}\sum_{i=1}^{4} k_i\,\frac{\mathbf{x}_i\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_i||^2} + \alpha\;(y,-x,0) + \beta\;(0,z,-y)\,, \tag{5.16}$$

with a given initial condition $\mathbf{x}(0) = \mathbf{x}_0$. The same direct numerical approach used in the previous problems described in Section 5.2 is used here to solve this control problem.

The results obtained with four vortices are shown in Tables 5.1 and 5.2 and the corresponding trajectories are shown in Figures 5.1(d) and 5.2(d). As in the previous problems ($N = 1, 2, 3$), the results of the two methods are very similar and the corresponding value of the objective function also decreases slightly with increasing n. The computation times also increase with n and are higher than those obtained for the problems with $N = 1, 2, 3$. The corresponding trajectories obtained by the two methods (see Figures 5.1(c) and 5.2(c)) are practically identical.

5.4 Concluding Remarks

This chapter is concerned with determining the optimal trajectory of a passive particle moving in a flow induced by one or more vortices between two points on the surface of a

sphere. The dynamics governing the motion of the passive particle and point vortex are modelled using Cartesian coordinates. These equations integrate the constraints on particle motion into the resulting optimal control problem. The numerical strategy presented to solve this problem is an alternative to the conventional methods of optimal control theory. The control problem is transformed into a discretized optimisation problem that is solved by a numerical method.

Two different optimisation methods were tested, the Interior Point and Active Set methods. Both could find optimal trajectories regardless of the time discretisation level of the problem and the number of point vortices involved. The computation times required by both methods are of the same order of magnitude. However, the Active Set requires more time when the number of control parameters to be determined is higher.

For each of the four problems analysed, the value of the objective function tends to decrease with the increase in the number of control parameters. This result shows that increasing the number of control parameters allows greater freedom in the trajectory choice, minimising the energy used for the displacement. On the other hand, the computation time is higher.

As the number of vortices increases, the computation time also increases. This is because more calculations are required to solve the ODEs that model the dynamics of point vortices and of the passive particle.

Chapter 6

Exploring and Comparing Controls

The results from this chapter have already been published as:
Balsa, C., Otero-Espinar, M. V., Gama, S. (2024). *Exploring Controlled Passive Particle Motion Driven by Point Vortices on a Sphere*. Computation, 12(2), 23. ISSN 2079-3197. https://doi.org/10.3390/computation12020023.

In Chapter 4 the problem of controlling the motion of a passive particle in a spherical flow induced by point vortices was solved by a direct numerical approach. In this approach the equations that govern the dynamics of the system composed of several vortices and a single passive particle were deduced using spherical coordinates (see also [16]). It was observed that, in scenarios involving one, two, and three vortices, it is possible to show the existence of near or quasi-optimal trajectories for the passive particle.

In Chapter 5 the problem was addressed using Cartesian coordinates, with controls introduced in the form of tangent vectors to the sphere (see also [19]). In this context as well, it was found that it is possible to obtain near or quasi-optimal trajectories for the passive particle, which is advected by one, two, three, or four vortices.

The possibility of modelling the problem through two different coordinate systems raises the question of whether the solution depends on the chosen coordinate system or if, on the contrary, it is unique regardless of the system. The study presented in this chapter was published in [22] and focuses on the answer to this question. The formulation of the

problem is deduced in spherical and Cartesian coordinates. It is then shown how it is possible to convert the problem formulation from one coordinate system to another.

The chapter is structured as follows: In Section 6.1, we undertake the derivation of equations governing the dynamics of the passive particle in Cartesian coordinates. Section 6.2 explores the modeling of dynamics in spherical coordinates. The formulation of optimization problems, depending on the type of coordinates, and the presentation of the numerical strategy employed for their resolution are detailed in Section 6.3. Results obtained from the solution of discretized optimization problems are outlined in Section 6.4.

6.1 Modelling with Cartesian Coordinates

The problem under consideration involves the movement of a passive particle between two specified points on the surface of a non-rotating sphere within a fixed time. The particle's autonomy in terms of displacement is governed by control, allowing it to move freely in any direction on the surface of the sphere.

In this section, we introduce the equations that govern the motion of controlled passive particles induced by a vortex on the sphere. The sphere is centered at the origin with a radius of R.

The evolution of vorticity, ω, on the unit sphere, S, is governed by the Euler equation $D_t\,\omega = 0$, where $\omega = \mathbf{x}\cdot(\nabla\times\mathbf{u})$ [94, 50]. Here, $D_t \equiv \partial_t + \mathbf{u}\cdot\nabla$ is the material derivative, $\mathbf{u}$ is a two-dimensional incompressible velocity field, and the vector $\mathbf{x}$ points from the center of the sphere to a point in the sphere. The incompressiblity condition $\nabla\cdot\mathbf{u} = 0$ indicates the presence of a streamfunction $\psi(\mathbf{x})$, where

$$\mathbf{u} = \mathbf{x}\times\nabla\psi\,. \tag{6.1}$$

The latter equality leads to the Poisson equation:

$$\nabla^2\psi = \omega\,. \tag{6.2}$$

The vorticity arising from point vortices corresponds to a specific scenario where the

vorticity field is localized at discrete points. In other words, assuming that: (i) there are N point vortices on the unit sphere, each characterized by Cartesian coordinates $\mathbf{x}_i = (x_i, y_i, z_i)$, and circulation k_i, $(i = 1, 2, \ldots, N)$, and (ii) the vorticity has the form

$$\omega = \frac{1}{\sin(\theta)} \sum_{i=1}^{N} k_i \, \delta(\theta - \theta_i) \, \delta(\phi - \phi_i) \,, \tag{6.3}$$

where $\delta(\cdot)$ denotes the $\delta-$Dirac function [45], the solution to Eq. (6.2) is expressed as:

$$\psi(\mathbf{x}) = \iint_S G(\mathbf{x}, \mathbf{x}') \, \omega(\mathbf{x}') \, dA \,, \tag{6.4}$$

where $G(\mathbf{x}, \mathbf{x}') = -(4\pi)^{-1} \ln \|\mathbf{x} - \mathbf{x}'\|^2$ represents the Green function of the Laplacian operator on a unit sphere, and $\| \cdot \|$ signifies the Euclidean norm in $\mathbb{R}^3$. Combining (6.1) with (6.4), one obtains the velocity field at location $\mathbf{x}$:

$$\mathbf{u}(\mathbf{x}) = \frac{1}{2\pi} \sum_{i=1}^{N} k_i \, \frac{\mathbf{x}_i \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_i\|^2} \,. \tag{6.5}$$

Let $\mathbf{x}_j = (x_j(t), y_j(t), z_j(t)) \equiv (x_j, y_j, z_j)$ denote the position of vortex j $(j = 1, 2, \cdots, N)$ on the sphere, where $||\mathbf{x}_j|| = R$. According to Eq. (6.5) the dynamics of this set of vortices is governed by the system of ordinary differential equations:

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \, \frac{\mathbf{x}_j \times \mathbf{x}_i}{\|\mathbf{x}_i - \mathbf{x}_j\|^2} \,, \qquad i = 1, 2, \ldots, N \,, \tag{6.6}$$

with the corresponding initial conditions. Here, k_j represents the circulation of vortex j, and N is the total number of vortices on the sphere. The distance along the chord between vortex i and j can be expressed as:

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2 \left(R^2 - \mathbf{x}_i \cdot \mathbf{x}_j \right) \,. \tag{6.7}$$

A passive particle, defined as a point vortex with circulation $k = 0$, contributes to the dynamics of a system, where P such particles are advected by N point vortices. This behaviour is governed by Eq. (6.6), supplemented by equations specifying the motion of

the passive particles:

$$\dot{\mathbf{x}}_p = \frac{1}{2\pi R} \sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_p}{\|\mathbf{x}_p - \mathbf{x}_j\|^2} \ , \qquad p = 1, 2, \ldots, P \, , \tag{6.8}$$

with the respective initial conditions.

Considering a single controlled passive particle ($P = 1$) moving in a spherical fluid induced by N vortices, the corresponding equation is

$$\dot{\mathbf{x}} = \frac{1}{2\pi R} \sum_{j=1}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_j\|^2} + \mathbf{U}^c \left(t\right) \ , \tag{6.9}$$

with the respective initial conditions. In the right-hand side of Eq. (6.9),

$$\mathbf{U}^c(t) = (u_x(t), u_y(t), u_z(t)) \equiv (u_x, u_y, u_z) \ , \tag{6.10}$$

is the control vector function. The interdependence among the entries of this control vector is evident. In the ensuing discussion, it will become clear that these entries must be configured in a manner to guarantee the particle's trajectory on the spherical surface.

Considering the single vortex located at the North Pole of the sphere, i.e., its position vector is

$$\mathbf{x}_1 = (0, 0, R) \, , \tag{6.11}$$

then

$$\mathbf{x}_1 \times \mathbf{x} = (0, 0, R) \times (x, y, z) = (-Ry, Rx, 0) \, , \tag{6.12}$$

and

$$\|\mathbf{x}_1 - \mathbf{x}\| = x^2 + y^2 + (R - z)^2, \tag{6.13}$$

which enables us to write row-wise the Eq. (6.9) for a single passive particle:

$$\begin{cases} \dot{x} = -\frac{k}{2\pi} \frac{y}{x^2+y^2+(R-z)^2} + u_x \\ \\ \dot{y} = \frac{k}{2\pi} \frac{x}{x^2+y^2+(R-z)^2} + u_y \\ \\ \dot{z} = u_z \end{cases} \tag{6.14}$$

The system (6.14), with the initial conditions ($\|\mathbf{x}(0)\| = \|(x(0), y(0), z(0))\| = R$):

$$\begin{cases} x(0) = x_0 \\ y(0) = y_0 \\ z(0) = z_0 \end{cases} \tag{6.15}$$

describes the dynamic of a controlled passive particle on a sphere advected by a single point vortex located at the North Pole.

The problem addressed here is the displacement of the passive particle between two given points of the sphere. Starting from the point P_0, the passive particle has to reach a final destination point P_f in a fixed time T. The autonomy of the particle in terms of displacement is given by the control $\mathbf{U}^c$. This control allows the particle to move in any direction on the surface of the sphere. Indeed, the particle should fully exploit the surface flow motion generated by the vortex, aiming to minimize the necessity for control input $\mathbf{U}^c$, since the application of control involves the expenditure of energy.

Since the particle $\mathbf{x}$ is constrained to move on the surface of the sphere with a radius of R , then

$$
\begin{aligned}
&\|\mathbf{x}\|^2 = R^2 \quad \Leftrightarrow \\
&\|(x,y,z)\|^2 = R^2 \quad \Leftrightarrow \\
&x^2+y^2+z^2 = R^2 \quad \Longrightarrow \\
&\frac{d}{dt}\left(x^2+y^2+z^2\right) = 0 \quad \Leftrightarrow \\
&2\left(x\dot{x}+y\dot{y}+z\dot{z}\right) = 0 \quad \Leftrightarrow \\
&x\dot{x}+y\dot{y}+z\dot{z} = 0
\end{aligned}
\tag{6.16}
$$

and, from (6.14), one obtains

$$
(u_x,u_y,u_z)\cdot(x,y,z) = 0\,. \tag{6.17}
$$

In other words, the control and position vectors are perpendicular to each other:

$$
(u_x,u_y,u_z) \perp (x,y,z)\ . \tag{6.18}
$$

This indicates that the control vector (u_x,u_y,u_z) lies within the tangent plane of the sphere at the point (x,y,z). In simpler terms, for the particle to move on the sphere, the exerted control must be orthogonal to the particle's position vector, i.e.,

$$
\mathbf{U}^c(t) \perp \mathbf{x}(t)\,, \quad \forall\, t \geq 0\,. \tag{6.19}
$$

The control vector can be defined as a linear combination of two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ belonging to tangent plane, i.e.,

$$
\mathbf{U}^c = \alpha\,\mathbf{w}_1 + \beta\,\mathbf{w}_2\,, \tag{6.20}
$$

with $\alpha \equiv \alpha(t)$ and $\beta \equiv \beta(t)$ two real (controls) scalars functions. Given two vectors in the tangent plane, $\mathbf{w}_1 = (y,-x,0)$ and $\mathbf{w}_2 = (0,z,-y)$, where $\mathbf{w}_1\cdot\mathbf{x} = 0$ and $\mathbf{w}_2\cdot\mathbf{x} = 0$, the

control vector indeed assumes the form

$$(u_x\,,u_y\,,u_z)=\alpha\,(y\,,-x\,,0)+\beta\,(0\,,z\,,-y)=(\alpha y\,,-\alpha x+\beta z\,,-\beta y)\,; \tag{6.21}$$

the dynamic of the passive particle is then given by

$$\begin{cases} \dot{x}=-\frac{k}{2\pi}\frac{y}{x^2+y^2+(R-z)^2}+\alpha y \\ \\ \dot{y}=\frac{k}{2\pi}\frac{x}{x^2+y^2+(R-z)^2}-\alpha x+\beta z \\ \\ \dot{z}=-\beta y \end{cases} \tag{6.22}$$

with initial conditions (6.15) and control

$$\mathbf{U}=(\alpha(t),\beta(t))\ . \tag{6.23}$$

The energy spent on the displacement aimed at minimization is represented by the integral over the time interval $[0,T]$ of the sum of the squares of the components of $\mathbf{U}$. Therefore, the objective function of the control problem is given by

$$\int_0^T \|\mathbf{U}\|^2\,dt=\int_0^T \left(\alpha^2(t)+\beta^2(t)\right)dt. \tag{6.24}$$

6.2 Modelling with Spherical Coordinates

In spherical coordinates, the dynamics of the passive particle under control in identical conditions, as detailed in the preceding section, is determined by (see Chapter 4 and [16]):

$$\begin{cases} \dot{\theta}=u_\theta \\ \dot{\phi}=\frac{k}{4\pi R^2}\,\frac{1}{1-\cos(\theta)}+u_\phi \end{cases} \tag{6.25}$$

with the given initial condition

$$\begin{cases} \theta(0) = \theta_0 \\ \phi(0) = \phi_0 \end{cases} \tag{6.26}$$

where $\theta \in [0, \pi]$ is the colatitude (or polar) angle of the particle position, i.e., the angle between the radius passing through the North Pole, $(0, 0, R)$, and the radius passing through the particle, and $\phi \in [0, 2\pi]$ is the longitude (or azimuthal) angle, i.e., the angle that the meridian passing through $(R, 0, 0)$ makes with the meridian passing by the particle position. The variables u_θ and u_ϕ represent the angular controls applied on the passive particle, i.e.,

$$\mathbf{U}^s = (u_\theta,\, u_\phi)\,, \tag{6.27}$$

and, as before, the circulation of the vortex located at the North Pole is given by k.

In Chapter 4, Eq. (6.25) was employed to solve the control problem of displacing the passive particle between two specified points of the superficial sphere within a fixed time. However, a direct comparison of results with those obtained using Cartesian coordinates is not feasible due to inherent differences in controls, given that the controls are independent of each other.

To understand the transformation of controls between Cartesian and spherical coordinates, we will derive the particle dynamics equations in spherical coordinates from their counterparts in Cartesian coordinates. This approach allows us to observe the changes in controls that occur during this transformation.

The correspondence between the position vectors of the passive particle in Cartesian coordinates (x, y, z) and spherical coordinates (θ, ϕ) is given by

$$\begin{cases} x = R\cos\phi\sin\theta \\ y = R\sin\phi\sin\theta \\ z = R\cos\theta \end{cases} \tag{6.28}$$

The time derivatives of these coordinates are:

$$\begin{cases} \dot{x} = -R\dot{\phi}\sin\phi\sin\theta + R\dot{\theta}\cos\phi\cos\theta \\ \\ \dot{y} = R\dot{\phi}\cos\phi\sin\theta + R\dot{\theta}\sin\phi\cos\theta \\ \\ \dot{z} = -R\dot{\theta}\sin\theta \end{cases} \tag{6.29}$$

The expression for the chord distance in spherical coordinates, as given by (6.13), can be written as

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}\| &= x^2 + y^2 + (R - z)^2 \\ &= x^2 + y^2 + z^2 + R^2 - 2zR \\ &= 2R^2 - 2R\,(R\cos\theta) \\ &= 2R^2\,(1 - \cos\theta) \end{aligned} \tag{6.30}$$

and the cross product (6.12) is modified to yield

$$\mathbf{x}_1 \times \mathbf{x} = (-Ry, Rx, 0) = \left(-R^2\cos\phi\sin\theta, R^2\cos\phi\sin\theta, 0\right). \tag{6.31}$$

Substituting (6.30) and (6.31) into Eq. (6.22) gives rise to

$$\begin{cases} \dot{x} = -\frac{k}{4\pi R}\frac{\sin\phi\sin\theta}{1-\cos\theta} + u_x \\ \\ \dot{y} = \frac{k}{4\pi R}\frac{\cos\phi\sin\theta}{1-\cos\theta} + u_y \\ \\ \dot{z} = u_z \end{cases} \tag{6.32}$$

Reciprocally, the controls specified in (6.21) can be transformed into Cartesian coordinates:

$$\begin{cases} u_x = \alpha y \\ u_y = -\alpha x + \beta z \\ u_z = -\beta y \end{cases} \Leftrightarrow \begin{cases} u_x = R\alpha \sin\phi \sin\theta \\ u_y = -R\alpha \cos\phi \sin\theta + R\beta \cos\theta \\ u_z = -R\beta \sin\phi \sin\theta \end{cases} \tag{6.33}$$

Replacing these controls in Eq. (6.32) leads to

$$\begin{cases} \dot{x} = -\frac{k}{4\pi R} \frac{\sin\phi \sin\theta}{1-\cos\theta} + R\alpha \sin\phi \sin\theta \\ \dot{y} = \frac{k}{4\pi R} \frac{\cos\phi \sin\theta}{1-\cos\theta} - R\alpha \cos\phi \sin\theta + R\beta \cos\theta \\ \dot{z} = -R\beta \sin\phi \sin\theta \end{cases} \tag{6.34}$$

The Eqs. (6.29) and (6.34) produce a system of equations that provides the basis for deriving the equations for $\dot{\theta}$ and $\dot{\phi}$. More precisely, the third equation of

$$\begin{cases} -R\dot{\phi} \sin\phi \sin\theta + R\dot{\theta} \cos\phi \cos\theta = -\frac{k}{4\pi R} \frac{\sin\phi \sin\theta}{1-\cos\theta} + R\alpha \sin\phi \sin\theta \\ R\dot{\phi} \cos\phi \sin\theta + R\dot{\theta} \sin\phi \cos\theta = \frac{k}{4\pi R} \frac{\cos\phi \sin\theta}{1-\cos\theta} - R\alpha \cos\phi \sin\theta + R\beta \cos\theta \\ -R\dot{\theta} \sin\theta = -R\beta \sin\phi \sin\theta \end{cases} \tag{6.35}$$

yields

$$\dot{\theta} = \beta \sin\phi \,, \tag{6.36}$$

while the first two equations of (6.35) provide us with

$$\begin{cases} \dot{\theta} = \beta \sin\phi \\ \dot{\phi} = \frac{k}{4\pi R^2} \frac{1}{1-\cos\theta} - \alpha + \beta \cos\phi \cot\theta \end{cases} \tag{6.37}$$

for which, when compared to Eq. (6.25), the result is

$$\mathbf{U}^s = (u_\theta,\, u_\phi) = (\beta \sin\phi,\, -\alpha + \beta \cos\phi \cot\theta)\ . \tag{6.38}$$

This expression allows for the computation of spherical controls $u_\theta(t)$ and $u_\phi(t)$ as functions of the same parameters $\alpha(t)$ and $\beta(t)$ utilized in Cartesian control.

6.3 Numerical Control for Particle Displacement

The control problem arising from controlling the particle's displacement on the surface of the sphere is numerically solved using a direct approach, akin to the one previously employed in the infinite plane [17]. Effectively, this control problem transforms into an optimization problem involving the minimization of the objective function, as defined by Eq. (6.24), while taking into account the constraint arising from the passive particle's displacement from an initial point P_0 to a target point P_f within precisely T units of time.

6.3.1 Optimization Problems

As explained in the preceding section, the dynamics of the passive particle can be modelled in various ways, contingent on the choice of coordinates and the type of control. In the scenario of Cartesian coordinates, the optimization problem is formulated as follows:

P1: Cartesian Optimization Problem	
Minimize :	$\int_0^T \|\mathbf{U}\,(t)\|^2\, dt$
subject to :	$\dot{\mathbf{x}} = \frac{k}{2\pi R} \frac{\mathbf{x}_1 \times \mathbf{x}}{\|\mathbf{x}-\mathbf{x}_1\|^2} + \mathbf{U}^c$
	$\mathbf{x}(0) = P_0$
	$\mathbf{x}(T) = P_f$
	$\|\mathbf{U}\| \leq u_{\max}$

where $\mathbf{U} = (\alpha, \beta)$, $P_0 = x_0$, and $P_f = x_f \in \mathbb{R}^3$ represent the initial and final position of the passive particles, $T > 0$ denotes the time available for the displacement, k is the circulation of the vortex $\mathbf{x} \in \mathbb{R}^3$, and $u_{\max} > 0$ is the upper limit value of the control.

In the generic optimization problem P1, the objective function to be minimized corresponds to the energy utilized for the displacement of the passive particle during the time interval $[0, T]$. The constraints are linked to the application of the equation governing particle dynamics, the specified initial and final positions, and the constraint on the maximum control norm. This last condition stems from the finite energy available for executing the displacement.

For spherical coordinates, the optimization problem can be formulated in the function of the controls $\mathbf{U}^s = (u_\theta, u_\phi)$ or $\mathbf{U} = (\alpha, \beta)$. The version corresponding to the first case is:

P2: Spherical Optimization Problem—Version 1

Minimize :

$$\int_0^T \|\mathbf{U}^s(t)\|^2 \, dt$$

subject to :

$$\dot{\mathbf{y}} = \left(u_\theta, \tfrac{k}{4\pi R^2} \tfrac{1}{1-\cos\theta} + u_\phi\right)$$
$$\mathbf{y}(0) = p_0$$
$$\mathbf{y}(T) = P_f$$
$$\|\mathbf{U}_s\| \leq u_{\max}$$

In the generic optimization problem P2, $\mathbf{y} = (\theta, \phi)$ is the position of the passive particle described by Eq. (6.25), $\mathbf{U}^s = (u_\theta, u_\phi)$ is the control, and $P_0 = (\theta_0, \phi_0)$ and $P_f = (\theta_T, \phi_T)$ are the initial and final point, respectively. The meaning of the objective function and constraints in the generic optimization problem P2 is identical to that in P1.

The version of the optimization problem in spherical coordinates formulated in function of the control $\mathbf{U} = (\alpha, \beta)$ is:

P3: Spherical Optimization Problem—Version 2

Minimize :

$$\int_0^T \|\mathbf{U}(t)\|^2 \, dt$$

subject to :

$$\dot{\mathbf{z}} = \left(\beta \sin\phi, \tfrac{k}{4\pi R^2} \tfrac{1}{1-\cos\theta} - \alpha + \beta\cos\phi\cot\theta\right)$$
$$\mathbf{z}(0) = P_0$$
$$\mathbf{z}(T) = P_f$$
$$\|\mathbf{U}\| \leq u_{\max}$$

In the optimization problem P3, $\mathbf{z} = (\theta, \phi)$ is the position of the passive particle described by Eq. (6.37), $\mathbf{U} = (\alpha, \beta)$ is the control, and $P_0 = (\theta_0, \phi_0)$ and $P_f = (\theta_T, \phi_T)$ are the initial and final point, respectively. The objective function and constraints of the generic optimization problem P3 have the same meaning as those in both P1 and P2.

6.3.2 Numerical Solutions by Direct Approach

To address the optimization problems, the generic control function $\mathbf{U}(\cdot)$ is discretized in n constant vector variables defined as

$$\begin{aligned} \mathbf{U}(t) &= \mathbf{u}_0 \quad \text{if} \quad t_0 \leq t < t_1\,, \\ \mathbf{U}(t) &= \mathbf{u}_1 \quad \text{if} \quad t_1 \leq t < t_2\,, \\ \mathbf{U}(t) &= \mathbf{u}_2 \quad \text{if} \quad t_2 \leq t < t_3\,, \\ &\vdots \\ \mathbf{U}(t) &= \mathbf{u}_{n-1} \quad \text{if} \quad t_{n-1} \leq t \leq t_n\,, \end{aligned} \tag{6.39}$$

with $t_0 = 0,\ t_n = T$. Each vector variable $\mathbf{u}_i$, $i = 1, 2, \cdots, n$, corresponds to a constant value of the control function exercised in the sub-interval $[t_{i-1}, t_i)$ with constant length $\Delta t = (t_n - t_0)/n$.

Therefore, the discretization of the objective function (6.24) in n time intervals lead to the approximation

$$\int_0^T \|\mathbf{U}\|^2\, dt \approx \Delta t\, \left(\|\mathbf{u}_0\|^2 + \|\mathbf{u}_1\|^2 + \ldots + \|\mathbf{u}_{n-1}\|^2\right)\,, \tag{6.40}$$

which represents the discretization of the energies function using the rule of rectangles.

The Cartesian optimization problem P1 is then replaced by its discretized version:

DP1: Discretized Cartesian Optimization Problem

Minimize :

$$f_n = \Delta t \sum_{i=0}^{n-1} \|\mathbf{u}_i\|^2$$

Subject to :

$$\dot{\mathbf{x}} = \tfrac{k}{2\pi R}\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_0^c, \quad \mathbf{x}(t_0) = x_0, \quad \|\mathbf{u}_0\| \le u_{\max}, \quad t_0 \le t < t_1$$
$$\dot{\mathbf{x}} = \tfrac{k}{2\pi R}\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_1^c, \quad \mathbf{x}(t_1) = x_1, \quad \|\mathbf{u}_1\| \le u_{\max}, \quad t_1 \le t < t_2$$
$$\vdots$$
$$\dot{\mathbf{x}} = \tfrac{k}{2\pi R}\tfrac{\mathbf{x}_1\times\mathbf{x}}{||\mathbf{x}-\mathbf{x}_1||^2} + \mathbf{u}_{n-1}^c, \quad \mathbf{x}(t_{n-1}) = x_{n-1}, \quad \|\mathbf{u}_{n-1}\| \le u_{\max}, \quad t_{n-1} \le t < t_n$$
$$\mathbf{x}(t_n) = x_f$$

Here, $\mathbf{u}_i^c = (\alpha_i y, -\alpha_i x + \beta_i z, -\beta_i y)$, $\mathbf{u}_i = (\alpha_i, \beta_i)$, $i = 0, 1, \ldots, n-1$ are the control variables, $x_0 = P_0$, $x_f = P_f$ are the initial and final point, respectively, and $t_n = T$.

The discretization approach for the spherical optimization problem is analogous. The only difference is that the dynamics of the passive particle in each sub-interval are governed by Eq. (6.25) in the case of Version 1, and by Eq. (6.37) in the case of Version 2.

The discretized form of Version 1 of the spherical optimization problem P2 is:

DP2: Discretized Spherical Optimization Problem—Version 1

Minimize :

$$f_n = \Delta t \sum_{i=0}^{n-1} \|\mathbf{u}_i^s\|^2$$

Subject to :

$$\dot{\mathbf{y}} = \left(0, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta}\right) + \mathbf{u}_0^s, \quad \mathbf{y}(t_0) = y_0, \quad \|\mathbf{u}_0^s\| \le u_{\max}, \quad t_0 \le t < t_1$$
$$\dot{\mathbf{y}} = \left(0, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta}\right) + \mathbf{u}_1^s, \quad \mathbf{y}(t_1) = y_1, \quad \|\mathbf{u}_1^s\| \le u_{\max}, \quad t_1 \le t < t_2$$
$$\vdots$$
$$\dot{\mathbf{y}} = \left(0, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta}\right) + \mathbf{u}_{n-1}^s, \quad \mathbf{y}(t_{n-1}) = y_{n-1}, \quad \left\|\mathbf{u}_{n-1}^s\right\| \le u_{\max}, \quad t_{n-1} \le t < t_n$$
$$\mathbf{y}(t_n) = y_f$$

In the discretized spherical optimization Problem DP2, $\mathbf{u}_i^s = (u_{\theta_i}, u_{\phi_i})$ are the control variables, $i = 0, 1, \ldots, n-1$, $y_0 = P_0$ and $y_f = P_f$ are the initial and final point, respectively, and $t_n = T$.

For Version 2 of the spherical optimization problem, the discretized optimization problem is:

DP3: Discretized Spherical Optimization Problem—Version 2

Minimize:

$$f_n = \Delta t \sum_{i=0}^{n-1} \|\mathbf{u}_i\|^2$$

Subject to:

$$\dot{\mathbf{z}} = \left(\beta_0 \sin\phi, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta} - \alpha_0 + \beta_0 \cos\phi \cot\theta\right), \ \mathbf{z}(t_0) = z_0, \ \|\mathbf{u}_0\| \le u_{\max}, \ \ t_0 \le t < t_1$$

$$\dot{\mathbf{z}} = \left(\beta_1 \sin\phi, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta} - \alpha_1 + \beta_1 \cos\phi \cot\theta\right), \ \mathbf{z}(t_1) = z_1, \ \|\mathbf{u}_1\| \le u_{\max}, \ \ t_1 \le t < t_2$$

$$\vdots$$

$$\dot{\mathbf{z}} = \left(\beta_{n-1} \sin\phi, \tfrac{k}{4\pi R^2}\tfrac{1}{1-\cos\theta} - \alpha_{n-1} + \beta_{n-1} \cos\phi \cot\theta\right), \ \ \mathbf{z}(t_{n-1}) = z_{n-1},$$
$$\|\mathbf{u}_{n-1}\| \le u_{\max}, \ \ t_{n-1} \le t < t_n$$

$$\mathbf{z}(t_n) = z_f$$

In the discretized spherical optimization Problem DP3, $\mathbf{u}_i = (\alpha_i, \beta_i)$ are the control variables, $i = 0, 1, \ldots, n-1$, $z_0 = P_0$ and $z_f = P_f$ are the initial and final point, respectively, and $t_n = T$.

To obtain the final position $\mathbf{x}(T)$ of the passive particle, the ordinary differential equation that models the dynamic of the passive particle is solved numerically in cascade by the fourth-order Runge–Kutta method. The initial position of the passive particle in each sub-interval is given by the final position in the previous one. The main restriction of the discretized optimization problem consists of reaching the targeted destination P_f in a given time T, imposed in the last sub-interval by the condition $\mathbf{x}(t_n) = P_f$. Because achieving an exact value is challenging in finite precision computation, the condition $\mathbf{x}(t_n) = P_f$ is replaced by

$$\|\mathbf{x}(T) - P_f\| < \varepsilon\,, \tag{6.41}$$

where ε denotes a predefined small tolerance. In spherical optimization problems, the equivalent condition is applied by substituting $\mathbf{x}(T)$ with $\mathbf{y}(T)$ in the case of Version 1, or with $\mathbf{z}(T)$ in the case of Version 2, in condition (6.41).

The discretized optimization problems are solved numerically by mean of the MATLAB nonlinear optimization solver `fmincon` [77] that provides the constrained optimization algorithms Interior Point and Active-Set (see, for instance, [123, 96]).

6.4 Experimental Results

The results presented in this section are related to the trajectory of a passive particle. The particle is required to move from an initial point $P_0 = \left(\frac{\pi}{6}, \pi\right)$ to a designated target point $P_f = \left(\frac{3\pi}{4}, 0\right)$ on the surface of a sphere – locations in spherical coordinates. For the sake of simplicity, the radius of the sphere is fixed at $R = 1$.

It is considered that the passive particle has reached the destination when its position $\mathbf{x}(T)$ is within a distance of $\varepsilon = 10^{-3}$ from P_f, as indicated in Eq. (6.41). The passive particles move in a flow that is induced by $N = 1, 2, 3$ point vortex with circulations $k_i = 1$, for $i = 1, \ldots, N$.

The discrete optimization problems (DP1, DP2, and DP3), previously outlined, have been computationally implemented using the MATLAB Optimization Toolbox [77]. The optimization tool `fmincon`, employed for nonlinear optimization, is utilized to identify optimal controls. Given that the solution is sensitive to the initial condition, which is randomly generated, multiple executions of this built-in function may yield diverse results. To mitigate the impact of the initial guess, the function is executed multiple times, and the solution leading to a lower value of the objective function (6.40) is chosen. For this purpose, the built-in function `multistart` is applied with twenty repetitions.

The nonlinear restrictions corresponding to the displacement of the passive particle in each sub-interval involve solving ordinary differential equations that govern both its dynamics and the dynamics of the vortices. These equations are numerically integrated using the `ode45` built-in function, which implements the fourth and fifth-order Runge–Kutta methods.

The computational runs were conducted on a KVM virtual machine equipped with 16 vCores of an Intel Xeon W-2195 CPU, 64 GB of RAM, and an SSD-based virtual disk, operating on Linux Ubuntu 20.04 LTS. It is important to note that the computing times provided in the tables below are indicative, as they are significantly influenced by the virtual machine's usage by other users.

Problems DP1, DP2, and DP3 are solved for the cases $N = 1, 2$, and 3 vortices. For the case of $N = 1$, the equations that govern the dynamics are those introduced in the previous section. For the case of $N = 2$ and $N = 3$, the equations that govern the dynamics of the passive particle and the vortices will be introduced.

6.4.1 Flow Created by a Single Vortex (N=1)

Table 6.1 displays the results from solving the discrete optimization problems DP1, DP2, and DP3 with $n = 3$ controls in the case of a flow induced by a single point vortex (N=1). It is noteworthy that the solutions for DP1 and DP3 produce identical values for both the controls and the objective function. This confirms the validity of the numerical algorithm based on the Discretization (6.39), which is employed for both problems DP1 and DP3. Due to their equivalence, given that one is derived from the other through a coordinate transformation, the algorithm is anticipated to yield consistent results for both cases.

The computational time is shorter when solving the DP1 problem. The DP2 problem takes about four times as long to resolve compared to the other two cases.

Table 6.1: Solution of the discrete optimization problems ($N = 1$) with $n = 3$ controls.

Problem	**Controls**	f_3	**CPU Time**
DP1	$\mathbf{u}_0 = (-0.189, -0.443)$ $\mathbf{u}_1 = (-0.225, -0.529)$ $\mathbf{u}_2 = (-0.344, -0.468)$	1.50	41.0
DP2	$\mathbf{u}_0^s = (0.695, -0.096)$ $\mathbf{u}_1^s = (0.674, -0.096)$ $\mathbf{u}_2^s = (0.673, -0.096)$	2.36	208.2
DP3	$\mathbf{u}_0 = (-0.189, -0.443)$ $\mathbf{u}_1 = (-0.225, -0.529)$ $\mathbf{u}_2 = (-0.344, -0.468)$	1.50	60.3

The trajectories corresponding to the solutions presented in Table 6.1 are illustrated in Figure 6.1. Trajectories resulting from the resolution of problems DP1 and DP3 are equal, as they both reach the final point by moving on the sphere through the same path. The trajectory resulting from DP2 leads to a longer path, leading to a larger value of the objective function. However, this value cannot be directly comparable with the objective function values resulting from the two other problems because the control parameters are different. It can be also seen that DP1 and DP3 present the same solution because they minimize the same objective function given by Eq. (6.40). The optimization algorithm found the same minimum for the two problems.

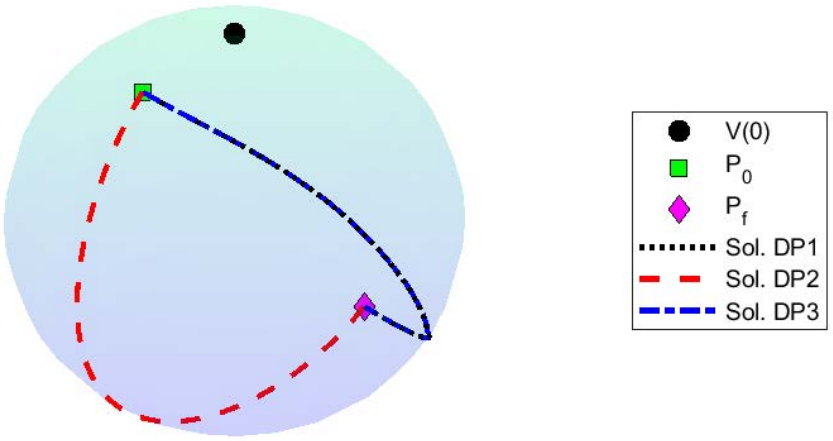


Figure 6.1: Trajectories resulting from the solutions of the optimization problems DP1, DP2, and DP3 with a set of $n = 3$ controls ($N = 1$).

Table 6.2: Discrete problem solutions in a flow generated by a single vortex ($N = 1$).

	DP1		**DP2**		**DP3**	
n	f_n	**CPUt [s]**	f_n	**CPUt [s]**	f_n	**CPUt [s]**
1	1.53	21.2	1.84	265.5	1.53	25.9
2	1.51	25.8	1.65	371.9	1.51	43.3
4	1.50	24.7	1.30	385.8	1.50	87.0
6	1.49	16.3	1.30	188.3	1.49	127.4
8	1.49	17.7	1.29	575.1	1.49	181.3
10	1.49	25.5	1.25	185.5	1.49	181.5
12	1.49	55.9	1.23	728.1	1.49	181.5

Table 6.2 displays the results obtained as a function of the number of controls n. As the number of control variables increases, the objective function shows a decreasing trend. The solution of DP1 and DP3 reaches a plateau after $n = 6$. This indicates that the incremental gains in the objective function become marginal, remaining below a few hundred units. Notably, DP1 and DP3 consistently yield the same values for the objective function. Regarding computing time, there is a prevailing trend of an increase with the number, n, of control variables, but with many irregularities. The computation times for solving DP1 are typically shorter compared to the other cases. Moreover, the times required for DP3 are lower than those observed for DP2.

6.4.2 Flow Created by Two Vortices (N=2)

In this section, the flow on the surface of the sphere is generated by two vortices. Considering Cartesian coordinates, the position of the two vortices, $\mathbf{x}_1$ and $\mathbf{x}_2$, are described by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{k_2}{2\,\pi\,R}\,\frac{\mathbf{x}_2\times\mathbf{x}_1}{\|\mathbf{x}_1-\mathbf{x}_2\|^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{k_1}{2\,\pi\,R}\,\frac{\mathbf{x}_1\times\mathbf{x}_2}{\|\mathbf{x}_2-\mathbf{x}_1\|^2} \end{cases} \tag{6.42}$$

with the respective initial conditions where the initial position of the vortices are $V_{10} = (-0.3090, 0, 0.9511)$ and $V_{20} = (0.8660, 0, -0.5000)$, and the passive particle, initially at $\mathbf{x}(0) = P_0$, is modeled by the equation

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1\times\mathbf{x}}{\|\mathbf{x}-\mathbf{x}_1\|^2} + k_2\,\frac{\mathbf{x}_2\times\mathbf{x}}{\|\mathbf{x}-\mathbf{x}_2\|^2}\right) + \alpha\;(y,-x,0) + \beta\;(0,z,-y)\,, \tag{6.43}$$

for $i = 0, 1, \ldots, n-1$, where α and β are the controllers, as presented in Section 6.1.

The determination of the optimal trajectory of the passive particle is obtained by solving a discrete optimization problem similar to DP1, where Eq. (6.22) is replaced by Eqs. (6.42) and (6.43).

In spherical coordinates, the dynamics of the two vortices $\mathbf{x}_1 = (\theta_1, \phi_1)$ and $\mathbf{x}_2 = (\theta_2, \phi_2)$ are given by

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi R^2}\,\frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \\ \dot{\phi}_1 = -\frac{k_2}{4\pi R^2}\,\frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{\sin(\theta_1)(1-\cos(\gamma_{12}))} \\ \\ \dot{\theta}_2 = -\frac{k_1}{4\pi R^2}\,\frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \\ \dot{\phi}_2 = -\frac{k_1}{4\pi R^2}\,\frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{\sin(\theta_2)(1-\cos(\gamma_{21}))} \end{cases} \tag{6.44}$$

with the given initial conditions of the vortices $V_{10} = (\pi/10,\, \pi)$ and $V_{20} = (2\pi/3,\, 0)$. In Eq. (6.44), the value of $\cos(\gamma_{12})$ and $\cos(\gamma_{21})$ are computed in agreement with

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,. \tag{6.45}$$

The motion of the passive particle is significantly influenced by the interaction between the two vortices and by the controls,

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi R^2}\left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\theta \\ \\ \dot{\phi}_p = \frac{1}{4\pi R^2}\left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})}\right) + u_\phi \end{cases} \tag{6.46}$$

with the initial conditions $P_0 = (\theta_{p0}, \phi_{p0})$. As in the case of a single vortex, the control applied to the passive particle is represented by the vector $\mathbf{U}^s = (u_\theta \; u_\phi)$. The determination of the optimal trajectory of the passive particle using spherical coordinates (Version 1) is obtained by solving a discrete optimization problem similar to DP2, where Eq. (6.25) is replaced by Eq. (6.44) together with Eq. (6.46).

The spherical controls $\mathbf{U}^s = (u_\theta,\ u_\phi)$ can be expressed in terms of the Cartesian controls α and β. Following an approach similar to that employed in Section 6.1, an analogous relationship between these two types of controls is derived:

$$\begin{cases} u_\theta = \beta \sin\phi \\ \\ u_\phi = -\alpha + \beta\cos\phi\cot\theta \end{cases} \tag{6.47}$$

By employing (6.47) in Eq. (6.46), the model of the passive particle's dynamics in a flow induced by two vortices with spherical coordinates is obtained through the use of Cartesian control (Version 2). The solution of this equation together with Eq. (6.44) enables us to determine the optimal trajectory of the passive particle.

Table 6.3 presents the results obtained from solving the discretised optimization problems DP1, DP2, and DP3 in the case of a flow induced by two vortices. As in the case of a single vortex, with two point vortices DP1 and DP3 also achieve the same values of the objective function of all number n of controls. In all three problems, the objective function values decrease as the number of controls increases. The computation times

Table 6.3: Discrete problem solutions in a flow generated by two vortices ($N = 2$).

	DP1		**DP2**		**DP3**	
n	f_n	**CPUt [s]**	f_n	**CPUt [s]**	f_n	**CPUt [s]**
1	1.35	181.5	1.37	106.2	1.35	41.8
2	1.27	181.8	1.33	225.4	1.27	101.6
4	1.27	181.7	1.25	112.4	1.27	101.2
6	1.26	181.8	1.23	101.7	1.26	101.2
8	1.26	183.3	1.36	101.8	1.26	101.4
10	1.25	182.8	1.66	101.0	1.25	101.8
12	1.25	182.4	1.58	207.2	1.25	101.5

exhibit considerable constancy with the variation of n. Generally, the computation times for DP3 are lower than those for DP2 and lower than those for DP1.

Figure 6.2 illustrates the trajectories corresponding to the minimum values of the objective function obtained in the resolution of DP1, DP2, and DP3 (see Table 6.3) in the case of a flow induced by two vortices. Notably, the trajectory resulting from DP1 is the same as the one resulting from DP3. The trajectory resulting from DP2 differs from the other two, being more extensive.

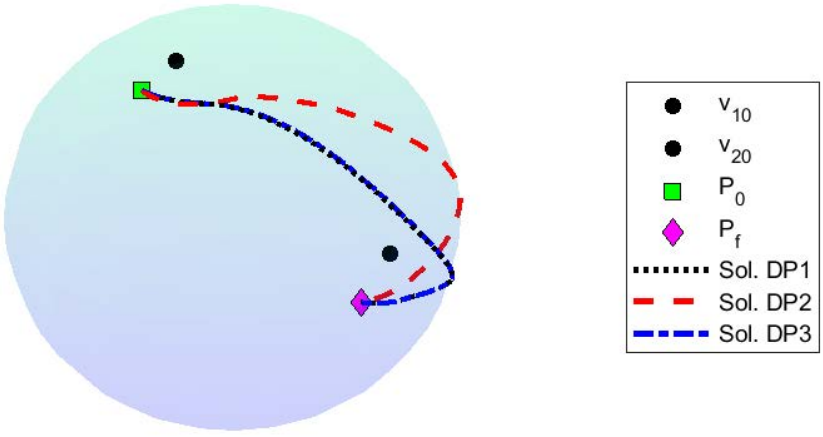


Figure 6.2: Trajectories corresponding to the minimum values of the objective function in the case of a flow induced by two vortices ($N = 2$).

6.4.3 Flow Created by Three Vortices (N=3)

This section focuses on a flow generated by three point vortices. Initially, we present the equations governing the displacement of a passive particle, introducing them in both Cartesian and spherical coordinates.

In Cartesian coordinates, the positions of the three vortices, $\mathbf{x}_1(t)$, $\mathbf{x}_2(t)$, and $\mathbf{x}_3(t)$, are governed by the following system of equations:

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{1}{2\,\pi\,R}\left(k_2 \frac{\mathbf{x}_2 \times \mathbf{x}_1}{||\mathbf{x}_1 - \mathbf{x}_2||^2} + k_3 \frac{\mathbf{x}_3 \times \mathbf{x}_1}{||\mathbf{x}_1 - \mathbf{x}_3||^2}\right) \\ \dot{\mathbf{x}}_2 = \frac{1}{2\,\pi\,R}\left(k_1 \frac{\mathbf{x}_1 \times \mathbf{x}_2}{||\mathbf{x}_2 - \mathbf{x}_1||^2} + k_3 \frac{\mathbf{x}_3 \times \mathbf{x}_2}{||\mathbf{x}_2 - \mathbf{x}_3||^2}\right) \\ \dot{\mathbf{x}}_3 = \frac{1}{2\,\pi\,R}\left(k_1 \frac{\mathbf{x}_1 \times \mathbf{x}_3}{||\mathbf{x}_3 - \mathbf{x}_1||^2} + k_2 \frac{\mathbf{x}_2 \times \mathbf{x}_3}{||\mathbf{x}_3 - \mathbf{x}_2||^2}\right) \end{cases} \tag{6.48}$$

with the respective initial conditions for the vortices $V_{10} = (-0.3090, 0, 0.9511)$, $V_{20} = (0.8660, 0, -0.5000)$, and $V_{30} = (0, 0, -1)$. The constants k_1, k_2, and k_3 represent the circulation of each vortex.

The dynamics of the passive particle is governed by the equation

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R} \sum_{i=1}^{3} k_i \, \frac{\mathbf{x}_i \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_i\|^2} + \alpha\,(y, -x, 0) + \beta\,(0, z, -y)\,, \tag{6.49}$$

with the given initial condition $\mathbf{x}(0) = P_0$.

In spherical coordinates, the dynamics of the point vortices positions $\mathbf{x}_1(t) = (\theta_1, \phi_1)$, $\mathbf{x}_2(t) = (\theta_2, \phi_2)$, and $\mathbf{x}_3(t) = (\theta_3, \phi_3)$ are given by

$$\begin{cases} \dot{\theta}_1 = -\frac{1}{4\pi R^2}\left(k_2\frac{\rho_{12}}{1-\cos(\gamma_{12})} + k_3\frac{\rho_{13}}{1-\cos(\gamma_{13})}\right) \\ \dot{\phi}_1 = \frac{1}{4\pi R^2}\left(k_2\frac{\sigma_{12}}{\sin(\theta_1)(1-\cos(\gamma_{12}))} + k_3\frac{\sigma_{13}}{\sin(\theta_1)(1-\cos(\gamma_{13}))}\right) \\ \dot{\theta}_2 = -\frac{1}{4\pi R^2}\left(k_1\frac{\rho_{21}}{1-\cos(\gamma_{21})} + k_3\frac{\rho_{23}}{1-\cos(\gamma_{23})}\right) \\ \dot{\phi}_2 = \frac{1}{4\pi R^2}\left(k_1\frac{\sigma_{21}}{\sin(\theta_2)(1-\cos(\gamma_{21}))} + k_3\frac{\sigma_{23}}{\sin(\theta_2)(1-\cos(\gamma_{23}))}\right) \\ \dot{\theta}_3 = -\frac{1}{4\pi R^2}\left(k_1\frac{\rho_{31}}{1-\cos(\gamma_{31})} + k_2\frac{\rho_{32}}{1-\cos(\gamma_{32})}\right) \\ \dot{\phi}_3 = \frac{1}{4\pi R^2}\left(k_1\frac{\sigma_{31}}{\sin(\theta_3)(1-\cos(\gamma_{31}))} + k_2\frac{\sigma_{32}}{\sin(\theta_3)(1-\cos(\gamma_{32}))}\right) \end{cases} \tag{6.50}$$

with the initial positions $V_{10} = (\pi/10,\, \pi)$, $V_{20} = (2\pi/3,\, 0)$, and $V_{30} = (\pi,\, \pi)$.

In Eq. (6.50), the values of γ_{ij} are given, as before, by (6.45), and ρ_{ij} and σ_{ij} are given by the Bogomolov notation [32, 66]:

$$\rho_{ij} = \sin(\theta_j)\sin(\phi_i - \phi_j)\,, \tag{6.51}$$

and

$$\sigma_{ij} = \sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)\,. \tag{6.52}$$

Hence, the dynamics of the controlled passive particle in a flow created by the three vortices is expressed as:

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi R^2}\left(k_1\frac{\rho_{p1}}{1-\cos(\gamma_{p1})} + k_2\frac{\rho_{p2}}{1-\cos(\gamma_{p2})} + k_3\frac{\rho_{p3}}{1-\cos(\gamma_{p3})}\right) + u_\theta \\ \dot{\phi}_p = \frac{1}{4\pi R^2}\left(k_1\frac{\sigma_{p1}}{\sin(\theta_p)(1-\cos(\gamma_{p1}))} + k_2\frac{\sigma_{p2}}{\sin(\theta_p)(1-\cos(\gamma_{p2}))} + k_3\frac{\sigma_{p3}}{\sin(\theta_p)(1-\cos(\gamma_{p3}))}\right) + u_\phi \end{cases} \tag{6.53}$$

with the given initial conditions $P_0 = (\theta_{p0},\, \phi_{p0})$. As in the previous cases, the variables u_θ and u_ϕ represent the angular controls applied on the passive particle.

Table 6.4 showcases the results obtained from solving the discrete optimization problems DP1, DP2, and DP3 for the spherical flow created by three vortices. In the resolution of DP1, the dynamics of the passive particle are determined by solving Eqs. (6.48) and (6.49). Similarly, for DP2, the dynamics of the passive particle are governed by resolving Eqs. (6.50) and (6.53). In DP3, analogous to the one and two vortex cases, the equations governing the displacement of passive particles are Eqs.(6.50) and (6.53). Here, the spherical controls u_θ and u_ϕ are substituted with Cartesian controls α and β, as indicated in Eq. (6.38).

Table 6.4: Discrete problem solutions in a flow generated by three vortices ($N = 3$).

	DP1		**DP2**		**DP3**	
$\boldsymbol{n}$	$\boldsymbol{f_n}$	**CPUt [s]**	$\boldsymbol{f_n}$	**CPUt [s]**	$\boldsymbol{f_n}$	**CPUt [s]**
1	1.46	181.3	2.13	105.9	1.46	101.3
2	1.37	181.3	1.90	101.1	1.37	101.8
4	1.35	181.4	1.68	101.6	1.35	101.4
6	1.34	181.9	1.68	101.5	1.34	101.9
8	1.34	182.5	1.62	102.5	1.34	101.9
10	1.33	182.4	1.62	104.9	1.33	103.1
12	1.33	183.9	1.59	104.8	1.33	103.1

In this scenario, a consistent pattern is observed in the gradual decrease of objective function values as the number of controls increases for all three problems. As before, DP1 and DP3 exhibit the same values of the objective function that attains the lowest value (1.33) for $n \geq 10$. DP2 attains the smallest value (1.59) for $n = 12$. Generally, computation times slightly increase with the number n of controls, with DP2 and DP3 achieving lower times than DP1.

Figure 6.3 depicts the trajectories corresponding to the minimum values of the objective function achieved in the solutions of DP1, DP2, and DP3 (refer to Table 6.4) for a flow induced by three vortices. As expected, the trajectories from DP1 and DP3 are the same. It is also observed that the trajectory resulting from DP2 is close to the other two.

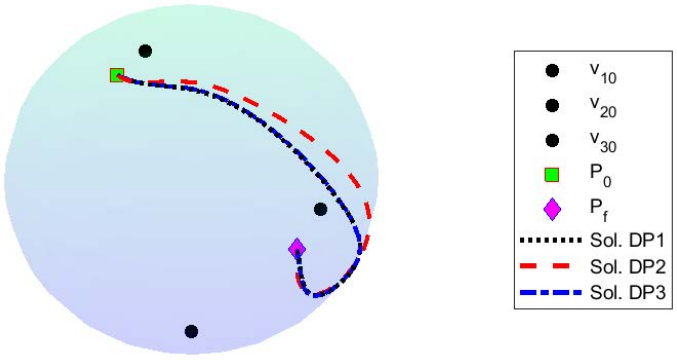


Figure 6.3: Trajectories corresponding to the minimum values of the objective function in the case of a flow induced by three vortices ($N = 3$).

6.5 Concluding Remarks

The results discussed in this chapter originate from an exploration into the modelling of surface displacement for a controlled passive particle. This displacement takes place in a flow generated by a set of point vortices, ranging from one to three, on the surface of a sphere.

The equations governing the dynamics of the passive particle are presented in both Cartesian and spherical coordinates. For each coordinate system, the task of determining the particle's trajectory between two points is translated into an optimization problem aimed at minimizing overall control. In the case of spherical coordinates, two forms of controls were applied. Initially, the problem was conceptualized with angular controls, and subsequently, the control parameters from the Cartesian formulation were employed in a second approach.

The three optimization problems arising from this study were solved using a direct method. This method involves dividing the available time for displacement into a predetermined number of sub-intervals. Throughout each sub-interval, the controls are constant.

The results obtained from resolving the three discretized problems suggest that each formulation produces a feasible solution. In simpler terms, each approach allows the passive particle to attain the desired displacement within the allocated time. These two versions of the same problem affirm the effectiveness of the algorithm used to convert the optimal control problem into an optimization problem, as demonstrated by the obtained controls.

The numerical resolution of these formulations through a direct approach consistently yields optimal solutions, irrespective of the chosen coordinate system.

Generally, as the number of controls increases, the objective function's value decreases. However, beyond a certain threshold, the objective function ceases to decrease and remains constant.

Computation times generally rise, when addressing the problem using spherical coordinates and angular controls. In general, the most efficient computational times are achieved through either the Cartesian formulation or the spherical formulation with Cartesian controls.

Chapter 7

Application to Conceptual Autonomous Technology

The results from this chapter have already been published as:
Balsa, C., Otero-Espinar, V., Gama, S. (2024). *An Approach to Environmental Cleanup Through Conceptual Autonomous Technology.* In: Gonçalves, J.A.d.C., Lima, J.L.S.d.M., Coelho, J.P., García-Peñalvo, F.J., García-Holgado, A. (eds) Proceedings of TEEM 2023. TEEM 2023. Lecture Notes in Educational Technology. Springer, Singapore. ISBN 978-981-97-1813-9. ISSN 2196-4963. https://doi.org/10.1007/978-981-97-1814-6_57.

This chapter centres around a conceptual mathematical toy model that describes the behaviour of autonomous technology (ATs) responsible for the garbage collection done by autonomous vehicles (AVs) in oceanic settings. This simplified mathematical model combines conceptual insights into the physical world with the symbolic structures of mathematics, thus serving as a valuable pedagogic instrument to engage students and emphasise the significance of mathematics and physics in addressing environmental concerns. Atmospheric or oceanic circulation is modelled using point vortices on a sphere, and ATs are modelled through optimal control of the trajectory of passive particles.

Mathematical toy models are highly valuable for mechatronics students and researchers, forming math-physics connections and developing physical intuition [99]. On the other hand, as these simplified models often satisfactorily describe the evolution of correspon-

ding complex physical systems, their computational implementation constitutes a low-cost alternative to developing complex computational fluid dynamics (CFD) models or to developing physical prototypes [8].

Point vortices simplify and approximate specific geophysical flows that exhibit long-term persistence, influenced by the Earth's curvature [120]. As point vortex models provide rapid answers to numerous questions concerning the fundamental dynamics of atmospheric flows [94], they are also employed to detect and assess physical phenomena derived from the interaction of atmospheric and oceanic vortices [84].

The relevance of these point vortices stems from their ability to provide a simplified approximation of specific geophysical flows. Point vortex models offer answers to numerous studies concerning the fundamental dynamics of atmospheric flows [94]. Conceptual models involving point vortices are employed to detect and assess physical phenomena that impact the structure and interaction of atmospheric and oceanic vortices [84]. Moreover, point vortices have recently been used in modelling pesticide dispersion in agricultural problems [131, 67].

The development of ATs, such as gliders, drones, or land vehicles, to solve environmental problems is a current priority. Ocean gliders are underwater AVs that are used for the collection of data or in remote locations. These little robots are uniquely suited for ocean sciences since they operate safely and at relatively low cost [91]. Another kind of AVs used in ocean studies is the Saildrone. This vehicle is an autonomous sailing drone currently used to provide high-quality oceanic and atmospheric observations, namely about the hurricane season [100].

In the present study, virtual AVs collect trash clusters accumulated in a given ocean region. The accumulation of plastics in the oceans is causing a growing environmental crisis for which immediate solutions must be found. The amount of plastic waste that ends up in the oceans each year is estimated at 8 million tons [62]. This massive pollution harms marine ecosystems and damages the most fragile habitats such as coral reefs [71]. According to [101], the absorption of microplastics by aquatic species, used in the production of agricultural fertilisers (fish, algae, etc.) [49], also represents a danger to human health. The Ocean Cleanup Project (OCP) is currently developing actions aimed at ridding the oceans of plastic waste through innovative technologies [116].

An action similar to the one undertaken by the OCP is modelled in the present chapter,

which is based on [23]. It is assumed to be an ideal scenario where AVs collect garbage. These vehicles are sufficiently small that they do not alter the local structure of the velocity field of the flow. These AVs are mathematically described as passive particles, and the ocean current, where the AVs move, is generated by the point vortices in a sphere.

Once the collection capacity of these vehicles is saturated, they move to a specific location where they deposit the collected material. The motion of these vehicles uses the ocean current as the primary source of displacement. But it also has a system of constant controls by parts. These controls consist of several predefined switch points allowing ATs to change their trajectory. As activating the controls has energy costs, the passive particles (AT representatives) need to optimise their displacement to minimise the sum of squares of the amplitude of the controls. Therefore, the focus of this work is the determination of the ideal relocation of the collected material between two fixed regions of the sphere, by minimising the energy expended on moving each AV within a predetermined time frame required for the relocation.

Similar to previous chapters (see also [18, 16, 19]), the resulting optimal control problem is solved by a direct numerical approach that consists of dividing the available time T for the displacement into n sub-intervals, where the controls remain constant. This approach leads to a non-linear programming problem that is solved using the built-in function `fmincon` from the MATLAB Optimisation Toolbox [77]. This function allows one to employ two optimisation methods, the Interior Point or the Active Set [34].

The structure of this chapter is as follows: Section 7.1 presents the simplified mathematical model for ATs. In Section 7.2, the article examines computational simulations using the model, specifically focusing on the cases of ocean currents induced by one and two point vortices. Finally, Section 7.3 is devoted to some final thoughts and considerations.

7.1 Dynamics in the Ocean by Point Vortices

This section describes the mathematical framework employed for constructing a simplified model of the dynamics of autonomous vehicles (AVs) moving on the ocean's surface. The AVs travel within oceanic (or atmospheric) currents. In this study, these currents are generated by point vortices.

The position of a vortex i on a sphere, centred in the origin with radius R, is given by the vector $\mathbf{x}_i$ that points from the centre of the sphere to the vortex location $\mathbf{x}_i = (x_i, y_i, z_i)$ on the spherical surface $\|\mathbf{x}\| = R$. The dynamics of the vortices satisfy

$$\dot{\mathbf{x}}_i = \frac{1}{2\pi R} \sum_{\substack{j=1 \\ j \neq i}}^{N} k_j \frac{\mathbf{x}_j \times \mathbf{x}_i}{\|\mathbf{x}_i - \mathbf{x}_j\|^2}\,, \qquad i = 1, 2, \ldots, N\,, \tag{7.1}$$

with the respective initial conditions, where k_j represents the circulation of vortex j, and N denotes the total number of point vortices on the sphere.

A passive particle is, by definition, a point vortex with circulation $k = 0$. The dynamics of a system with a single passive particle advected by N point vortices is described by Eq. (7.1) together with the equation for the passive particle

$$\dot{\mathbf{x}} = \frac{1}{2\pi R} \sum_{i=1}^{N} k_i \frac{\mathbf{x}_i \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_i\|^2}\,, \tag{7.2}$$

supplemented with the initial condition $\mathbf{x}(0)$, which denotes the position of a trash's element.

7.1.1 Autonomous Technology as Passive Particles

The AT idealised in this study moves on the earth's surface (ocean or atmosphere), advected by the fluid in which it is inserted (water or air), therefore, being modelled by a passive particle. But the AT can also control its trajectory. Therefore, the control problem of a single particle advected by N point vortices is formulated from the Eqs. (7.1)-(7.2) adding to the right-hand sides of Eq. (7.2) the control vector function $\mathbf{U}(t)$, leading to

$$\dot{\mathbf{x}} = \frac{1}{2\pi R} \sum_{i=1}^{N} k_i \frac{\mathbf{x}_i \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_i\|^2} + \mathbf{U}\,(t)\,, \tag{7.3}$$

with the respective initial conditions.

The vector position has norm equal to R, that is, $\mathbf{x} \cdot \mathbf{x} = R^2$. Since the motion of the vortices and particles takes place on a sphere with radius R, this implies that the control vector $\mathbf{U}$ must be orthogonal to the position vector of the passive particle $\mathbf{x}$, i.e.

$\mathbf{x} \cdot \mathbf{U} = 0$ (see previous chapters and [19] for details). Considering $\mathbf{x}(t) = (x(t), y(t), z(t))$, the control vector will be given by

$$\mathbf{U}(t) = \alpha(t)\ (y(t), -x(t), 0)\ +\ \beta(t)\ (0, z(t), -y(t))\,, \tag{7.4}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are two scalar control functions that enable the particle to move in any direction on the surface of the sphere. Since the control $\mathbf{U}(\cdot)$ depends on $\alpha(\cdot)$ and $\beta(\cdot)$, from now on we assume $\mathbf{U}(\cdot) = (\alpha(\cdot), \beta(\cdot))$, and Eq. (7.3) is rewritten as

$$\dot{\mathbf{x}} = \frac{1}{2\pi R} \sum_{i=1}^{N} k_i \, \frac{\mathbf{x}_i \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_i\|^2} + \alpha\ (y, -x, 0) + \beta\ (0, z, -y)\,. \tag{7.5}$$

Defining $\|\mathbf{U}(\cdot)\|^2 = \alpha(\cdot)^2 + \beta(\cdot)^2$, the function to be minimized is

$$f(\mathbf{U}) = \int_0^T \|\mathbf{U}(\cdot)\|^2 \, dt\,, \tag{7.6}$$

where T is the exact time the passive particle has to make the displacement from an initial point P_0 to the end point P_f, which models the land serving constraint. The minimisation of (7.6) corresponds to the optimal reduction of the cost of moving the particle between the initial (P_0) and the final point (P_f) so that the particle reaches the destination at the prescribed final time T, as in [17, 75, 74, 18]. Here, P_f is deliberately chosen far from P_0 so that we can better observe the optimal trajectories obtained.

As in the previous chapters, this control problem is solved by a direct approach based on the discretisation of the control function $\mathbf{U}(\cdot)$ that is replaced by n control vectors $\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_{n-1}$. Each control vector $\mathbf{u}_i$ is constant over a time interval $\Delta t = T/n$. Therefore, the objective function is given by

$$f_n = \Delta t \left(\|\mathbf{u}_0\|^2 + \|\mathbf{u}_1\|^2 + \ldots + \|\mathbf{u}_{n-1}\|^2 \right), \tag{7.7}$$

which corresponds to the discretisation of the function (7.6) using the rule of rectangles. The discretised problem is then numerically solved using the Matlab nonlinear optimisation solver `fmincon` [77]. The main restriction consists of reaching the targeted destination P_f and is given by

$$\|\mathbf{x}(T) - P_f\| \leq 10^{-2}. \tag{7.8}$$

To obtain the final position $\mathbf{x}(T)$ of the passive particle, Eq. (7.5) is solved numerically in cascade by the fourth-order Runge-Kutta method. The final position in the previous time subinterval gives the initial position of the passive particle in the next subinterval. In the first subinterval, the starting point is P_0, and in the last subinterval, the condition (7.8) must be verified.

7.1.2 Passive Particle Advected by a Single Vortex

In this section, we study the dynamics of a passive particle (AT) in a spherical flow with motion induced by a single vortex, $N = 1$, localised in the North Pole of the sphere, $V(0) = (0,\, 0,\, 1)$, with circulation k.

As a consequence of the orthogonality condition between the control vector and the position vector, given by Eq. (7.4), the Eq. (7.5) can be written row-wise

$$\begin{cases} \dot{x} = \frac{k}{2\,\pi\,R}\,\frac{-y}{x^2+y^2+(z-R)^2} + \alpha_i\, y \\ \\ \dot{y} = \frac{k}{2\,\pi\,R}\,\frac{x}{x^2+y^2+(z-R)^2} - \alpha_i\, x + \beta_i\, z \quad , \qquad i = 0, 1, \ldots, n-1\,. \\ \\ \dot{z} = -\beta_i\, y \end{cases} \tag{7.9}$$

The dynamics of the passive particles in each time subinterval are then given by Eq. (7.9) with the respective initial condition. The solution of the constrained optimisation problem previously described will enable to find the optimal controls $\mathbf{u}_i = (\alpha_i, \beta_i)$, $i = 0, 1, \ldots, n-1$, which allow the passive particle to move from the starting point, $P_0 = (x_0, y_0, x_0)$, to the target point, $P_f = (x_f, y_f, x_f)$, in exactly T units of time.

7.1.3 Passive Particle Advected by Two Vortices

In this section, we study the dynamics of a passive particle (AV) in a spherical flow with motion induced by two vortices, $N = 2$, initially localised at $\mathbf{x}_1(0) = V_{01}$ and $\mathbf{x}_2(0) = V_{02}$, with circulations k_1 and k_2.

The position of the two vortices ($\mathbf{x}_1$ and $\mathbf{x}_2$) are governed by

$$\begin{cases} \dot{\mathbf{x}}_1 = \frac{1}{2\,\pi\,R}\; k_2\; \frac{\mathbf{x}_2 \times \mathbf{x}_1}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \\ \\ \dot{\mathbf{x}}_2 = \frac{1}{2\,\pi\,R}\; k_1\; \frac{\mathbf{x}_1 \times \mathbf{x}_2}{\|\mathbf{x}_2 - \mathbf{x}_1\|^2} \end{cases} \tag{7.10}$$

with the respective initial conditions $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$.

The passive particle, initially at $\mathbf{x}(0) = P_0$, is governed by the equation

$$\dot{\mathbf{x}} = \frac{1}{2\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_1\|^2} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{\|\mathbf{x} - \mathbf{x}_2\|^2}\right) + \alpha_i\;(y, -x, 0) + \beta_i\;(0, z, -y)\,, \tag{7.11}$$

for $i = 0, 1, \ldots, n-1$, where α_i and β_i are the controllers.

This control problem is solved similarly to the problem with a single vortex ($N = 1$). The problem is discretised over time and reformulated as an optimisation problem. Since the control vectors have the form $\mathbf{u}_i = (\alpha_i, \beta_i)$, the problem reduces to determining the values of the parameters α_i and β_i , for every individual subinterval identified by $i = 0, 1, \ldots, n-1$.

7.2 Application to Oceanic Garbage Collection

As mentioned in the Introduction, the concept of dynamics of passive particles induced by vortex points in a sphere is applied to computationally simulate the removal of a set of $n_p = 30$ trash accumulations from a garbage patch located at a given oceanic region around the point C_0 of the sphere. The garbage is collected by a set of 30 AVs (here, AV can be interpreted as a realisation of AT), which behave like passive particles capable of controlling their trajectory to transport the collected garbage to the targeted location, defined by the point P_f, for further processing (recycling, storage, etc.).

The initial location of the $n_p = 30$ trash accumulations is randomly generated in a square, centred on the point $C_0 = (0.866, 0, -0.5)$, with 0.15 units on a side. The destination point of all the trash accumulations is $P_f = (-0.866, 0, 0.5)$.

It is considered that each AV has reached the destination if its position $\mathbf{x}(T)$ is at a distance from P_f less than 10^{-2}, i.e. $\|\mathbf{x}(T) - P_f\| < 10^{-2}$. Each AV has to do the

displacement in exactly $T = 10$ units of time.

The trajectories were determined by the method described in Section 7.1 in which the dynamics of the AVs are given by Eq. (7.9), if the flow is induced by a single vortex, or by Eqs. (7.10)-(7.11), if two vortices induce the flow. The controls are determined to minimise the objective function f_n, defined by Eq. (7.7).

7.2.1 Flow Induced by a Single Vortex

This section considers that the fluid in which the AVs are inserted has a movement induced by a single point vortex located at the North Pole of the sphere, with a circulation $k = 2$.

Figure 7.1 shows the trajectory performed by the 30 AVs for transporting the collected trash. The trajectories were determined by the method described in Section 7.1 in which the dynamics of the AVs are given by Eq. (7.9). In the left of Figure 7.1 two ($n = 2$) constant controls $\mathbf{u}_0 = (\alpha_0, \beta_0)$ and $\mathbf{u}_1 = (\alpha_1, \beta_1)$ were used for each displacement, and in right of Figure 7.1 six ($n = 6$) constant controls $\mathbf{u}_i = (\alpha_i, \beta_i)$, $i = 0, 1, ..., 5$, were used.

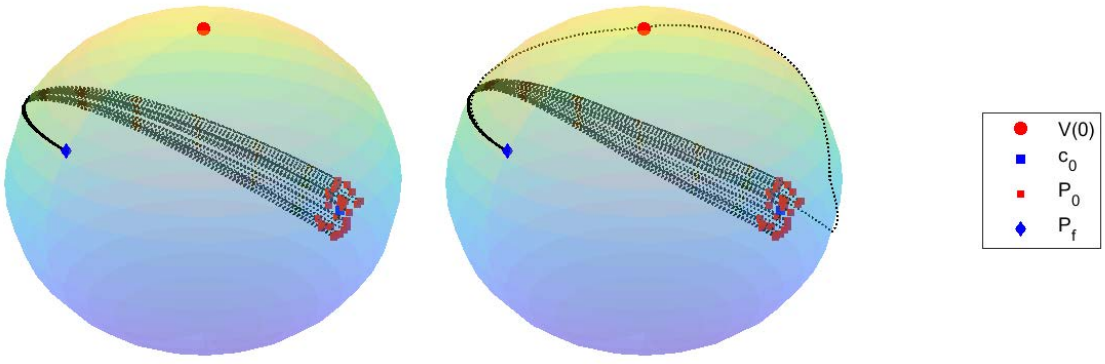


Figure 7.1: Trajectories of the thirty AVs in a flow induced by $N = 1$ vortex: (i) left $n = 2$, (ii) right $n = 6$.

The results obtained with $N = 1$ indicate that, in the case of $n = 2$ controls, the

average optimal controls are $\mathbf{u}_0 = (-0.1075, -0.1660)$, and $\mathbf{u}_1 = (-0.0758, -0.1452)$ and the average objective function is $f_2 = 0.3322$. In the case of $n = 6$, the average optimal controls are $\mathbf{u}_0 = (-0.0952, -0.1332)$, $\mathbf{u}_1 = (-0.0888, -0.1271)$, $\mathbf{u}_2 = (-0.0948, -0.1333)$, $\mathbf{u}_3 = (-0.0887, -0.1267)$, $\mathbf{u}_4 = (-0.0951, -0.1327)$, and $\mathbf{u}_5 = (-0.0890, -0.1262)$ and the average objective function is $f_6 = 0.3647$. Thus, a slight increase in average energy costs is observed with an increase in the number of controls. The increase in controls leads to alternative trajectories to the most frequent ones. However, alternative paths have higher energy costs.

7.2.2 Flow Induced by Two Vortices

In this section, the AVs are inserted in flow induced by a two vortices initially located at points $V_1(0) = (0.2185, 0.2185, 0.9511)$ and $V_2(0) = (0.4330, 0.7500, -0.5000)$, and with equal circulations $k_1 = k_2 = 2$.

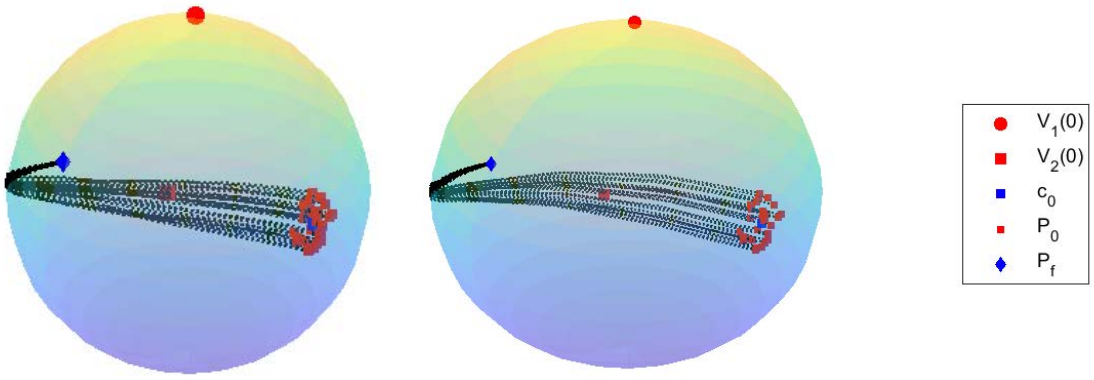


Figure 7.2: Trajectories of thirty AVs in a flow induced by $N = 2$ vortices: (i) left $n = 2$, (ii) right $n = 6$.

Figure 7.2 shows the trajectory performed by the 30 AVs for transporting the collected trash. The trajectories were determined by the method described in Section 7.1 in which

the dynamics of the AVs are given by Eq. (7.9). In the left of Figure 7.2 two ($n = 2$) constant controls $\mathbf{u}_0 = (\alpha_0, \beta_0)$ and $\mathbf{u}_1 = (\alpha_1, \beta_1)$ were used for each displacement, and in the right of Figure 7.2 six ($n = 6$) constant controls $\mathbf{u}_i = (\alpha_i, \beta_i)$, $i = 0, 1, ..., 5$, were used.

In the case of $n = 2$, the average optimal controls are $\mathbf{u}_0 = (-0.2042, 0.0360)$, $\mathbf{u}_1 = (-0.2214, 0.0497)$ and the average objective function is $f_2 = 0.4754$.

In the case of $n = 6$, the average optimal controls are $\mathbf{u}_0 = (-0.1759, -0.0400)$, $\mathbf{u}_1 = (-0.2114, 0.0247)$, $\mathbf{u}_2 = (-0.2195, 0.0686)$, $\mathbf{u}_3 = (-0.2243, 0.0807)$, $\mathbf{u}_4 = (-0.2161, 0.0430)$, and $\mathbf{u}_5 = (-0.1647, -0.0300)$ and the average objective function is $f_6 = 0.4433$. Thus, a slight decrease in average energy costs is observed with an increase in the number of controls.

7.3 Concluding Remarks

A simple mathematical model is proposed to simulate a collection of ocean waste by autonomous technology. Passive particles model the autonomous technology responsible for trash transport, and the ocean current is generated by the motion of point vortices on a sphere.

This toy model can be used for educational purposes and as an introduction to a research activity since it blends concepts from physics and mathematics. It is also a low-cost alternative to experimental studies with real prototypes.

This simplified model also gives the student community the perspective and awareness that technological fields such as mathematics, physics, programming, control, etc., can collaborate to address terrestrial ecosystem challenges.

The preliminary findings focus on optimising the displacement of autonomous vehicles interacting with vortices on a sphere's surface. The results indicate the existence of multiple trajectories for transporting garbage particles. However, energy costs are lower for the most direct trajectories between the starting and destination points. The number of controls corresponds to increased vehicle autonomy and can help define trajectories that minimise energy costs.

It is necessary to incorporate specifications such as autonomy and degrees of freedom of autonomous technology to improve the mathematical model and better adapt to reality.

Chapter 8

Application to Steering Oceanic Debris to a Targeted Region

The results from this chapter have already been published as:
Balsa, C., Otero-Espinar, M.V., Gama, S. (2024). *A Simple Mathematical Model to Steering Oceanic Debris to a Targeted Region.* In: Guarda, T., Portela, F., Diaz-Nafria, J.M. (eds) Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2023. Communications in Computer and Information Science, vol. 1937. Springer, Cham. ISBN 978-3-031-48929-7. ISSN 1865-0929. https://doi.org/10.1007/978-3-031-48930-3_5.

Ocean plastic pollution is a growing environmental crisis that requires immediate attention. According to estimates, approximately eight million tons of plastic waste are introduced into the oceans every year [62]. This excessive amount of plastic waste devastates marine life and ecosystems. Plastic pollution also damages fragile habitats such as coral reefs [71]. Initiatives such as the Ocean Cleanup Project aim to rid the oceans of plastic waste using innovative technologies [116]. The article [102] suggests that the Ocean Cleanup's recommended method of utilising 29 plastic collectors to track microplastic movement and density in the ocean indicates that cleanup initiatives should emphasise addressing microplastic flow in coastal areas instead of focusing solely on regions where plastic waste gathers.

The absorption of ocean debris (for example, microplastics) by marine species has an influence on agriculture [49], and hence poses a risk to human health, according to [101]. Indeed, ocean pollution impacts agricultural techniques that rely on marine fertilisers (e.g., fish, seaweed, etc.).

In the chapter presented here, we will assume an idealised scenario where garbage collection is done by a set of autonomous vehicles initially located where the garbage has a higher density. Let's assume that the autonomous vehicles are relatively small, such that they can be treated as passive particles, meaning they do not alter the flow's local structure of the velocity field.

Once the holds of these autonomous vehicles are full, they must travel to a specific area/location to deposit the collected material there. The movement of these vehicles uses the ocean current as the primary source of displacement. It has a system of piecewise constant controls, based on a small number of pre-defined switching points, whose sum of the squares of the amplitudes (objective function) is minimal. Here, the motion of point vortices on a sphere generates the ocean current.

The relevance of these point vortices stems from their ability to provide a simplified approximation of specific geophysical flows. These flows exhibit long-term persistence and are influenced by the Earth's curvature [120]. Vortex point models offer answers to numerous studies concerning the fundamental dynamics of atmospheric flows [94]. Point vortex conceptual models are utilised to identify and analyse physical phenomena that influence the structure and interactions of atmospheric and oceanic vortices [84]. The presence of boundaries can also be addressed for path planning of an autonomous underwater vehicle, like in [42] for steady flows, and in [43] for a flow with tidal variations.

This work focuses on optimising the displacement of a set of passive particles (autonomous vehicles) interacting with vortices on the surface of a sphere [21]. More precisely, our focus lies in determining the ideal relocation of this group, specifically on the sphere, between two fixed regions, by minimising the energy expended on moving each particle within a predetermined time frame required for the relocation.

Similar to the previous chapters, we tackle this problem by treating the displacement of the passive particle as an optimisation problem. To accomplish this, we employ a direct numerical approach. Contrary to Chapter 7 where Carteisan coordinates are used for modelling the dynamics, the present chapter focuses on the dynamics of autonomous

vehicles based on spherical coordinates.

The chapter is structured as follows: Section 8.1 introduces the equations that model the dynamics of a passive particle and the point vortices on a sphere. Section 8.2 presents the numerical method to solve the optimal control problem of determining passive particle trajectories. Section 8.3 discusses applying the mathematical model to the problem of ocean debris removal. The analysis includes cases where one or two vortices induce ocean currents. Finally, the chapter concludes with some final considerations in Section 8.4.

8.1 Point Vortices and Passive Particles on a Spherical Surface

The dynamics of N point vortices on a sphere, with their initial positions specified in spherical coordinates at $t = 0$, is described by the following system of ordinary differential equations ($i = 1, 2, \ldots, N$):

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_i - \phi_j)}{1 - \cos(\gamma_{ij})} \,, \tag{8.1}$$

$$\dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)}{\sin(\theta_i)\,(1 - \cos(\gamma_{ij}))} \,. \tag{8.2}$$

The present study is focused on a sphere with a constant radius indicated as R. The colatitude angle, denoted as θ, varies from 0 (at the North Pole) to π (at the South Pole) and corresponds to the angle formed between the radius passing through the North Pole, located at $(0, 0, R)$, and the radius passing through the vortex. The longitude angle ϕ lies within the interval $[0, 2\pi)$ and denotes the angular difference between the prime and meridian passing through the vortex position. Furthermore, γ_{ij} represents the central angle between the ith and jth vortex points and can be defined as follows:

$$\cos(\gamma_{ij}) = \cos(\theta_i)\cos(\theta_j) + \sin(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j) \,. \tag{8.3}$$

Eqs. (8.1) and (8.2) can be expressed more concisely using the Bogomolov notation [32] as follows:

$$\dot{\theta}_i = -\frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\alpha_{ij}}{1-\cos(\gamma_{ij})}\,, \quad \dot{\phi}_i = \frac{1}{4\pi R^2} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j \, \frac{\beta_{ij}}{\sin(\theta_i)\,(1-\cos(\gamma_{ij}))}\,, \tag{8.4}$$

with $\alpha_{ij} = \sin(\theta_j)\sin(\phi_i - \phi_j)$ and $\beta_{ij} = \sin(\theta_i)\,\cos(\theta_j) - \cos(\theta_i)\sin(\theta_j)\cos(\phi_i - \phi_j)$.

By definition, a point vortex with zero circulation is referred to as a passive particle. Hence, the motion of a system consisting of a single passive particle located at coordinates (θ_p, ϕ_p) and being advected by N point vortices, follows the Eqs. (8.1)-(8.2), with the additional equations that describe the behavior of the passive particle:

$$\dot{\theta}_p = -\frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})}\,, \tag{8.5}$$

$$\dot{\phi}_p = \frac{1}{4\pi R^2} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})}. \tag{8.6}$$

After applying appropriate initial conditions, we can introduce a time reparametrisation ($\tau := t/R^2$) in Eqs. (8.1)-(8.2) and (8.5)-(8.6). This allows us to set $R = 1$ throughout the subsequent analysis.

8.2 Numerical Solution of the Control Problem

The control problem at hand is derived from Eqs. (8.5)-(8.6), where we augment their right-hand sides with the angular controls u_θ and u_ϕ :

$$\dot{\theta}_p = -\frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\sin(\theta_j)\sin(\phi_p - \phi_j)}{1-\cos(\gamma_{pj})} + u_\theta\,, \tag{8.7}$$

$$\dot{\phi}_p = \frac{1}{4\pi} \sum_{j=1}^{N} k_j \, \frac{\cos(\theta_j) - \cot(\theta_p)\sin(\theta_j)\cos(\phi_p - \phi_j)}{1 - \cos(\gamma_{pj})} + u_\phi \, . \tag{8.8}$$

The function that we seek to minimise is

$$f(\mathbf{U}) = \int_0^T \|\mathbf{U}\|^2 \, dt \, , \tag{8.9}$$

where $\mathbf{U} = (u_\theta \; u_\phi)$, T is the time that the passive particle has to make the displacement from an initial point P_0 to the end point P_f. The displacement has to be done in exactly T units of time. The objective behind minimising (8.9) is to find out the most effective strategy for guiding the particle from its initial point, P_0, to the final point, P_f, within the given time constraint T . This optimisation for cost reduction has been studied in previous literature [17, 75, 74, 18].

The discretized problem is then solved using the Matlab function `fmincon` (as in [17]) with the objective function:

$$f_n = \Delta t \left(\|\mathbf{u}_0\|^2 + \|\mathbf{u}_1\|^2 + \ldots + \|\mathbf{u}_{n-1}\|^2 \right) \, . \tag{8.10}$$

The method employed here is the discretisation of the function (8.9) using the rule of rectangles. For constrained optimisation, the solver offers multiple algorithms, including the Interior Point and Active Set methods, which are documented in [77]. Within each subinterval, the vortices' dynamics are integrated using the numerical scheme of fourth-order Runge-Kutta [46].

8.2.1 Transport Dynamics for a Single Vortex

For the scenario of the one vortex problem ($N = 1$), with the vortex positioned at the North Pole of the sphere and having a circulation of k, the motion of the passive particle is described by the following dynamics:

$$\begin{cases} \dot{\theta}_p = u_\theta \,, \\ \\ \dot{\phi}_p = \frac{k}{4\pi} \frac{1}{1-\cos(\theta_p)} + u_\phi \,, \end{cases} \tag{8.11}$$

given the initial condition $P_0 = (\theta_{p0},\, \phi_{p0})$. The variables u_θ and u_ϕ correspond to the angular controls exerted on the passive particle.

8.2.2 Transport Dynamics for Two Vortices

In the two vortices ($N = 2$) problem, the dynamics of the vortices positions $V_1(\tau) = (\theta_1, \phi_1)$ and $V_2(\tau) = (\theta_2, \phi_2)$ are given by

$$\begin{cases} \dot{\theta}_1 = -\frac{k_2}{4\pi} \frac{\sin(\theta_2)\sin(\phi_1-\phi_2)}{1-\cos(\gamma_{12})} \\ \\ \dot{\phi}_1 = -\frac{k_2}{4\pi} \frac{\cos(\theta_2)-\cot(\theta_1)\sin(\theta_2)\cos(\phi_1-\phi_2)}{\sin(\theta_1)(1-\cos(\gamma_{12}))} \\ \\ \dot{\theta}_2 = -\frac{k_1}{4\pi} \frac{\sin(\theta_1)\sin(\phi_2-\phi_1)}{1-\cos(\gamma_{21})} \\ \\ \dot{\phi}_2 = -\frac{k_1}{4\pi} \frac{\cos(\theta_1)-\cot(\theta_2)\sin(\theta_1)\cos(\phi_2-\phi_2)}{\sin(\theta_2)(1-\cos(\gamma_{21}))} \end{cases} \tag{8.12}$$

with the given initial conditions $V_1(0) = (\theta_{10},\, \phi_{10})$ and $V_2(0) = (\theta_{20},\, \phi_{20})$. The value of $\cos(\gamma_{12})$ and $\cos(\gamma_{21})$ are computed in agreement with Equation (8.3).

The interaction between the two vortices and the control significantly influences the dynamics of the passive particle, as depicted in

$$\begin{cases} \dot{\theta}_p = -\frac{1}{4\pi} \left(k_1 \frac{\sin(\theta_1)\sin(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\sin(\theta_2)\sin(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})} \right) + u_\theta \\ \\ \dot{\phi}_p = \frac{1}{4\pi} \left(k_1 \frac{\cos(\theta_1)-\cot(\theta_p)\sin(\theta_1)\cos(\phi_p-\phi_1)}{1-\cos(\gamma_{p1})} + k_2 \frac{\cos(\theta_2)-\cot(\theta_p)\sin(\theta_2)\cos(\phi_p-\phi_2)}{1-\cos(\gamma_{p2})} \right) + u_\phi \end{cases} \tag{8.13}$$

with the initial conditions $P_0 = (\theta_{p0},, \phi_{p0})$. As in the previous problem, the control applied to the passive particle is represented by the vector $\mathbf{U} = (u_\theta\ u_\phi)$.

8.3 Application to Trash Collectors

The dynamics of passive particles induced by point vortices is used to model the dynamics of a set of n_p trash collectors from a garbage patch located at a given point p_0 of the sphere. We consider all trash collectors to be of equal loads. These particles are transported individually to a point P_f on the sphere for further processing (recycling, storage, etc.).

The initial location of the n_p trash particles is randomly generated in a square, centred on the point $C_0 = \left(\frac{2\pi}{3}, 0\right)$, with 0.15 units on a side.

The destination point of trash particles is $P_f = (\frac{\pi}{3}, \pi)$. The particle is considered to have reached the destination if its position $P(T)$ is at a distance from P_f less than 0.01, i.e. $\|P(T) - P_f\| < 0.01$.

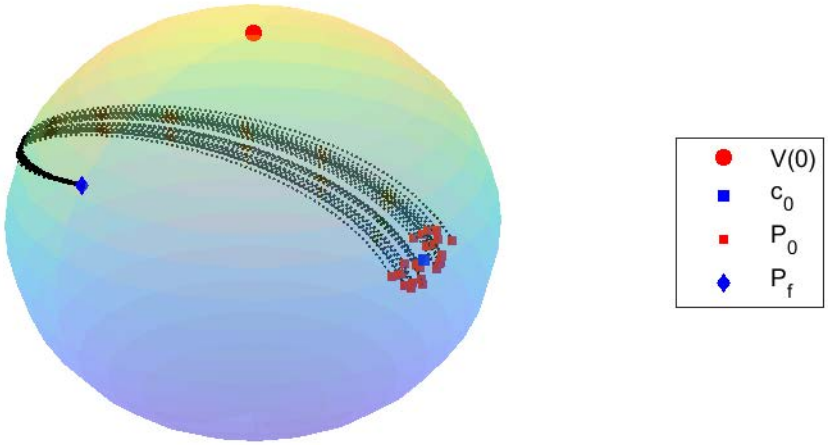
The trajectories were determined by the method described in Section 8.2 in which the dynamics of each particle are given by Eq. (8.11) or Eq. (8.13), and the controls are determined to minimise the objective function (8.10).

8.3.1 Motion Advected by a Single Vortex

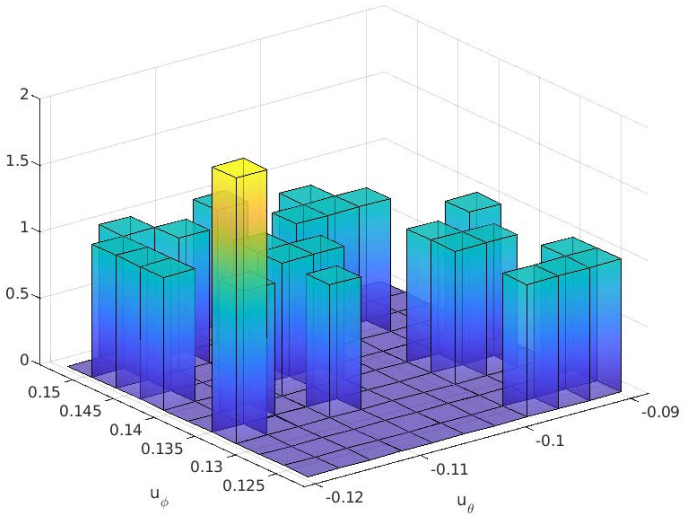
In Figure 8.1a, we can observe the optimal trajectory of $n_p = 25$ trash particles being advected by a single vortex with circulation $k = 2$, situated at the North Pole. A time limit of $T = 10$ constrains the displacement of each trash particle. For each displacement a single constant control $\mathbf{u}_0 = (u_\theta, u_\phi)$ was used (n=1). The average optimal control found was $\mathbf{u}_0 = (-0.105, 0.139)$ and the average objective function $f_1 = 1.16$.

Figure 8.1b illustrates the histogram of the single control $\mathbf{u}_0 = (u\theta,\, u_\phi)$ responsible for the trajectories depicted in the preceding figure. The histogram demonstrates that u_θ values vary between -0.12 and -0.09, while the u_ϕ values range from 0.125 to 0.150 .

In Figure 8.2a, the optimal trajectory of $n_p = 25$ trash particles is depicted, advected by a single vortex with circulation $k = 2$, situated at the North Pole. Each trash particle is allocated a time limit of $T = 10$ for movement. The trajectories have been generated using four constant controls ($n = 4$) for each displacement. The averages of the control values are $\mathbf{u}_0 = (-1.06 \times 10^{-7}, 2.87 \times 10^{-8})$, $\mathbf{u}_1 = (-1.85 \times 10^{-8}, 1.66 \times 10^{-8})$, $\mathbf{u}_2 = (-4.19 \times 10^{-1}, 5.52 \times 10^{-1})$, and $\mathbf{u}_3 = (-7.37 \times 10^{-8}, 1.48 \times 10^{-7})$, and the corres-



(a) Trajectories



(b) Histogram of $\mathbf{u}_0$

Figure 8.1: Transport with $n = 1$ control of $n_p = 25$ trash particles advected by a single vortex ($N = 1$) located in the North Pole.

ponding average objective function is $f_4 = 1.30 \times 10^{-6}$.

The histograms of the four controls, $\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2$ and $\mathbf{u}_3$, responsible for the trajectories depicted in the Figure 8.2a are presented in Figure 8.2b. It is possible to observe in this histogram that the controls $\mathbf{u}_0$, $\mathbf{u}_1$, and $\mathbf{u}_3$ assume minimal values in comparison with the values assumed by $\mathbf{u}_2$.

8.3.2 Motion Advected by Two Vortices

Figure 8.3a shows the optimal trajectory of $n_p = 25$ trash particles advected by two vortices initially located at $(\pi/10,\ \pi/4)$ and $(2\pi/3,\ \pi/3)$, both with circulation $k = 2$. The time limit for moving each trash particle is $T = 10$. A single constant control ($n = 1$) has been used for each displacement. The average of the control $\mathbf{u}_0$ and of the objective function f_1 are, respectively, $(0.0189, 0.0333)$ and 0.0257.

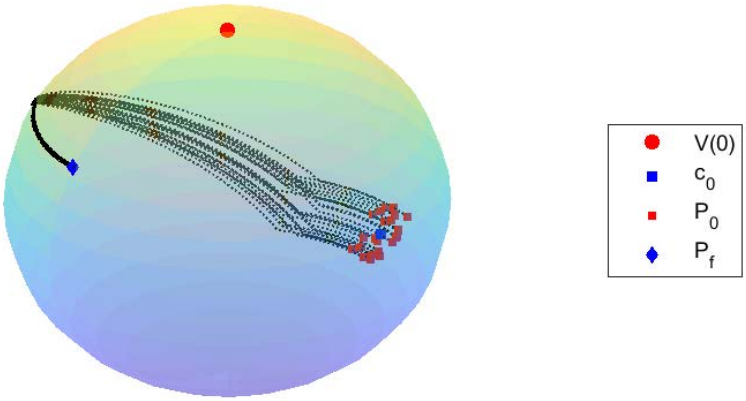
The histograms of the control $\mathbf{u}_0$ responsible for the trajectories shown in Figure 8.3a are presented in Figure 8.3b. This histogram shows that the u_θ values vary between -0.04 and 0.08, while the u_ϕ values range from approximately 0.02 to 0.05 .

Figure 8.4a shows the trajectory of $n_p = 25$ trash particles advected by two point vortices ($N = 2$) initially located at $(\pi/10,\ \pi/4)$ and $(2\pi/3,\ \pi/3)$, both with circulation $k = 2$. Each trash particle has an available time of $T = 10$. Four constant controls ($n = 4$) have been used for each displacement. The average of the control values are $\mathbf{u}_0 = (0.0058, -0.0251)$, $\mathbf{u}_1 = (-0.0281, -0.0077)$, $\mathbf{u}_2 = (0.0185, 0.0301)$, and $\mathbf{u}_3 = (-0.0035, 0.0749)$, and the corresponding average objective function is $f_4 = 0.3392$.

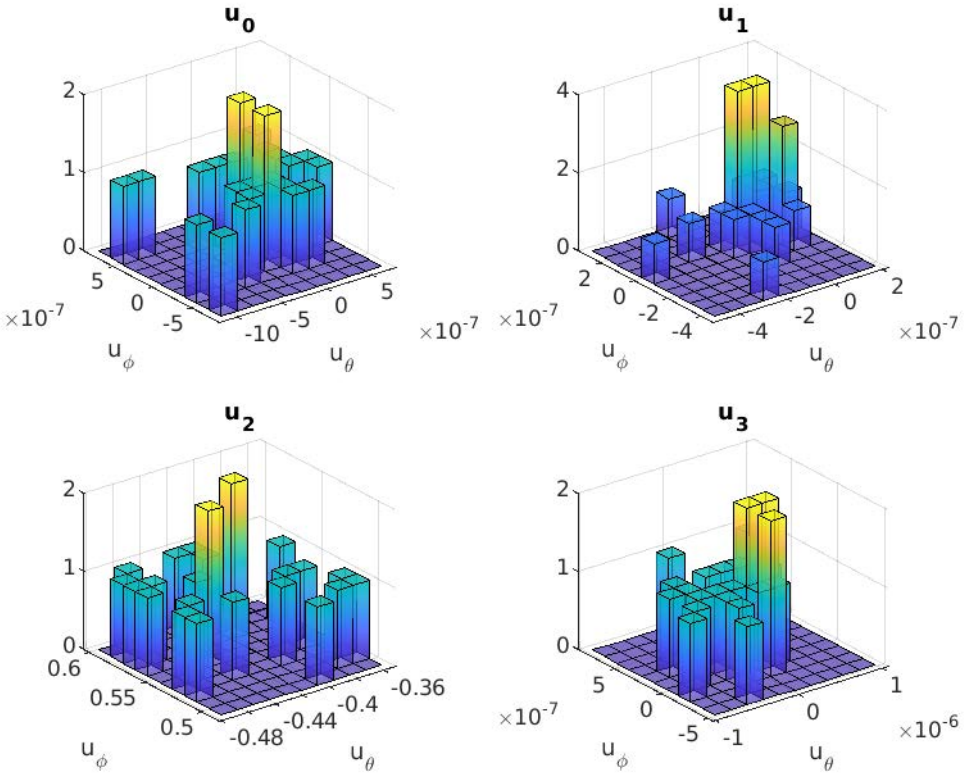
When comparing the outcomes obtained using a single vortex, it is evident that both the control values and the objective function show higher values. Additionally, some divergent trajectories were observed in the majority of cases.

Figure 8.4b showcases the histograms of the four controls, namely $\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2$ and $\mathbf{u}_3$, associated with the trajectories depicted in Figure 8.4a. Upon analysing this histogram, it becomes apparent that certain values significantly deviate from the most common values, which correspond to alternative trajectories that offer less direct connections between the starting and arrival points.

Figure 8.5 depicts the histograms illustrating the objective function values derived from

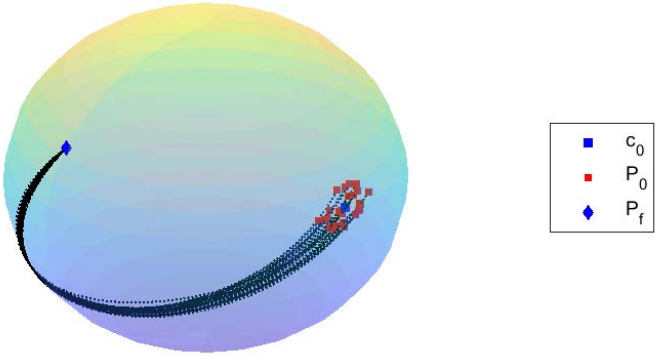


(a) Trajectories

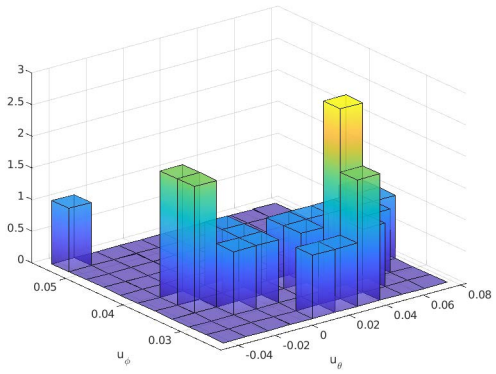


(b) Histograms of $\mathbf{u}_0$, $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$

Figure 8.2: Transport with $n = 4$ controls of $n_p = 25$ trash particles advected by $N = 1$ vortex located in the North Pole.

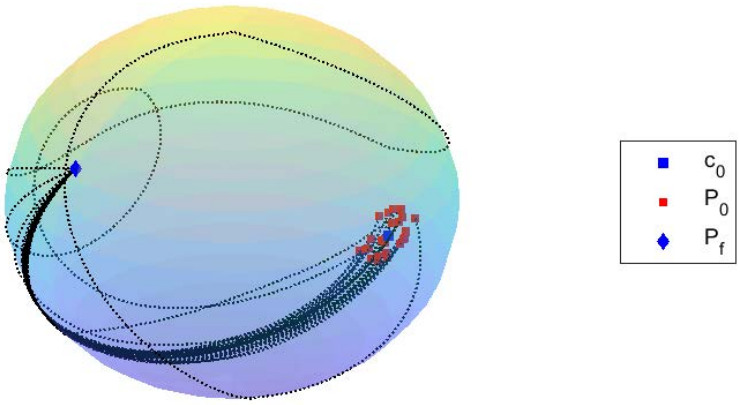


(a) Trajectories

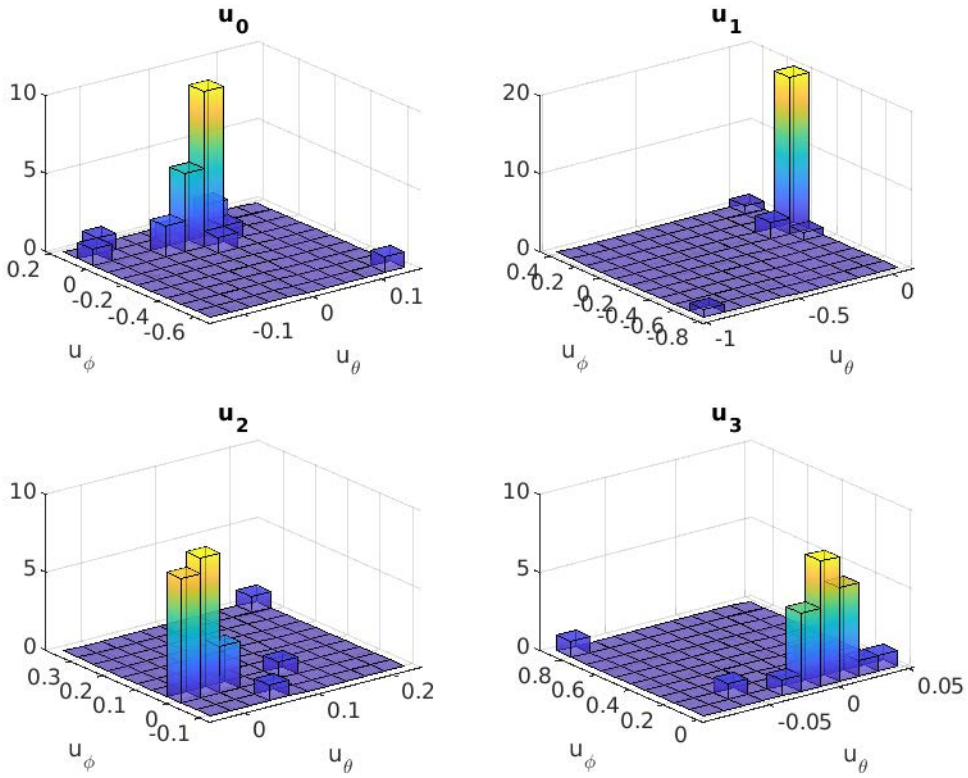


(b) Histogram of $\mathbf{u}_0$

Figure 8.3: Transport with $n = 1$ control of $n_p = 25$ trash particles advected by $N = 2$ vortices.



(a) Trajectories



(b) Histograms of $\mathbf{u}_0$, $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$

Figure 8.4: Transport with $n = 4$ control of $n_p = 25$ trash particles advected by $N = 2$ vortices.

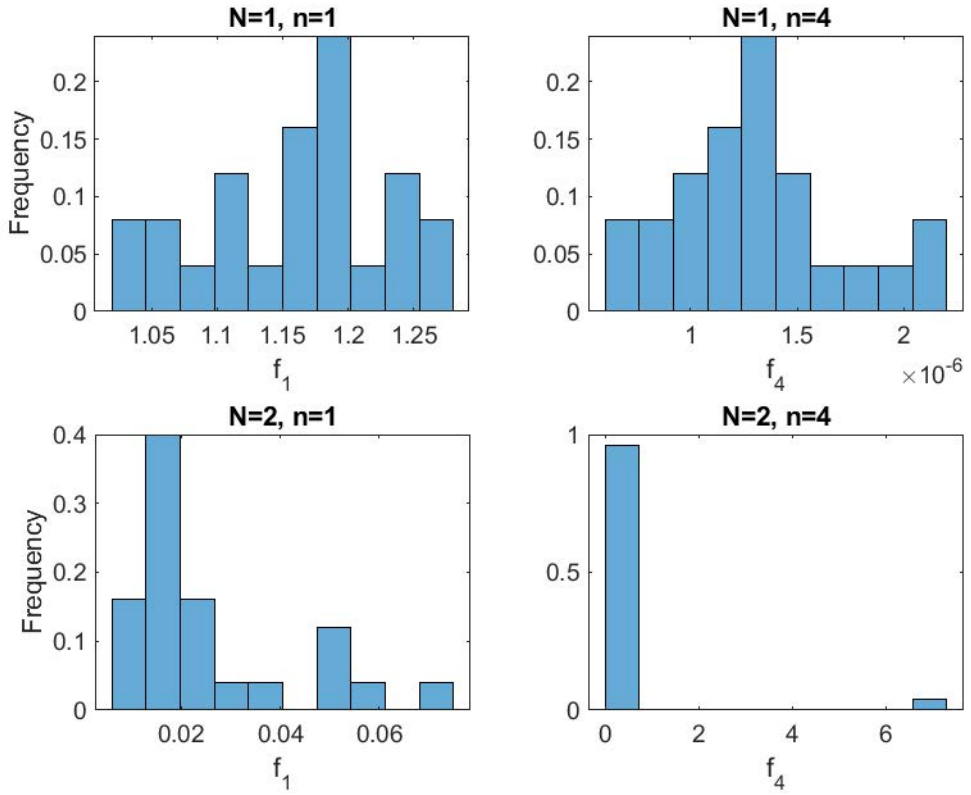


Figure 8.5: Histograms of the objective function f_n resulting from the transport of $n_p = 25$ trash particles for $N = 1, 2$ and $n = 1$ and 4.

the analysis of four specific scenarios: $N = 1$ with $n = 1$ and $n = 2$, and $N = 2$ with $n = 1$ and $n = 2$. It is evident that there is no consistent pattern in the fluctuation of f_n concerning N or n. The minimum values occur when $N = 1$ and $n = 4$. Conversely, the maximum values are observed when $N = 2$ and $n = 4$. However, in the latter case, the majority of f_n values are situated near 0, with only a small subset exceeding 6. These higher values correspond to alternative trajectories deviating from the direct trajectories depicted in Figure 8.4a. Consequently, the findings highlight that the trajectories characterised by lower energy consumption (lower objective function values) are the ones that follow the most direct trajectories.

8.4 Concluding Remarks

The cleaning of accumulated garbage in the oceans is a pressing problem. Waste, especially plastic, is affecting marine ecosystems and human food chains. This chapter introduces a simple mathematical model for oceanic garbage collection. Passive particles model the autonomous vehicles responsible for the trash transport, and the ocean current is generated by the motion of point vortices on a sphere.

A system of piecewise constant controls is employed to achieve vehicle autonomy, utilising a limited number of pre-defined switching points that determine the vehicle's trajectory. Each control incurs an energy cost that is aimed to be minimised. This optimisation is accomplished by solving a nonlinear problem on the spherical surface.

The findings focus on optimising the displacement of a group of passive particles (representing autonomous vehicles) interacting with vortices on a sphere's surface. The results indicate the existence of multiple trajectories for transporting garbage particles. However, energy costs are lower for the most direct trajectories between the starting and destination points.

Refining the mathematical model to better align with the reality of autonomous vehicles is crucial to progress further. This entails incorporating specifications such as autonomy and degrees of freedom.

Chapter 9

Controlled Particle Motion on a Rotating Sphere

The results from this chapter have already been published as:
Balsa, C., Otero-Espinar, M.V., Gama, S. (2025). *Particle Movement on a Rotating Sphere*. In: A. P. Aguiar at al. (Eds.): CONTROLO 2024, Lecture Notes in Electrical Engineering 1325, pp. 1-9, 2025. ISBN 978-3-031-81723-6. ISSN 1876-1100. https://doi.org/10.1007/978-3-031-81724-3_29.

This chapter presents a complementary study to the previous ones. This research explores the complex motion of a passive particle within a flow induced by two-point vortices located on the surface of a rotating sphere.

We analyse the problem of controlling the displacement of a passive particle on a rotating spherical surface. For this purpose, two point vortices are considered to induce movement. The optimal control problem corresponding to the displacement of a particle immersed in a flow caused by two vortex points on the surface of a rotating sphere is formulated in Cartesian coordinates.

As in previous chapters, where the studies were conducted on an irrotational sphere, the control problem is converted into a non-linear optimisation problem that is solved numerically through a direct approach. The time to make the displacement is discretised

into a certain number n of sub-intervals. In each of these sub-intervals, the control is considered constant.

The structure of this chapter is as follows: Section 9.1 introduces the equations governing the displacement of vortices and the passive particle, along with formulating the optimisation control problem. Section 9.2 outlines the numerical methodology to address the optimal control problem. Numerical results, dependent on the number of constant controls and sphere rotation speed, are discussed in Section 9.3. Finally, concluding remarks are presented in Section 9.4.

9.1 Dynamics of the Driven Passive Particle

Consider two point vortices, denoted by $\mathbf{x}_i \equiv \mathbf{x}_i(t) = (x_i(t), y_i(t), z_i(t))\,,\ i = 1, 2\,,$ initially located at $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$. It is assumed that these vortices reside on the surface of a sphere undergoing rotation about the $z-$axis.

The dynamics of these two point vortices is given by the differential equations [89]:

$$\begin{cases} \dot{\mathbf{x}}_1 = \dfrac{k_2}{4\,\pi\,R}\,\dfrac{\mathbf{x}_2 \times \mathbf{x}_1}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}_1} + \Omega\,(-y_1, x_1, 0) \\ \\ \dot{\mathbf{x}}_2 = \dfrac{k_1}{4\,\pi\,R}\,\dfrac{\mathbf{x}_1 \times \mathbf{x}_2}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}_2} + \Omega\,(-y_2, x_2, 0) \end{cases} \tag{9.1}$$

where R is the radius of the sphere, k_i is the circulation of the two vortex i (i=1,2), and Ω is the angular velocity of the rotation. The position of the controlled passive particle $\mathbf{x} \equiv (x, y, z)$ is then governed by

$$\dot{\mathbf{x}} = \frac{1}{4\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}}\right) + \Omega\,(-y, x, 0) + \mathbf{U}^c\,, \tag{9.2}$$

with the initial condition $\mathbf{x}(0)$. Here, $\mathbf{U}^c \equiv (u_x, u_y, u_z)$ is the control that acts in the passive particle and has the particular form

$$\begin{cases} u_x = \alpha y \\ u_y = -\alpha x + \beta z \\ u_z = -\beta y \end{cases} \tag{9.3}$$

to maintain the motion on the spherical surface (see Chapter 6). Since the control relies directly on $\alpha \equiv \alpha(t)$ and $\beta \equiv \beta(t)$, it is straightforwardly assumed that the control is given by

$$\mathbf{U} = (\alpha, \beta) \ . \tag{9.4}$$

The energy spent on the displacement is given by the integral over the time interval $[0, T]$ of the sum of the squares of the components of $\mathbf{U}$

$$\int_0^T \|\mathbf{U}\|^2 \, dt = \int_0^T \left(\alpha^2 + \beta^2\right) dt. \tag{9.5}$$

Therefore, (9.5) constitutes the objective function of the control problem, to minimise its value.

9.2 Numerical Solution

The optimal control problem, derived from governing the displacement of the passive particle within a flow induced by two vortices located on the surface of the rotating sphere, can be formulated as the Optimization Problem ($\mathcal{OP}$), presented below.

In the optimization problem $\mathcal{OP}$, $\mathbf{x}_0 \, , \mathbf{x}_f \in \mathbb{R}^3$, such that $\|\mathbf{x}_0\| = \|\mathbf{x}_f\| = R$, represent the initial and final point, $T > 0$ denotes the available time for the displacement, and $u_{\max} > 0$ indicates a given upper limit value for the control.

Optimization Problem ($\mathcal{OP}$)

Minimize :

$$\int_0^T \|\mathbf{U}(t)\|^2 \, dt$$

subject to :

$$\begin{cases} \dot{\mathbf{x}}_1 = \dfrac{k_2}{4\,\pi\,R}\,\dfrac{\mathbf{x}_2 \times \mathbf{x}_1}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}_1} + \Omega\,(-y_1, x_1, 0) \\ \dot{\mathbf{x}}_2 = \dfrac{k_1}{4\,\pi\,R}\,\dfrac{\mathbf{x}_1 \times \mathbf{x}_2}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}_2} + \Omega\,(-y_2, x_2, 0) \\ \dot{\mathbf{x}} = \frac{1}{4\,\pi\,R}\left(k_1\,\frac{\mathbf{x}_1 \times \mathbf{x}}{R^2 - \mathbf{x}_1 \cdot \mathbf{x}} + k_2\,\frac{\mathbf{x}_2 \times \mathbf{x}}{R^2 - \mathbf{x}_2 \cdot \mathbf{x}}\right) + \Omega\,(-y, x, 0) + \mathbf{U}^c \end{cases}$$

$$\mathbf{x}(0) = \mathbf{x}_0$$

$$\mathbf{x}(T) = \mathbf{x}_f$$

$$\|\mathbf{U}\| \leq u_{\max}$$

In this optimisation problem, the objective function to be minimised represents the energy expended in transporting the passive particle from $\mathbf{x}_0$ and $\mathbf{x}_f$, during the time interval $[0, T]$. The restrictions apply to the equations that govern the dynamics of the two vortices and the particle between the initial and final positions, as well as the maximum control norm (maximum available energy).

A direct approach similar to the one used in the previous chapters is applied to solve numerically the optimisation problem $\mathcal{OP}$. The control function $\mathbf{U}(\cdot)$ is discretized in n constant vectors:

$$\begin{aligned} \mathbf{U}(t) &= \mathbf{u}_0 \quad \text{if} \quad t_0 \leq t < t_1\,, \\ \mathbf{U}(t) &= \mathbf{u}_1 \quad \text{if} \quad t_1 \leq t < t_2\,, \\ \mathbf{U}(t) &= \mathbf{u}_2 \quad \text{if} \quad t_2 \leq t < t_3\,, \\ &\vdots \\ \mathbf{U}(t) &= \mathbf{u}_{n-1} \quad \text{if} \quad t_{n-1} \leq t \leq t_n\,, \end{aligned} \tag{9.6}$$

Each constant vector variable $\mathbf{u}_i$, $i = 1, 2, \cdots, n$, corresponds to constant values of the control variables α_i and β_i exercised in the sub-interval $[t_{i-1}, t_i)$ with constant length of time $\Delta t = (t_n - t_0)/n$.

Therefore, the discretisation of the energy function (9.5) using the rectangular rule with

n time sub-intervals leads to the approximation

$$\int_0^T \|\mathbf{U}\|^2 \, dt \approx \Delta t \, \left[\left(\alpha_1^2 + \beta_1^2\right) + \left(\alpha_2^2 + \beta_2^2\right) + \ldots + \left(\alpha_{n-1}^2 + \beta_{n-1}^2\right) \right] \equiv f_n \,, \tag{9.7}$$

and the optimisation problem is replaced by its discretised version (for details see Chapter 6).

The initial position of the passive particle, $\mathbf{x}(t_i)$, in each sub-interval $[t_i, t_{i+1})$, is given by the final position in the previous sub-interval, i.e. $\mathbf{x}(t_{i-1})$. The terminal (or final) condition $\mathbf{x}(T) = \mathbf{x}_f$, in ($\mathcal{OP}$), is considered to be satisfied, if

$$\|\mathbf{x}(T) - \mathbf{x}_f\| < \varepsilon \,, \tag{9.8}$$

where ε denotes a small positive tolerance that is given. In other words, the condition $\mathbf{x}(T) = \mathbf{x}_f$, in $\mathcal{OP}$, is relaxed to (9.8).

The method used to find the optimal trajectory is similar to the shooting method. Uniformly distributed random numbers (see [104]) are used to initialise the controls in Eq. (9.2) that govern the dynamics of the passive particle. This ordinary differential equation together with Eq. (9.1) are numerically integrated in each time sub-interval, $[t_{i-1}, t_i) \,, i = 0, 1, \ldots, n-1$, by a fourth-fifth's order Runge-Kutta method included in the Matlab built-in function `ode45` [78].

The feasible controls and trajectories are the controls and respective trajectories that lead to the final position $\mathbf{x}_f$ that comply with condition (9.8). Among a finite set of possible trajectories, the optimal one yields the minimum value for the objective function. This trajectory and associated controls are termed quasi-optimal. Feasible trajectories are determined using the built-in `fmincon` function, which allows solving non-linear optimisation problems [77].

9.3 Computational Simulations

In all the cases detailed here, we have considered $R = 1$, $T = 5$, $k_1 = k_2 = 1$, $u_{\max} = 2\pi$, and $\varepsilon = 10^{-4}$.

Figure 9.1 shows the trajectories obtained in the displacement of the passive particle

from the initial point $\mathbf{x}_0 = \left(-1/2, 0, \sqrt{3}/2\right)$ to the final point $\mathbf{x}_f = \left(\sqrt{2}/2, 0, -\sqrt{2}/2\right)$, considering that the two vortices are initially located at $\mathbf{v}_1(0) = (0, 0, -1)$ and $\mathbf{v}_2(0) = \left(0, -\sqrt{3}/2, 1/2\right)$, and that the sphere rotation is $\Omega = 1$. Each trajectory corresponds to a different number n of control variables, ranging from 1 to 50. The corresponding values of the objective function are $f_1 = 2.44$, $f_{10} = 1.80$, $f_{20} = 1.79$, and $f_{50} = 1.79$. A general trend towards a decrease in the energy used in displacement f_n is observed with the increase in the number of controls, n. This corresponds to what is expected, as increasing the value of n means more control is available to guide the passive particle to reach its destination more efficiently.

Figure 9.2 shows 300 solutions of the optimisation problem obtained with the direct numerical method described in the previous section in the case of $n = 20$ controls. In Figure 9.2 a) the trajectories are represented, and in Figure 9.2 b) there is the histogram of the objective function values corresponding to the 300 trajectories. It can be observed that most of the trajectories (more than 200) correspond to objective function values close to the minimum ($f_{20} = 1.79$). Most of the remaining solutions obtained correspond to objective function values slightly higher than the minimum value. Only a few solutions give rise to high objective function values (greater than 4). Given that the algorithm starts with a random distribution of controls, this means that the solution found by the algorithm converged to a local minimum that corresponds to a greater energy expenditure. The left-skewed distribution observed in the histogram indicates that most of the trajectories provide optimal solutions near the minimum found value of $f_{50} = 1.79$. Thus, it suggests a high probability that the algorithm will find a trajectory close to the desired minimum.

Figure 9.3 shows the trajectories obtained with different values of the sphere rotation ($\Omega = 0, 1, 2, 5$; counterclockwise rotation). It is observed that for the lowest rotation values ($\Omega = 0$ and $\Omega = 1$), the particle's trajectory is shorter. In contrast, for higher rotation values ($\Omega = 2$ and $\Omega = 5$), the particle circulates several times around the sphere before reaching the destination point. The values of the objective function are $f_{20} = 1.79$ for $\Omega = 0$, $f_{20} = 1.62$ for $\Omega = 1$, $f_{20} = 0.91$ for $\Omega = 2$, and $f_{20} = 2.63$ for $\Omega = 5$. These findings demonstrate that specific rotation values of the sphere aid the particle in reaching its destination. However, when this rotation becomes excessive, the particle requires more energy to accomplish its objective.

Figure 9.4 shows the trajectories obtained with zero and different negative values of the sphere rotation ($\Omega = 0, -1, -2, -5$; clockwise rotation). It can be observed that the

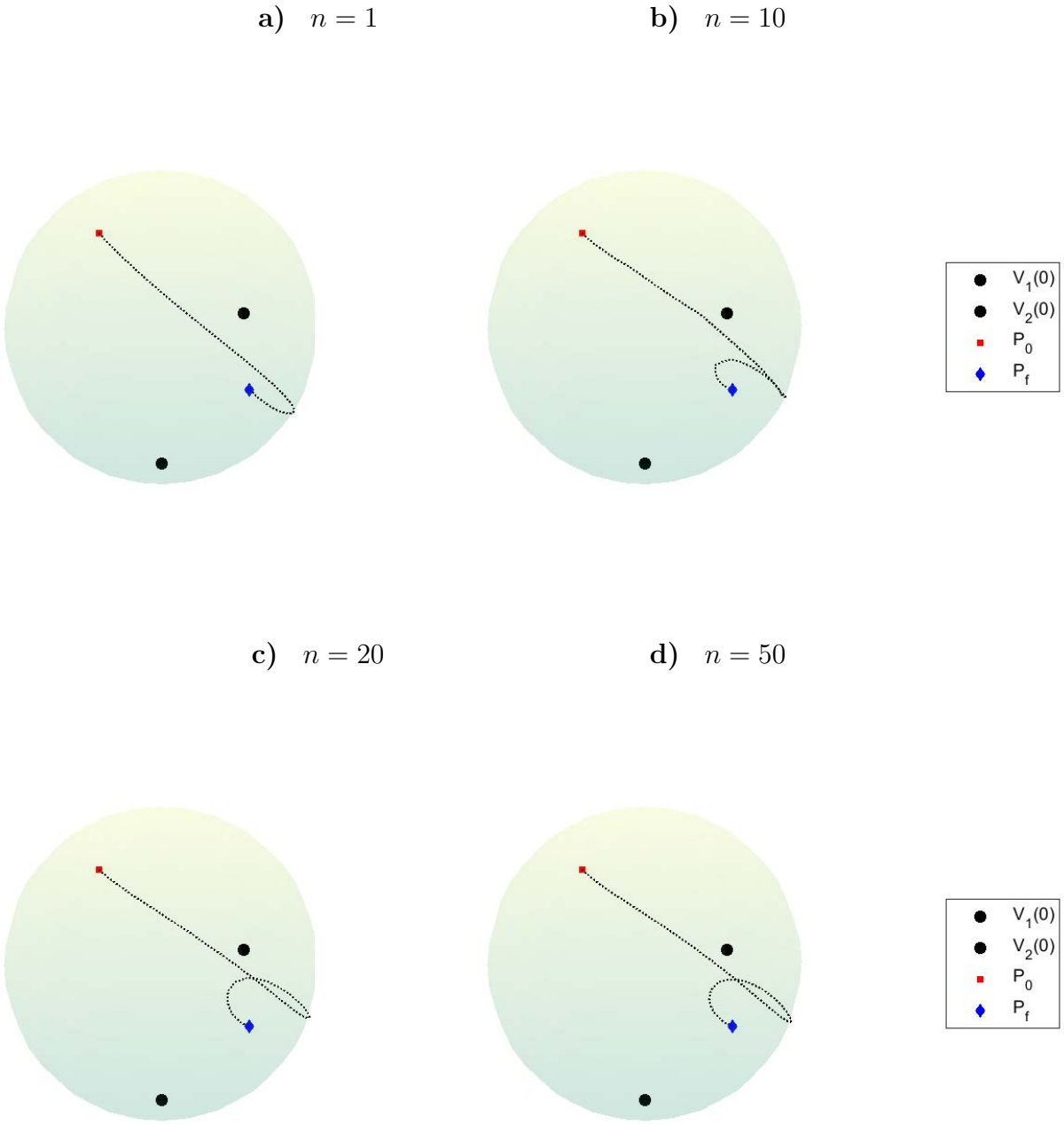


Figure 9.1: Optimal trajectories obtained with different values of n $(= 1, 10, 20, 50)$, with $\Omega = 1$. The objective function yields $f_1 = 2.44$, $f_{10} = 1.80$, $f_{20} = 1.79$, $f_{50} = 1.79$.

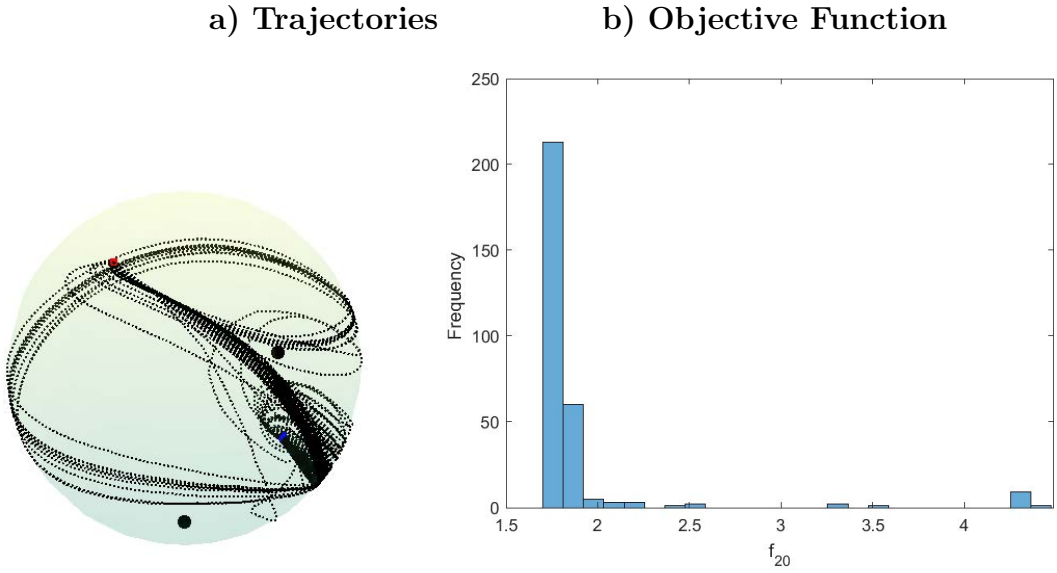


Figure 9.2: Trajectories and their respective objective function values for 300 solutions achieved with $n = 20$ controls and $\Omega = 1$.

trajectories become more complex (the particle circulates several times around the sphere before reaching the endpoint) with the increase in the absolute value of the rotation. However, the value of the objective function is not proportional to the rotation value. The smallest value ($f_{20} = 1.31$) is obtained with $\Omega = -2$. The remaining values of the objective function are $f_{20} = 1.62$ for $\Omega = 0$, $f_{20} = 3.61$ for $\Omega = -1$, and $f_{20} = 2.75$ for $\Omega = -5$.

9.4 Concluding Remarks

The study presented in this chapter outlines the optimal control problem associated with the movement of a particle within a flow induced by two vortex points located on the surface of a rotating sphere.

The problem is numerically solved using a direct approach, whereby the available time for displacement is divided into a specific number of sub-intervals. Within each of these sub-intervals, the control remains constant.

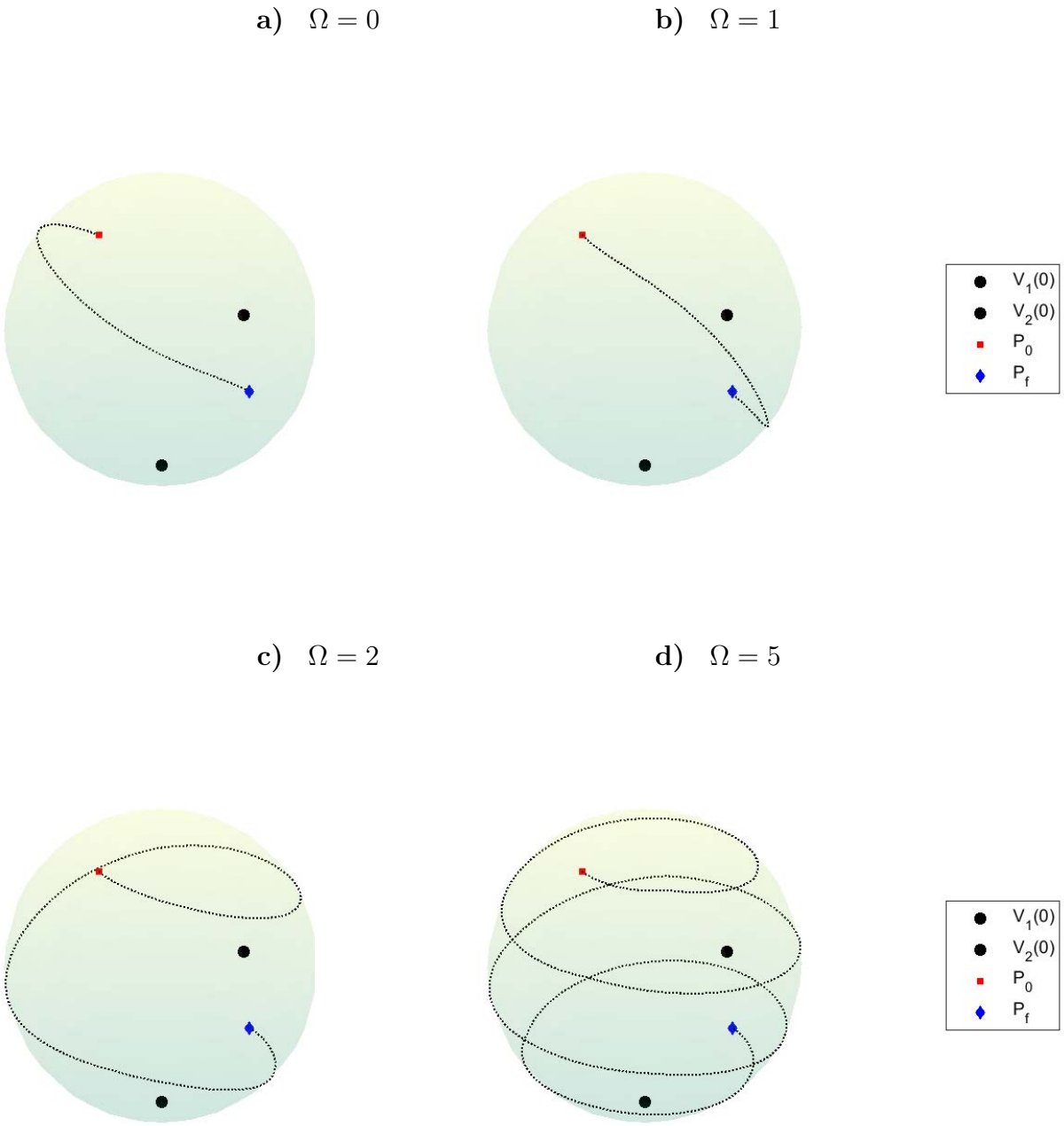


Figure 9.3: Optimal trajectories under different sphere rotation values ($\Omega = 0, 1, 2, 5$), with $n = 20$ controls considered across all cases. The objective function values are $f_{20} = 1.62$, $f_{20} = 1.80$, $f_{20} = 0.91$, $f_{20} = 2.63$, respectively.

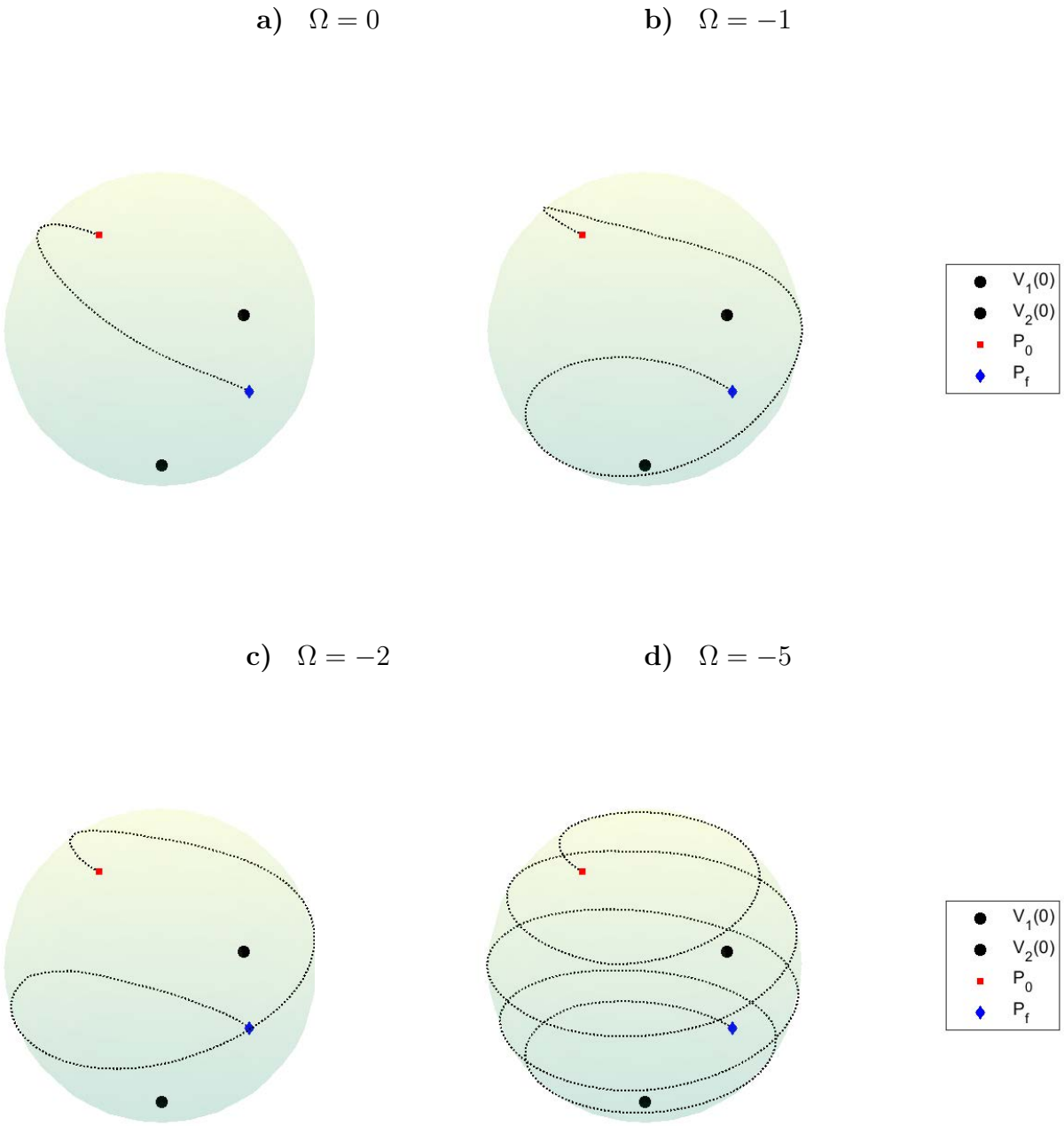


Figure 9.4: Optimal trajectories under different sphere rotation values ($\Omega = 0, -1, -2, -5$), with $n = 20$ controls considered across all cases. The objective function values are $f_{20} = 1.62$, $f_{20} = 3.61$, $f_{20} = 1.31$, $f_{20} = 2,75$, respectively.

It is shown that the problem has a quasi-optimal solution and that the solution depends on the number of control variables used. In general, the greater the number of controls, the less energy is spent to move the particle between two fixed points on the sphere.

The method presented is stochastic. The solution attained depends on the randomly generated initial controls. Iterating the technique enables us to derive various solutions, among which the optimal one corresponds to the minimum energy.

Changing the sphere's rotation speed changes the solution's nature, allowing for displacement with reduced energy requirements or increased energy demands. Typically, when the rotation speed is relatively high, the particle must complete multiple rotations around the sphere before reaching the destination.

Part II

Meteorological Data Reconstruction

Chapter 10

Analogue Ensembles of Principal Components

The results from this chapter have already been published as:
Balsa, C., Breve, M.M., André, B., Rodrigues, C.V., Rufino, J. (2023). *PCAnEn - Hindcasting with Analogue Ensembles of Principal Components*. In: Garcia, M.V., Gordón-Gallegos, C. (eds) CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI). CSEI 2022. Lecture Notes in Networks and Systems, vol. 678. Springer, Cham. ISBN 978-3-031-30591-7. ISSN 2367-3370. https://doi.org/10.1007/978-3-031-30592-4_13.
and the Section 10.1 is based on:
Balsa, C., Rodrigues, C.V., Araújo, L., Rufino, J. (2022). *Cluster-Based Analogue Ensembles for Hindcasting with Multistations*. Computation, 10(6): 91. https://doi.org/10.3390/computation10060091.

Filling gaps in observed time series is an essential problem in many areas of applied sciences that depend on data analysis. Without this filling, data reconstruction is difficult or even impossible. This assumption is particularly true in weather forecasting, where stored information grows four times faster than the world economy [79]. Given this, big data analytics can help improve predictions by uncovering patterns and correlations in the data [61] and reconstructing missing data in areas with limited information. Conversely,

this growth in data also means that the amount of missing data is increasing, which makes accurate reconstruction a crucial task. To handle this challenge, forecasting methods must be able to handle large amounts of data, multiple data sources and a wide variety of meteorological variables. This requires advanced methodologies that can adapt to the particular characteristics of big data in weather forecasting.

Despite the abundance of weather data available, many regions still lack historical data records. These locations, which may be remote or underdeveloped, have the potential to be significant generators of renewable energy. However, without historical weather data, it is difficult to accurately predict the potential for energy generation in such places. Therefore, there is a growing need for methods to generate weather data from limited inputs and locations to run simulations of environmentally driven systems that target these locations. This may significantly enhance our understanding of the potential for renewable energy generation, and may facilitate the development of sustainable energy systems in such regions [69].

Weather prediction often faces two considerable challenges: (i) missing or absent weather data, and (ii) handling large volumes of data. The first challenge can be addressed through weather data reconstruction techniques known as hindcasting. Hindcasting reconstructs missing historical data (non-recorded observations) using a generic prediction model to recreate past weather conditions. For this reason, hindcasting is also used to validate forecast models by comparing their output to past observations.

Besides reconstructing missing data, research in hindcasting also aims to improve various aspects of meteorology, such as downscaling and forecasting methods. One of the key techniques employed in meteorological data reconstruction is the Analog Ensemble (AnEn) method [86, 85]. Although the original AnEn method was first proposed for postprocessing Numerical Weather Predictions (NWPs), this technique has been applied in several areas, as the production of renewable energies (wind and solar) [5, 38].

Given the factors discussed above, it is clear that the AnEn method is up-and-coming for addressing challenges in hindcasting, forecasting and downscaling. Several studies have also been conducted to compare the effectiveness of AnEn with other methods, namely, convolutional neural networks (CNNs), to provide further evidence. For example, studies such as [80, 105] have found that the AnEn method can improve prediction accuracy equally or even outperform CNNs when used to post-process results of regional weather prediction models, such as the Weather Research and Forecasting (WRF). Furthermore,

implementing the AnEn method is relatively straightforward, and its results are easy to understand and explain compared to machine learning methods [3]. These findings provide a strong basis for the continued use of the AnEn methodology in future weather prediction and reconstruction research.

As the AnEn method benefits from large training datasets, there is a great interest in improving its computational efficiency [121, 59]. One possible approach is to use clustering techniques (as recommended by [82]), which implies previously grouping the analogous data records, reducing the number of operations needed to identify the analogues and compute the reconstructed value (c.f. [24, 127]).

Another approach to overcome the difficulty of handling a large data set is the reduction techniques, which reduce the dimension of the original dataset without losing essential information. The Principal Components Analysis (PCA) method is widely used in multivariate statistics to reduce the dataset size, retaining only the most relevant information. In forecasting, PCA has already been applied with post-processing techniques like Neural Networks (NN) and the AnEn method, to forecast wind power and solar radiation [47].

This chapter presents an exploratory study combining the PCA with the AnEn methods into a new process (PCAnEn). First, PCA is applied to the input data to reduce the dataset's dimension. Then, AnEn is used to reconstruct the data from a station using the neighbouring station's data. This allows more stations/information, but without compromising processing time. This study was initially published in [12] and follows preliminary ideas that were presented in [13].

This chapter is organised as follows. Section 10.1 is devoted to the AnEn method and its different variants. Two alternative ways (dependent or independent) of using predictor time series are presented, and the cluster-based version of the AnEn method, in line with previous work published in [25]. Section 10.2 presents the characterisation of the dataset used and introduces the PCA technique combined with the AnEn method (PCAnEn). Section 10.3 is dedicated to applying the new PCAnEn method to the principal components, for the reconstruction of meteorological variables of a single station; the results of accuracy and computational efficiency tests are presented. Section 10.4 presents final considerations and future work directions.

10.1 Hindcasting with the AnEn Method

Classically, weather hindcasting involves applying a forecast model to a past starting point to validate the model by comparing its forecasts with available observations (*reanalysis*). If some observations are unavailable, their time series can be complemented (reconstructed) using the corresponding forecasts as a substitute for the missing observations. However, the combination of hindcasting with the AnEn method also makes it possible to reconstruct meteorological data. For example, a variable at a meteorological station can be reconstructed based on the data of correlated variables from the same station and/or other nearby stations.

The application of the AnEn method to the reconstruction of missing data, from time series of real-world observations, is illustrated by Figure 10.1, for the scenario of a single predictor station. In the figure, the *historical dataset* is a complete record of past observations of a specific meteorological variable collected at the predictor station, and the *observation dataset* is an incomplete record of the same or a correlated variable at the predicted station (this record is complete for a *training period* but incomplete or absent for a *prediction period*).

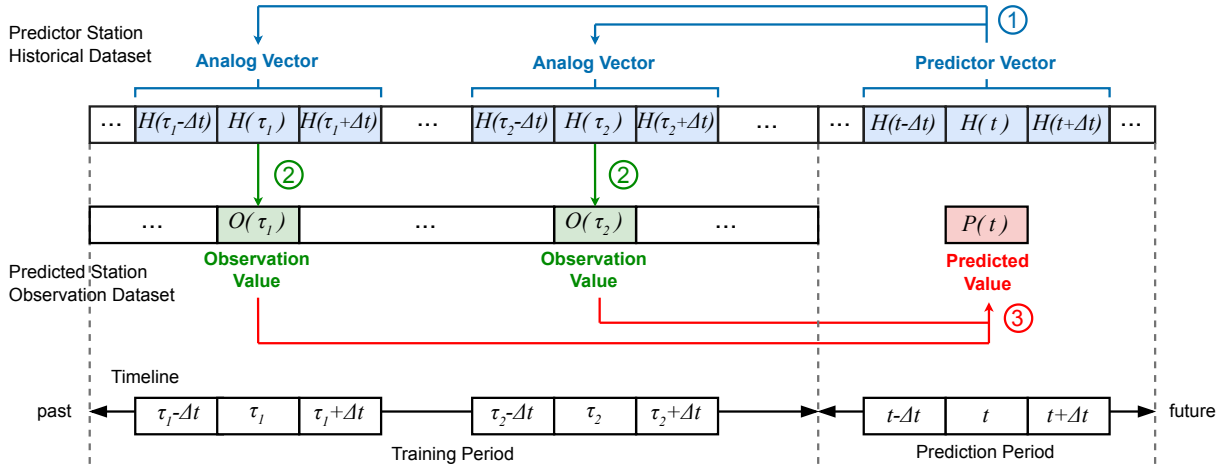


Figure 10.1: Illustration of the reconstruction of meteorological data with the Analog Ensemble (AnEn) method, considering a window size with $k = 1$.

The reconstruction of a missing value (predicted value) of the predicted station, concerning a certain instant t in time, unfolds in three steps.

In step ①, a certain number of analogues are selected from the historical dataset based on the similarity of the past observations to a predictor at instant t. The predictor and each analogue are vectors of $2\,k + 1$ elements sampled at successive intervals of Δt time-step. Each element is the value of a meteorological variable, and k is a positive integer that represents the width of each half-window (into the past and future) around the central instant of the time window ($k = 1$ in the scenario of Figure 10.1). Comparing vectors instead of single values considers the meteorological variable's evolutionary trend around the time window's central instant, allowing for the selection of analogues based on short-range weather patterns instead of single isolated events.

In step ②, the analogue map is onto observations in the training period of the predicted station. This mapping is done only for the central instant of the analogue time window, meaning that, for each analogue vector, only a single observational value is selected.

Finally, in step ③, the observations selected are used to estimate the missing predicted value. If this value is available as real observational data (assumed in this paper), assessing the prediction/reconstruction error becomes possible.

In the ensuing text, the following notation is used: H stands for the historical time series from which analogues are defined, O represents the measurements/observations time series for the feature to be predicted, and P is the outcome prediction. Whilst O and P can be viewed solely as function of time t, when using multistations the history H will be an aggregate of time-series of multiple predictor stations, in which case, it will be a function also of the station s—that is, $H = H(s,\, t)$.

The different variants of the AnEn method that are relevant in the scope of this work are presented below. The various techniques that can be used to obtain analogues in step ① of the scheme depicted in Figure 10.1 are discussed in detail. We begin with the classical error-based approaches and end with the cluster-based technique initially proposed in [25], for which a complete formal description is provided.

10.1.1 Classical Error-Based Analogues

The classical techniques are based on the score obtained through given error metrics (ϵ). These metrics differ depending on the predictor stations being used in an *independent* or *dependent* way: in the first scenario, the choice of analogues for a predictor station is

independent of the choice made for other predictor stations; in the second scenario, the analogues chosen for the different predictor stations must coincide in time.

10.1.1.1 Independent Analogues

With a single predictor station, a univariate similarity metric can be defined [86] as

$$\epsilon(t,\,\tau) = \sqrt{\sum_{r=-k}^{k} \left[H(t + r\,\Delta t) - H(\tau + r\,\Delta t)\right]^2} \tag{10.1}$$

where H is a single historical time-series from which analogues are chosen, t is a time instant in the prediction period of H, τ is a time instant defining a possible analogue in the training period of H, and Δt is the time-series time-step. Both t and τ are the central instants of time windows with $2\,k + 1$ consecutive instants that are Δt apart in time. For each time window, there is a vector with $2\,k + 1$ consecutive records of a meteorological variable. The metric $\epsilon(t,\,\tau)$ is a Euclidean distance between a possible analogue and the predictor, such that the best analogues are those that yield the lowest values (scores) of $\epsilon(t,\,\tau)$. This metric can be adapted to a scenario with multiple predictor stations at different locations, identified by an index s, each with its own time-series $H(s)$, turning into

$$\epsilon_I(s,\,t,\,\tau) = \sqrt{\sum_{r=-k}^{k} \left[H(s,\,t + r\,\Delta t) - H(s,\,\tau + r\,\Delta t)\right]^2} \tag{10.2}$$

where $s = 1,\,\ldots,\,N_s$, and N_s is the total number of predictor stations. The metric $\epsilon_I(s,\,t,\,\tau)$ allows us to identify the best analogues for each station s, independently of those found for other stations; hence, this metric is designated as an *independent* score. This is equivalent to finding, for each predictor station s, the N_a analogues with the lowest scores, i.e.,

$$a_{s,n} = \underset{\tau}{\operatorname{argmin}}\ \left[\epsilon_I(s,\,t,\,\tau)\ \text{ if } \tau \notin \{a_{s,1},\,\ldots,\,a_{s,n-1}\}\right] \qquad \text{for } s = 1,\,\ldots,\,N_s \text{ and } n = 1,\,\ldots,\,N_a\,. \tag{10.3}$$

The prediction follows from the arithmetic mean of the observations in the time series

O, corresponding to the central instant of the time window of the selected analogues:

$$P(t) = \frac{1}{N_s\, N_a} \sum_{s=1}^{N_s} \sum_{n=1}^{N_a} O(a_{s,n})\,. \tag{10.4}$$

This assumes that each predictor station has the same number of best analogues. However, it is also possible to consider, for each station, a different amount (cf. [44, 26]).

Algorithm 1 Classic AnEn for independent stations.

1. **inputs**: H, O, N_p, N_a, m and k
2. **for** $i = 1, \ldots, N_p$
3. **for** $s = 1, \ldots, N_s$
4. **for** $j = 1, \ldots, m$
5. $\epsilon_I(s,\, t_i,\, \tau_j) \leftarrow \sqrt{\sum_{r=-k}^{k} \left[H(s,\, t_i + r\,\Delta t) - H(s,\, \tau_j + r\,\Delta t)\right]^2}$
6. **endfor**
7. **for** $n = 1, \ldots, N_a$
8. $a_{s,n} \leftarrow \underset{\tau}{\operatorname{argmin}} \left[\epsilon_I(s,\, t_i,\, \tau) \text{ if } \tau \notin \{a_{s,1},\, \ldots,\, a_{s,n-1}\}\right]$
9. **endfor**
10. **endfor**
11. $P(t_i) \leftarrow \frac{1}{N_s\, N_a} \sum_{s=1}^{N_s} \sum_{n=1}^{N_a} O(a_{s,n})$
12. **endfor**

Algorithm 1 sums up the classic AnEn method with independent stations. The most demanding part from the computational point of view corresponds to the inner loops between lines 3 and 6; in them, the metric ϵ_I is calculated m times per each predictor station, where $m = N_\tau - 2k$, and N_τ is the total number of records in the training period of H, a period that may span several years (the number of subsets is $m = N_\tau - 2k$ because each one of them has dimension $2k+1$ and is constituted by the central register $H(\tau)$ plus

the k previous and the k following records, therefore not being possible to form subsets for the first k and the last k records of H); once these inner loops are repeated for each prediction i (outer loop between lines 2 and 12), the computational demand increases further, in direct proportion to the overall number of predictions, N_p. However, this algorithm is easily parallelisable due to the data independence of the iterative operations in its various loops.

10.1.1.2 Dependent Analogues

When using several predictor stations, the analogues from different stations can be forced to overlap in time, meaning there is a time dependency between the analogues. In this scenario, the score metric for the same instant τ in all time-series $H(s)$ is given by

$$\epsilon_D(t,\,\tau) = \sqrt{\sum_{s=1}^{N_s} \epsilon_I(s,\,t,\,\tau)^2}\,. \tag{10.5}$$

Overall, there are now only N_a best analogues to identify, and they are given by

$$a_n = \underset{\tau}{\operatorname{argmin}}\ [\epsilon_D(t,\,\tau)\ \text{ if } \tau \notin \{\,a_1,\,\ldots,\,a_{n-1}\}] \quad \text{ for } n = 1,\,\ldots,\,N_a\,. \tag{10.6}$$

The best analogues thus correspond to historical vectors from different stations that coincide in time and, when considered together in Formula 10.5, ensure the lowest ϵ_D values.

In turn, this translates into predictions based only on N_a observations, given by

$$P(t) = \frac{1}{N_a} \sum_{n=1}^{N_a} O(a_n)\,. \tag{10.7}$$

Algorithm 2 describes the classic AnEn method with dependent stations. Simpler, this algorithm ends up involving similar computational effort as Algorithm 1 used with independent stations, as it also requires ϵ_I to be computed $m \times N_s$ times.

Algorithm 2 Classic AnEn for dependent stations.

1. **inputs**: H, O, N_p, N_a, m and k
2. **for** $i = 1, \ldots, N_p$
3. **for** $j = 1, \ldots, m$
4. $\epsilon_D\left(t_i, \tau_j\right) \leftarrow \sqrt{\sum_{s=1}^{N_s} \epsilon_I\left(s, t_i, \tau_j\right)^2}$
5. **endfor**
6. **for** $n = 1, \ldots, N_a$
7. $a_n \leftarrow \underset{\tau_j}{\operatorname{argmin}} \left[\epsilon_D\left(t_i, \tau_j\right) \text{ if } \tau_j \notin \{a_1, a_2, \ldots, a_{n-1}\}\right]$
8. **endfor**
9. $P(t_i) \leftarrow \frac{1}{N_a} \sum_{n=1}^{N_a} O(a_n)$
10. **endfor**

10.1.2 Cluster-Based Analogue Ensembles

The search for analogues in the training period of the historic time-series H may take considerable time due to (*i*) the need to go through every instant τ_j in the training period and (*ii*) compare the vector of records centred in that instant with the predictor vector to compute the metrics ϵ_I or ϵ_D. In addition, parallelisation (an alternative fast method to define the best analogues) was achieved by employing clustering techniques. Next, this approach is described, assuming the usage of K-means clustering.

10.1.2.1 Independent K-Means Analogues

The historic time-series $H(s)$ of a station s may be broken into smaller overlapping vectors or subsets of size $2\,k+1$, such that each subset j of the station s is given by

$$\mathbf{x}_{s.j} = \{H(s,\,(j-k)\,\Delta t),\,\ldots,\,H(s,\,(j+k)\,\Delta t)\} \qquad \text{for } s = 1,\,\ldots,\,N_s \text{ and } j = k+1,\,\ldots,\,N_\tau - k \tag{10.8}$$

where N_τ is, as already stated, the dimension of the time series of the training period. There are thus $N_\tau - 2k$ subsets (vectors) $\mathbf{x}_j$ per station. The set of these subsets for a station s is

$$\mathbf{X}_s = \{\mathbf{x}_{s.j}\} = \{\mathbf{x}_{s.(k+1)},\,\ldots,\,\mathbf{x}_{s.(N_\tau-k)}\}\,. \tag{10.9}$$

The clustering method is a function f that maps $\mathbf{X}_s$ into a set of clusters $\mathbf{c}_{s.q}$, for a maximum number of clusters N_c—that is, $f : \mathbf{X}_s \to \{\mathbf{c}_{s.q}\}$, with $q = 1,\ldots,N_c$, or identically

$$f : \mathbf{X}_s \to \{\mathbf{c}_{s.1},\,\ldots,\,\mathbf{c}_{s.N_c}\} \tag{10.10}$$

where each cluster $\mathbf{c}_{s.q}$ will include a certain number of $\mathbf{x}_{s.j}$ subsets that share an aggregation criterion (for instance, minimising the coherence within each cluster as will be stated later). The aggregation of the $\mathbf{x}_{s.j}$ subsets into a cluster will thus depend on the clustering algorithm employed and the efficiency metric used. The number N_c of clusters to be formed

may be specified *a priori*, or estimated from $\mathbf{X}_s$, depending on the technique adopted.

After the application of the clustering algorithm, each cluster $\mathbf{c}_{s.q}$ will have a centroid $\overline{\mathbf{c}}_{s.q}$, corresponding to the mean of the cluster subsets (vectors). This centroid is given by

$$\overline{\mathbf{c}}_{s.q} = \frac{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} \mathbf{x}_{s.j}}{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} 1} \quad \text{for } q = 1, \ldots, N_c \quad \text{where } \overline{\mathbf{c}}_{s.q} \equiv \overline{\mathbf{c}}_{s.q}(\tau) \quad \tau \in [-k\,\Delta t,\, k\,\Delta t] \tag{10.11}$$

meaning the centroid of a cluster is a vector, and each element of that vector is given by the average of the corresponding elements in the vectors that belong to the cluster.

Each centroid vector $\overline{\mathbf{c}}_{s.q}$ acts as an individual analogue that may be compared against the historic value $H(s,\, t)$ for a prediction time t, using a metric similar to Formula (10.2):

$$\epsilon_C(s,\, t,\, \overline{\mathbf{c}}_{s.q}) = \sqrt{\sum_{r=-k}^{k} \left[H(s,\, t + r\,\Delta t) - \overline{\mathbf{c}}_{s.q}(r\,\Delta t)\right]^2}\,. \tag{10.12}$$

Having ranked all clusters of a station s by the Euclidean distance of their centroids to $H(s,\, t)$, it becomes possible to select the N_{ac} best clusters. These will be the clusters $\mathbf{c}_{s.q}$ whose centroids $\overline{\mathbf{c}}_{s.q}$ ensure the N_{ac} lowest values of $\epsilon_C(s,\, t,\, \overline{\mathbf{c}}_{s.q})$:

$$\mathbf{c}_{s,q} = \underset{\overline{\mathbf{c}}_{s.q}}{\operatorname{argmin}}\ [\epsilon_C(s,\, t,\, \overline{\mathbf{c}}_{s.q}) \text{ if } \overline{\mathbf{c}}_{s.q} \notin \{c_{s,1},\, \ldots,\, c_{s,q-1}\}] \quad \text{for } s = 1, \ldots, N_s \text{ and } q = 1, \ldots, N_{ac}\,. \tag{10.13}$$

Moreover, for each of the N_{ac} clusters selected, one may consider all its members (vectors) as analogues—the approach adopted in this work—or only the N_a best (the N_a vectors closest to the cluster centroid, based on the clustering algorithm used).

Each subset (vector) $\mathbf{x}_{s.j}$ of a cluster $\mathbf{c}_{s.j}$ has a time correspondence to the observation time-series O that can be mapped into a matching subset of observations $o_{s.j}$, given by

$$o_{s.j} = O(s, j\,\Delta t) \quad \text{for } s = 1, \ldots, N_s \text{ and } j = k+1, \ldots, N_t - k \tag{10.14}$$

It follows that each centroid $\overline{\mathbf{c}}_{s.q}$ will have an associated observation $\overline{o}_{s.q}$, which is the average of all the observations $o_{s.j}$ that matches the central time of the vectors $\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}$:

$$\overline{o}_{s.q} = \frac{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} o_{s.j}}{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} 1} \quad \text{for } q = 1, \ldots, N_{ac}\,. \tag{10.15}$$

Considering the contribution of all predictor stations, the prediction is thus given by

$$P(t) = \frac{1}{N_s\, N_{ac}} \sum_{s=1}^{N_s} \sum_{q=1}^{N_{ac}} \overline{o}_{s.q}\,. \tag{10.16}$$

In this work, a single analogue cluster is used per predictor station: $c_{s.1}$, the cluster with the best score $\epsilon_C(s,\, t,\, \overline{\mathbf{c}}_{s.q})$. Thus, the prediction with N_s predictor stations is simply

$$P(t) = \frac{1}{N_s} \sum_{s=1}^{N_s} \overline{o}_{s.1}\,. \tag{10.17}$$

The full sequence of the steps described above can be found in Algorithm 3. Clusterisation is performed only once (lines 2 to 4), for each historical dataset (one dataset per predictor station). Compared to Algorithm 1, the performance advantage of Algorithm 3 lies in the fact that the metric $\epsilon_C(s,\, t_i,\, \overline{\mathbf{c}}_{s.q})$ is computed (line 8) for several clusters N_c that is usually much smaller than the number m of vectors for which the metric $\epsilon_I(s,\, t_i,\, \tau)$ must be calculated in Algorithm 1 (line 5). Moreover, as in the classical algorithm, the main loops of the cluster-based algorithm may also be easily parallelised.

10.1.2.2 Dependent K-Means Analogues

The previous approach may be classified as independent because the historical dataset of each predictor station s is clustered autonomously. As a result, the vectors of the best clusters of each station are not required (nor expected) to be perfectly aligned in time or even to overlap. It is possible to apply clustering to enforce some temporal correlation between the vectors of different stations, as described next.

Algorithm 3 Cluster-based AnEn for independent stations.

1. **inputs**: $\mathbf{X_s}$, H, O, N_p, N_c, N_a and k
2. **for** $i = 1, \ldots, N_s$
3. $\quad \{\mathbf{c}_{s.1}, \ldots, \mathbf{c}_{s.N_c}\} \leftarrow f\left(\mathbf{X_s}\right)$
4. **endfor**
5. **for** $i = 1, \ldots, N_p$
6. $\quad$ **for** $s = 1, \ldots, N_s$
7. $\quad\quad$ **for** $q = 1, \ldots, N_c$
8. $$\quad\quad\quad \epsilon_C(s,\, t_i,\, \overline{\mathbf{c}}_{s.q}) \leftarrow \sqrt{\sum_{r=-k}^{k} \left[H(s,\, t_i + r\,\Delta t) - \overline{\mathbf{c}}_{s.q}(r\,\Delta t)\right]^2}$$
9. $\quad\quad$ **endfor**
10. $\quad\quad$ **for** $q = 1, \ldots, N_{ac}$
11. $$\quad\quad\quad \mathbf{c}_{s,q} \leftarrow \underset{\overline{\mathbf{c}}_{s.q}}{\operatorname{argmin}} \left[\epsilon_C(s,\, t,\, \overline{\mathbf{c}}_{s.q}) \text{ if } \overline{\mathbf{c}}_{s.q} \notin \{\mathbf{c}_{s,1},\, \ldots,\, \mathbf{c}_{s,q-1}\}\right]$$
12. $$\quad\quad\quad \overline{o}_{s.q} \leftarrow \frac{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} o_{s.j}}{\sum\limits_{\mathbf{x}_{s.j} \in \mathbf{c}_{s.q}} 1}$$
13. $\quad\quad$ **endfor**
14. $\quad$ **endfor**
15. $$\quad P(t_i) \leftarrow \frac{1}{N_s\, N_{ac}} \sum_{s=1}^{N_s} \sum_{q=1}^{N_{ac}} \overline{o}_{s.q}$$
16. **end**

Start by joining the vectors $\mathbf{x}_{s\cdot j}$ for the same j and different stations s. This produces a new vector $\mathbf{x}_j$ with $(2k+1)N_s$ elements:

$$\mathbf{x}_j = \bigcup_{s=1}^{N_s} \mathbf{x}_{s\cdot j} \quad \text{with } j = k+1 \ \ldots, \ N_\tau - k\,. \tag{10.18}$$

The new vector $\mathbf{x}_j$ may be viewed as a *stripe* made of *slices* $\mathbf{x}_{s\cdot j}$. The set of all *stripes* is

$$\mathbf{X} = \{\mathbf{x}_j\} = \{\mathbf{x}_{k+1}, \ \ldots, \ \mathbf{x}_{N_\tau - k}\}\,. \tag{10.19}$$

Now, the clustering f will create N_c clusters, bringing together correlated *stripes* (which implies their composing *slices* must also be correlated), into the same cluster:

$$f : \mathbf{X} \to \{\mathbf{c}_1, \ \ldots, \ \mathbf{c}_{N_c}\}\,. \tag{10.20}$$

Each cluster $\mathbf{c}_q$ will have its own centroid $\overline{\mathbf{c}}_q$. This centroid is still a vector of averages; however, with dependent stations, it has $(2k+1)N_s$ elements, being given by:

$$\overline{\mathbf{c}}_q = \frac{\sum\limits_{\mathbf{x}_j \in \mathbf{c}_q} \mathbf{x}_j}{\sum\limits_{\mathbf{x}_j \in \mathbf{c}_q} 1} \quad \text{for } q = 1, \ \ldots, \ N_c\,. \tag{10.21}$$

In turn, the predictor vectors $H(s,t)$ are also joined, by the same order the vectors $\mathbf{x}_{s\cdot j}$ were joined; thus, there is now a single unified predictor vector, $H(t)$, with $(2k+1)N_s$ elements:

$$H(t) = \bigcup_{s=1}^{N_s} H(s,t)\,. \tag{10.22}$$

The predictor vector $H(t)$ is compared against all N_c centroids, using the metric:

$$\epsilon_C(\,t,\,\overline{\mathbf{c}}_q) = \sqrt{\sum_{r=-k}^{k} \left[H(t + r\,\Delta t) - \overline{\mathbf{c}}_q(r\,\Delta t)\right]^2}\,. \tag{10.23}$$

Only the N_{ac} best clusters/centroids are selected, being the ones that minimise $\epsilon_C(\, t,\, \overline{\mathbf{c}}_q)$:

$$\mathbf{c}_q = \underset{\overline{\mathbf{c}}_q}{\operatorname{argmin}} \; [\epsilon_C(\, t,\, \overline{\mathbf{c}}_q) \text{ if } \overline{\mathbf{c}}_q \notin \{c_1,\, \ldots,\, c_{q-1}\}] \quad \text{for } q = 1,\, \ldots,\, N_{ac}\,. \tag{10.24}$$

Once a set of clusters is chosen, the selection of the observations and the calculation of the prediction proceed similarly to how it is performed for independent stations, except that now the clusters and the corresponding observations are not iterated by station:

$$\overline{o}_q = \frac{\sum\limits_{\mathbf{x}_j \in \mathbf{c}_q} o_j}{\sum\limits_{\mathbf{x}_j \in \mathbf{c}_q} 1} \quad \text{for } q = 1,\, \ldots,\, N_{ac}\,. \tag{10.25}$$

$$P(t) = \frac{1}{N_{ac}} \sum_{q=1}^{N_{ac}} \overline{o}_q\,. \tag{10.26}$$

Algorithm 4 resumes the main steps of the cluster-based AnEn method, considering dependent stations. Compared to Algorithm 3, clustering is done only once, regardless of the number of predictor stations, albeit the unification of their datasets must precede clustering. Thus, depending on these datasets' size and original organisation, such reorganisation may offset the gains of a single clusterisation process (which will also have to deal with N_s times more data).

Fewer clusters are formed than in the independent stations scenario; thus, the error metric is computed fewer times. The same applies to the number of average observations that need to be calculated once fewer centroids are considered. Algorithm 4 should be, nevertheless, faster than its classic counterpart (Algorithm 2), due to the reasons already discussed that make clusterisation approaches inherently faster than classic ones.

10.1.2.3 K-Means Clusterisation

In the cluster-based AnEn method proposed in [25], clustering is achieved using the K-means algorithm, whereby $m = (N_\tau - 2k)N_s$ historical data subsets $\mathbf{x}_j \in \mathbb{R}^{2k+1}$ are to be classified in N_c clusters. The data is organised as lines in a matrix $\mathbf{X} \in \mathbb{R}^{m\times(2k+1)}$. To describe the K-means method as proposed in [53], a partition of the subsets vectors

Algorithm 4 Cluster-based AnEn for dependent stations.

1. **inputs**: $\mathbf{X}$, H, O, N_p, N_c, N_a and k
2. $\{\mathbf{c}_1, \ldots, \mathbf{c}_{N_c}\} \leftarrow f(\mathbf{X})$
3. **for** $i = 1, \ldots, N_p$
4. **for** $q = 1, \ldots, N_c$
5. $\epsilon_C(t_i, \overline{\mathbf{c}}_q) \leftarrow \sqrt{\sum_{r=-k}^{k} [H(t_i + r\,\Delta t) - \overline{\mathbf{c}}_q(r\,\Delta t)]^2}$
6. **endfor**
7. **for** $q = 1, \ldots, N_{ac}$
8. $\mathbf{c}_q \leftarrow \underset{\overline{\mathbf{c}}_q}{\operatorname{argmin}} \left[\epsilon_C(t, \overline{\mathbf{c}}_q) \text{ if } \overline{\mathbf{c}}_q \notin \{\mathbf{c}_1, \ldots, \mathbf{c}_{q-1}\}\right]$
9. $\overline{o}_q \leftarrow \dfrac{\sum_{\mathbf{x}_j \in \mathbf{c}_q} o_j}{\sum_{\mathbf{x}_j \in \mathbf{c}_q} 1}$
10. **endfor**
11. $P(t_i) \leftarrow \dfrac{1}{N_{ac}} \sum_{q=1}^{N_{ac}} \overline{o}_q$
12. **endfor**

$\mathbf{x}_1, \ldots, \mathbf{x}_m$ in N_c clusters is denoted as $\prod = \{\mathbf{c}_1, \ldots, \mathbf{c}_{N_c}\}$, where

$$\mathbf{c}_j = \{\ell : \mathbf{x}_\ell \in \text{cluster } j\} \tag{10.27}$$

defines the set of vectors in cluster j. The centroid (arithmetic mean) of the cluster j is then

$$\overline{\mathbf{c}}_j = \frac{1}{n_j} \sum_{\ell \in \mathbf{c}_j} \mathbf{x}_\ell \tag{10.28}$$

where n_j is the number of elements in cluster j. The sum of the squared distance between the vectors and the j cluster centroid, known as *coherence*, is

$$\mathbf{Q}_j = \sum_{\ell \in \mathbf{c}_j} \|\mathbf{x}_\ell - \overline{\mathbf{c}}_j\|^2 . \tag{10.29}$$

The closer the vectors are to the centroid, the smaller the value of the *coherence* $\mathbf{Q}_j$. The quality of a clustering process can be measured as the *overall coherence*:

$$\mathbf{Q}\left(\prod\right) = \sum_{j=1}^{N_c} \mathbf{Q}_j . \tag{10.30}$$

K-means is an optimisation method. It searches for an optimal partition that minimises $\mathbf{Q}(\prod)$. The problem of minimising *overall coherence* is NP-hard [55]; consequently, there is no guarantee that it will converge to the global optimum. The basic algorithm for K-means clustering is a two-step heuristic procedure. First, each vector is assigned to the group that is closest to it. Then, new centroids are calculated based on the assigned vectors. In the K-means version of Algorithm 5, adapted from [53], these steps are alternated until the changes in the *overall coherence* are smaller than a certain predefined tolerance.

As the initial partition is generated randomly, each execution of the K-means algorithm can lead to a different solution, all corresponding to quasi-optimal solutions that verify the convergence criterion.

Algorithm 5 K-means Algorithm.

1. Start with an initial random partitioning $\prod^{(0)}$ and compute the corresponding centroid vectors $\overline{\mathbf{c}}_j^{(0)}$ for $j = 1, \ldots, N_c$. Compute $Q(\prod^{(0)})$. Set $z = 1$.
2. For each $\mathbf{x}_i$ find the closest centroid. If the closest centroid is $\overline{\mathbf{c}}_p^{z-1}$ assign $\mathbf{x}_i$ to $\mathbf{c}_p^{(z)}$.
3. Compute the centroids $\overline{\mathbf{c}}_j^{(z)}$ for $j = 1, \ldots, N_c$ of the new partitioning $\prod^{(z)}$.
4. If $\left|\mathbf{Q}(\prod^{(z)}) - \mathbf{Q}(\prod^{(z-1)})\right| <$ tolerance, stop; Else $z = z + 1$ and return to step 2.

10.2 Dimension Reduction with PCA

In this section, the Principal Components Analysis (PCA) technique is introduced and applied to reduce the dataset's dimensionality and clarify how the AnEn method is combined with the PCA technique. The section begins by characterising the meteorological dataset used and presents the correlations between the meteorological variables and stations selected.

10.2.1 Meteorological Dataset Characterization

The National Data Buoy Centre (NDBC), located in the southern United States, operates and maintains a network of data collection buoys and coastal stations, with collected data being publicly available [90]. The buoy network is spread worldwide, with the most significant numbers in North America.

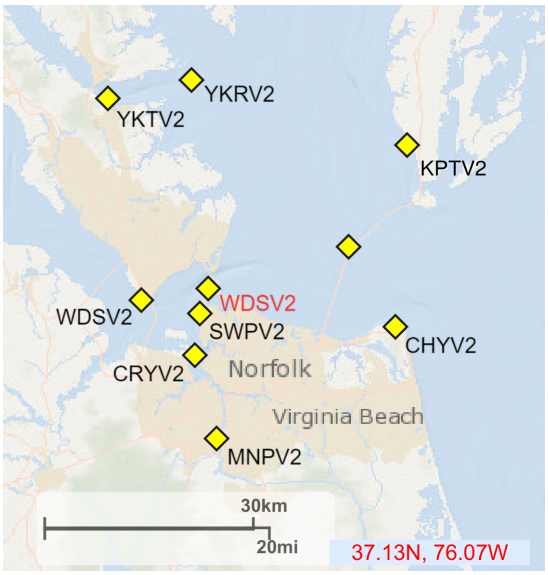


Figure 10.2: Geolocation of the NDBC meteorological stations in the region near Hampton and Newport News [90].

Figure 10.2 shows weather stations maintained by NDBC in the region near Hampton and Newport News. In this work, the predicted station is the WDSV2 station (in red). The predictor stations (in black) are within a radius of approximately 30 km from the WDSV2 station. The experiments' stations were ordered based on their proximity to the predictor station. The closest are SWPV2, CRYV2, and MNPV2, and they were used first in the test setups.

The meteorological variables available in the NDBC dataset, measured at each station, are: air pressure ($PRES$) [bar]; air temperature ($ATMP$) [°C]; wind speed ($WSPD$) [m/s] averaged over 6 minutes; peak gust speed (GST) [m/s] during the same 6 minutes period.

The characterisation of these variables is shown in Table 10.1. Variable's time-series with more than 85% data availability were selected for the analysis. However, variables have different availability at different stations. Thus, we choose different station combinations for each variable to maximise data availability.

10.2.2 Data Correlation

A correlation study between variables and stations defined the test setups. Variables and stations that are sufficiently correlated with each other can be used together in the PCA

Table 10.1: Meteorological dataset characterisation.

WSPD					***GST***			
Station	**Min**	**Mean**	**Max**	**Availab.(%)**	**Min**	**Mean**	**Max**	**Availab.(%)**
WDSV2	0.0	5.7	26.7	**97.5**	0.0	6.6	32.2	**97.5**
YKRV2	0.0	5.9	27.6	**98.0**	0.0	6.9	39.6	**98.0**
YKTV2	0.0	4.3	23.8	**97.7**	0.0	5.4	32.8	**97.7**
MNPV2	0.0	2.6	18.6	**96.4**	0.0	4.1	30.7	**96.5**
CHYV2	0.0	5.4	29.7	**95.5**	0.0	6.9	34.9	**95.5**
DOMV2	0.0	3.9	24.3	**97.5**	0.0	5.3	32.1	**97.5**
KPTV2	0.0	4.7	29.6	**97.4**	0.0	6.0	35.6	**97.5**
SWPV2	*NA*	*NA*	*NA*	0	*NA*	*NA*	*NA*	0
CRYV2	0.0	4.1	22.2	82.5	0.0	15.6	30.5	80.5
PRES					***ATMP***			
Station	**Min**	**Mean**	**Max**	**Availab.(%)**	**Min**	**Mean**	**Max**	**Availab.(%)**
WDSV2	970.1	1017.4	1044.9	**93.6**	-12.7	16.5	44.4	**87.9**
YKRV2	972.6	1017.4	1043.9	**98.6**	-12.8	15.9	36.3	**98.5**
YKTV2	974.7	1017.3	1044.3	**98.4**	-13.5	16.0	37.8	**98.2**
MNPV2	968.5	1017.5	1044.1	**97.9**	-13.8	16.8	37.3	**97.7**
CHYV2	985.2	1017.0	1042.7	31.1	-12.2	16.1	36.5	**97.0**
DOMV2	972.8	1017.8	1044.5	**98.3**	-12.6	16.1	37.2	**98.2**
KPTV2	*NA*	*NA*	*NA*	0	*NA*	*NA*	*NA*	0
SWPV2	972.0	1017.7	1044.1	**96.1**	*NA*	*NA*	*NA*	0
CRYV2	970.3	1017.6	1044.3	82.8	-10.5	16.5	36.3	34.3

technique, as more correlation allows more information to be kept in fewer dimensions.

The original multivariable historical dataset can be represented by the matrix $\mathbf{H}_0 \in \mathbb{R}^{m\times n}$, where m is the number of records of n meteorological variables:

$$\mathbf{H}_0 = \left[\mathbf{h}_0^1\, \mathbf{h}_0^2\, \cdots\, \mathbf{h}_0^n\right] \tag{10.31}$$

Each column of $\mathbf{H}_0$ includes the historical dataset of one of the n variables.

The matrix of the means of the observations for each variable is given by

$$\bar{\mathbf{H}}_0 = \left[\bar{\mathbf{h}}_0^1\, \bar{\mathbf{h}}_0^2\, \cdots\, \bar{\mathbf{h}}_0^n\right] \tag{10.32}$$

where $\bar{\mathbf{h}}_0^i \in \mathbb{R}^{m\times 1}$, for $i = 1, \ldots, n$, is a constant vector with value $\bar{\mathbf{h}}_0^i$. Including the standard deviation s_i of each variable $i = 1, \ldots, n$, in a diagonal matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ and subtracting the means and dividing by the standard deviation of each observation leads to the matrix of the scaled meteorological variables:

$$\begin{aligned}\mathbf{H} =& \mathbf{S}^{-1}\left(\mathbf{H}_0 - \bar{\mathbf{H}}_0\right)\\ =& \left[s_1^{-1}(\mathbf{h}_0^1 - \bar{\mathbf{h}}_0^1) \quad s_2^{-1}(\mathbf{h}_0^2 - \bar{\mathbf{h}}_0^2)\, \cdots\, s_n^{-1}(\mathbf{h}_0^n - \bar{\mathbf{h}}_0^n)\right]\\ =& \left[\mathbf{h}_1\, \mathbf{h}_2\, \cdots\, \mathbf{h}_n\right]\end{aligned} \tag{10.33}$$

The matrix $\mathbf{H}$ is then used to obtain the correlation matrix, given by

$$\mathbf{C} = \frac{1}{m}\mathbf{H}^T\mathbf{H} \tag{10.34}$$

where each (i, j)-entry of the matrix $\mathbf{C}$ is the correlation between the meteorological variables $\mathbf{h}_0^i$ and $\mathbf{h}_0^j$.

Figures 10.3 and 10.4 present the correlations between different meteorological variables within the same station. In stations KPTV2 and CHYV2 (Figure 10.3) there are only records of two ($WSPD$ and GST) and three ($WSPD$, GST and $ATMP$) variables, respectively. In the other stations (Figure 10.4) there are records of four variables: $WSPD$, GST, $ATMP$ and $PRES$.

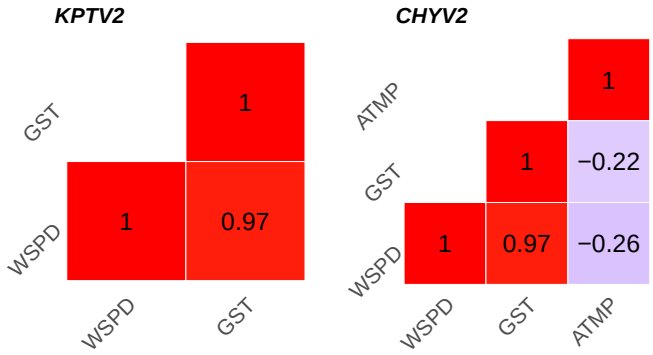


Figure 10.3: Correlation between variables at stations with less data available.

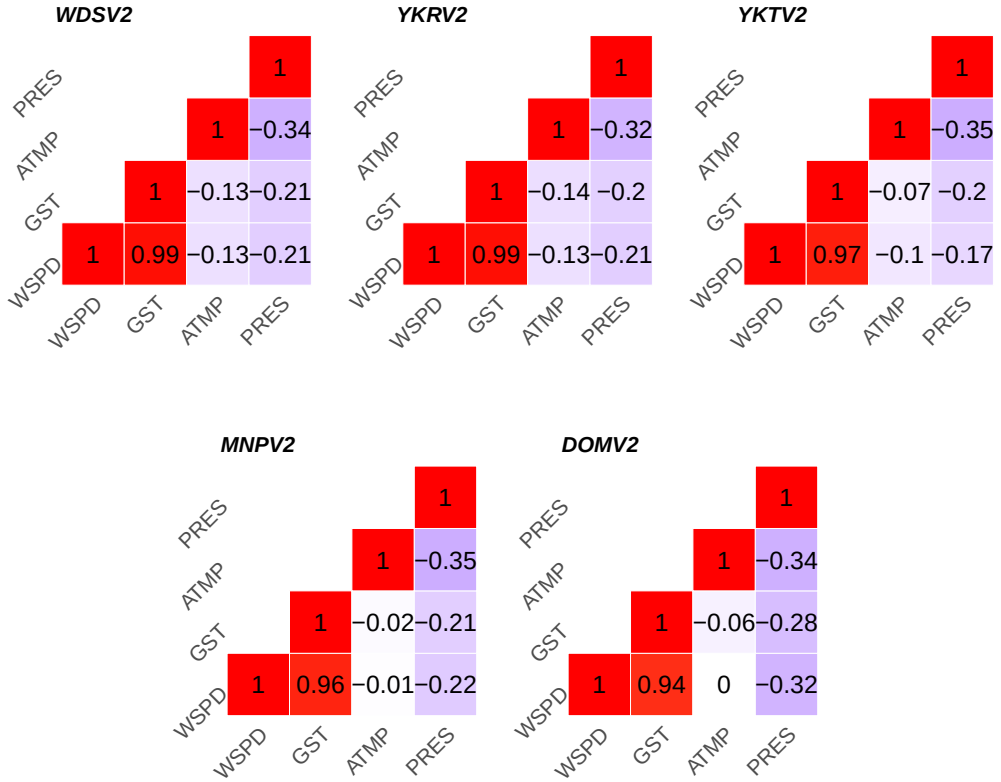


Figure 10.4: Correlation between variables at stations with more data available.

All stations have a high correlation between $WSPD$ and GST. A mild inverse correlation is observable between $ATMP$ and $PRES$. The other variable interactions showed low and inconsistent correlation among the stations.

Complementarily, Figures 10.5 and 10.6 show the correlations for the same variable between different stations. It can be observed that, with minor exceptions, correlations are high, with variables $PRES$ and $ATMP$ showing the highest correlations between various stations.

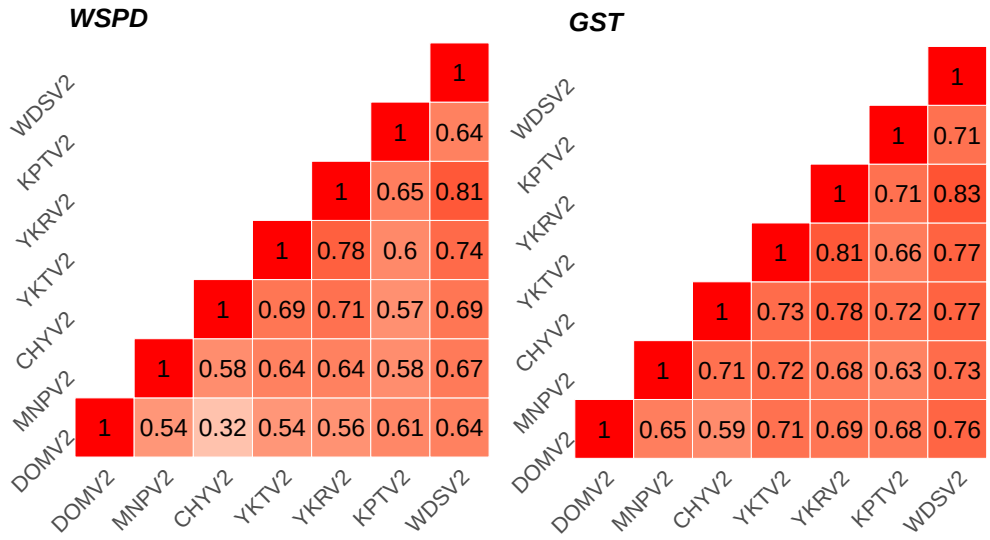


Figure 10.5: Correlation between stations for the variables $WSPD$ and GST.

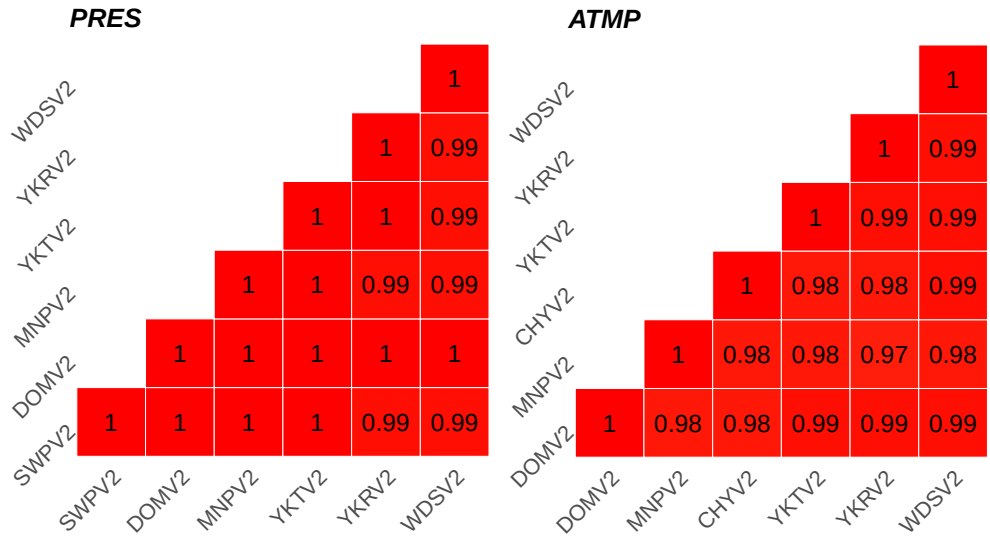


Figure 10.6: Correlation between stations for the variables $PRES$ and $ATMP$.

10.2.3 PCA - Principal Components Analysis

The PCA technique identifies the dimensions along which the data are most dispersed. In this way, we can identify the dimensions that best differentiate the dataset under analysis: its principal components. This can be achieved by the singular value decomposition (SVD) – thin version – of the scaled data matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, previously introduced in Eq. (10.34), that is given by

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{10.35}$$

where $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ (for detail see, for instance, [53]). The diagonal matrix $\mathbf{\Sigma}$ contains the singular values σ_i of $\mathbf{H}$, for $i = 1, \ldots, n$, where $\sigma_1 > \sigma_2 > \ldots > \sigma_n$. The right singular vectors v_i are the *principal component directions* of $\mathbf{H}$.

The vector

$$\mathbf{z}_1 = \mathbf{H}\mathbf{v}_1 \tag{10.36}$$

has the largest sample variance, given by σ_1^2/m, amongst all normalized linear combinations of the columns of $\mathbf{H}$. Vector $\mathbf{z}_1$ represents the first new variable called the first principal component (PC_1). The second principal component (PC_2) is $\mathbf{z}_2 = \mathbf{H}\mathbf{v}_2$ because $\mathbf{v}_2$ corresponds to the second largest variance (σ_2^2/m), and the remaining principal components are defined similarly. The new variables are linear combinations of the columns of $\mathbf{H}$, i.e, they are linear combinations of the normalized variables $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$,

$$\mathbf{z}_i = v_{1i}\mathbf{h}_1 + v_{2i}\mathbf{h}_2 + \ldots + v_{ni}\mathbf{h}_n \quad \text{for} \quad i = 1, 2, \ldots, n \tag{10.37}$$

where the coefficients v_{ji}, $j = 1, 2, \ldots, n$ (called *loadings*) are the elements of the vector $\mathbf{v}_i$. The magnitude of a coefficient is related to the relative importance of the corresponding variable in the principal component.

The substitution criterion of the original variables by a few new variables must consider the new variables' influence on the original data's variance. This influence is directly proportional to the magnitude of the corresponding singular values. The first few principal components, corresponding to the most significant singular values, are expected to account for a large proportion of the total variance, so they are all that is needed for future analyses [112].

10.2.4 Dimension Reduction of the Dataset

A decomposition into principal components (PCs) of the original meteorological variables from different stations was performed. Tables 10.2 and 10.3 show the standard deviations of each PC for various numbers of input stations. In Table 10.2, PCs were calculated from the two variables $WSPD$ and GST coming from several neighbouring stations between 2 and 6. In Table 10.3, PCs were calculated from each variable $PRES$ and $ATMP$ from several neighbouring stations ranging from 2 to 5. The variables from the WDSV2 station were not included in the original variables because WDSV2 was used only as the predicted station.

Table 10.2: Standard deviation of the PCs generated from the variables $WSPD$ and GST coming from different stations together.

Standard Deviation						
# Stations	**WSPD and GST**					
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
2	**1.771**	0.881	0.234	0.169	—	—
3	**2.059**	**1.027**	0.782	0.207	0.172	0.148
4	**2.386**	**1.030**	0.807	0.688	0.207	0.178
5	**2.689**	**1.038**	0.848	0.691	0.609	0.207
6	**2.913**	**1.047**	0.885	0.834	0.652	0.607

In Table 10.2, the PCA is performed from the data matrix that includes two meteorological variables, GST and $WSPD$, coming from different stations. This is because GST and $WSPD$ are highly correlated (recall section 10.2.2), so using them together in the PCA is possible. In this table, standard deviations above one are highlighted. When this occurs, the corresponding PC has a higher variance than the original scaled variables and, consequently, more information.

Table 10.3: Standard deviation of the PCs generated from the variables $PRES$ and $ATMP$ coming from different stations together.

Standard Deviation								
# Stations	**PRES**				**ATMP**			
	PC_1	PC_2	PC_3	PC_4	PC_1	PC_2	PC_3	PC_4
2	**1.414**	0.039	—	—	**1.408**	0.137	—	—
3	**1.731**	0.040	0.039	—	**1.721**	0.154	0.126	—
4	**1.998**	0.063	0.039	0.028	**1.987**	0.154	0.129	0.102
5	**2.233**	0.086	0.050	0.037	**2.220**	0.185	0.139	0.102

It can also be seen in Table 10.2 that, for most cases, the standard deviation is greater than 1 for PC_1 and PC_2, meaning that these two new variables concentrate most of the information contained in all the original variables (included in the data matrix $\mathbf{H}$). As in [47], we choose the PCs with standard deviations higher than 1 to represent the original dataset. For $WSPD$ and GST, PC_1 and PC_2 showed values higher than 1, except for the 2-station configuration. As expected, more components are needed to represent the original dataset by increasing the number of input stations.

Table 10.3 shows the standard deviations of the PCs computed from the $ATMP$ and $PRES$ variables. It is important to note that, unlike Table 10.2, these variables were analysed separately because they do not correlate sufficiently. The same pattern of values was observed for both variables; PC_1 was sufficient to represent the data in all configurations.

10.2.5 Combining PCA with AnEn (PCAnEn)

The AnEn method introduced in Section 10.1 can be used to reconstruct missing or incomplete data of a weather station. The reconstruction is done with data from nearby predictor stations. This method allows using more than one predictor station (or variable). In this case, the data from the predictor stations (or variables) can be used either dependently or independently (i.e., with the analogues selected in different predictor series having to overlap in time, or not – see Section 10.1.1).

The proposed approach replaces the original predictor variables with some principal components obtained through PCA decomposition. Figure 10.7 represents the AnEn classical method combined with the PCA technique, in the experimental scenario considered in this work. The combined PCAnEn method uses the time series with the chosen prin-

cipal components (PCs) to reconstruct a period (prediction period) of another time series (station WDSV2). Both time series are completed during the training period. In the prediction period, only the PCs generated by neighbouring stations have data (actually, data from the predicted station along the prediction period are also available and are used to measure the accuracy of the reconstruction method).

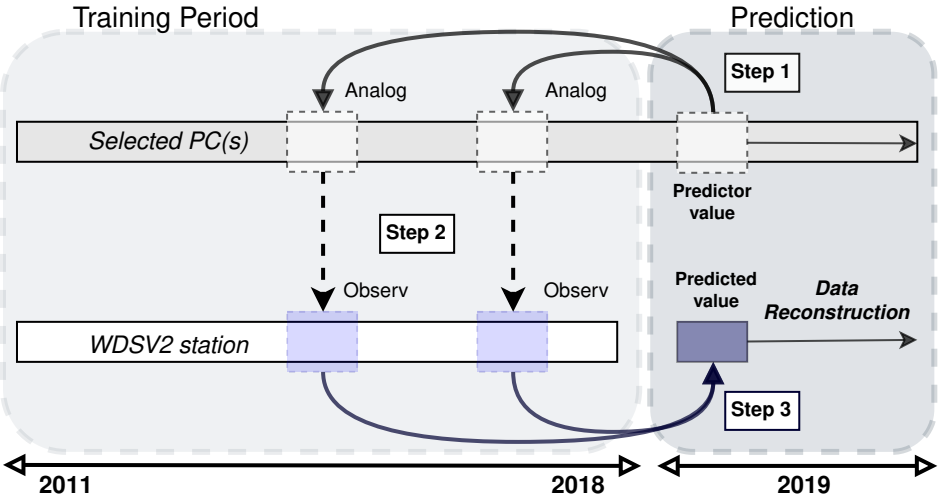


Figure 10.7: Hindcasting with the PCAnEn method.

The methodology starts by selecting analogous values according to similarity with the predictor value (step 1). Both the predictor and the analogues are vectors of $2k+1$ elements, where each element is the value of a weather variable at successive $2k+1$ instants of the same time window, and $k > 0$ is an integer representing the width of each half-window (into the past, and the future) around the central instant.

In step 2, each analogue value has a corresponding observed value, simultaneously instant as the central analogue value. The comparison of vectors, rather than single values, accounts for the evolutionary trend of the weather variable around the central instant of the time window. It thus allows the selection of analogues to consider weather patterns, rather than single isolated values. For the experiments, $k = 1$ was used; each vector represents 12 minutes (data are distributed every 6 minutes).

Finally, in step 3, the observations selected in the training period are used to predict (hindcast) the missing value, through its average, weighted or not. When this value is available as real observational data (as in this work), it becomes possible to assess the prediction/reconstruction error.

This study's available historical data ranges from 2011 to the last hour of 2018, and the reconstruction period is 2019. Because of the high resolution (6 minutes) and extensive data, we opted to make predictions only between 10 am and 4 pm, every 6 minutes. For comparaison, the results obtained with the classical AnEn methods are also presented. The experiments with the AnEn method use the original predictor data, in a *dependent* way (see Subsection 10.1.1), instead of the PCs.

10.3 Experiments with the PCAnEn Method

This section applies the PCAnEn method to a hindcasting problem with the dataset presented in Section 10.2.1. Several experiments were conducted to evaluate the effects of using principal components instead of the original historical data. The accuracy of the reconstructed values is assessed by comparison to the exact values recorded at the WDSV2 station during the prediction period. In Subsection 10.3.1, we begin by introducing the error metrics used to evaluate the accuracy of the reconstruction.

All tests were duplicated, using two different implementations of the methods, one in R [98] and another in MATLAB [78]. This provided confidence in the numerical results obtained (expected and verified to be identical) and allowed for the respective computational performance comparison. The computer system used in the experiments was a virtual machine hosted on the CeDRI/IPB virtualisation cluster, running Ubuntu 20.04.4 LTS. The resources associated with the virtual machine were 16 virtual cores of an Intel Xeon W-2195 CPU, 64GB of RAM and 256GB of SSD-based secondary storage.

The tests were divided among different numbers of predictor stations. In addition, the results obtained with the classical AnEn method applied to the original variables are also presented, allowing a comparison with the results obtained with the PCAnEn methodology. Note that for AnEn, the same stations as the PCAnEn (2-station configuration) were chosen for the prediction; the variable is predicted from the same variable located in the two closest stations to ensure the most favourable configuration of the AnEn method.

The Subsection 10.3.2 presents and discusses the accuracies obtained from the experiments. The Subsection 10.3.3 compares performance between the AnEn and PCAnEn methods, and between R and MATLAB implementations.

10.3.1 Error Metrics for Accuracy

The error metrics used to evaluate the accuracy are the Bias and the Root Mean Square Error (RMSE), as recommended by [39]. The Bias is a basic indicator of the *systematic error* in a prediction. The RMSE indicates the corresponding global error, i.e, the RMSE error includes the *systematic error* and the *random error*. In the artificial intelligence (AI) field, the *systematic error* is the *approximation error* and the *random error* is the *estimation error* [48].

As real data was available for the predicted period, it was possible to compare the predictions (p_i) with the observed values (o_i). Following Chai and Draxler [39] recommendations, multiple metrics were used to assess the model accuracy, starting with the Bias,

$$\text{Bias} = \frac{1}{N_p}\sum_{i=1}^{N_p}(p_i - o_i), \tag{10.38}$$

where N_p is the number of predictions, p_i is a prediction, and o_i is the corresponding truth value. The Bias measures the average error compared to the truth, however, allowing over- and under-predictions to cancel out. This metric can be interpreted as a rough approximation of the systematic error in the prediction. Thus, complementary to the Bias, the Root Mean-Squared Error (RMSE) is also used:

$$\text{RMSE} = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(p_i - o_i)^2}. \tag{10.39}$$

The RMSE is helpful because the squared terms give a higher weight to higher errors. Thus, the RMSE will be higher if the model makes predictions far from the truth, even if these erroneous predictions are few.

The Standard Deviation of the Error (SDE) is computed from the Bias and RMSE metrics, corresponding to:

$$\text{SDE} = \sqrt{\text{RMSE}^2 - \text{Bias}^2}. \tag{10.40}$$

The SDE represents the error due to variance; hence, it can be used as a rough approximation of the random error in the prediction.

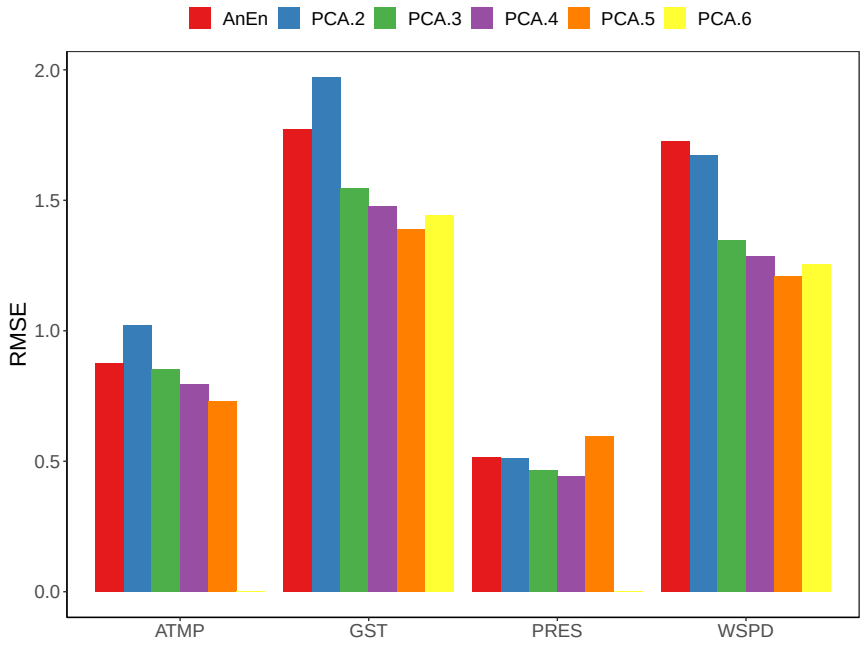


Figure 10.8: Comparison of the RMSE for different variables and number of stations.

10.3.2 Comparing Accuracy

Figure 10.8 allows for comparing the accuracy of the AnEn and PCAnEn methods, with different numbers of stations, for the four meteorological variables considered in this study. The number of PCs used for each combination is 1 or 2, as indicated in Tables 10.2 and 10.3. The chart is based on the accuracies provided by the R implementation; however, they are very similar to those of the MATLAB implementation, as may be seen in Table 10.4; this table provides the full accuracy results, including also the Bias in addition to the RMSE.

The minor errors were obtained by the PCAnEn method in the configurations with 4 or 5 stations. For the variables $ATMP$, GST and $WSPD$, five stations showed better RMSE. In contrast, for $PRES$, the 4-station configuration provided the most accurate prediction.

For instance, as shown in Table 10.4, the predictions of $WSPD$ and GST with the PCAnEn method generated 30% and 21.8% lower RMSE errors compared to the classical AnEn method, respectively. To a lesser extent, the reconstructions of the $PRES$ and $ATMP$ variables showed a reduction of 13.6% and 16.7% with the PCAnEn method. Moreover, the lowest Bias measurements were obtained with four stations in all variables.

Table 10.4: Comparison between the PCAnEn and AnEn methods.

Method	#St	Errors	R				MATLAB			
			WSPD	*GST*	*ATMP*	*PRES*	*WSPD*	*GST*	*ATMP*	*PRES*
PCAnEn	2	Bias	-0.086	-0.349	0.204	0.444	-0.085	-0.347	0.204	0.444
		RMSE	1.674	1.972	1.020	0.512	1.675	1.972	1.019	0.512
	3	Bias	-0.065	-0.327	0.179	0.367	-0.065	0.327	0.179	0.367
		RMSE	1.347	1.548	0.851	0.467	1.347	1.549	0.851	0.467
	4	Bias	**-0.038**	**-0.295**	**0.139**	**0.357**	**-0.038**	**-0.295**	**0.139**	**0.357**
		RMSE	1.287	1.478	0.795	**0.442**	1.288	1.479	0.795	**0.442**
	5	Bias	-0.048	-0.297	0.119	0.508	-0.049	-0.297	0.119	0.508
		RMSE	**1.210**	**1.388**	**0.729**	0.597	**1.210**	**1.389**	**0.729**	0.597
	6	Bias	-0.066	-0.313	—	—	-0.066	-0.313	—	—
		RMSE	1.255	1.442	—	—	1.256	1.443	—	—
AnEn	2	Bias	-0.147	-0.353	0.164	0.410	-0.147	-0.353	0.164	0.410
		RMSE	1.728	1.774	0.875	0.516	1.728	1.774	0.875	0.516

The non-*PRES* variables showed better predictions with all PCAnEn configurations than AnEn, except for the one with two stations. This is expected since these stations are the same as the ones used in the tests with AnEn. The PCs can represent the data in fewer components, but some information is always lost.

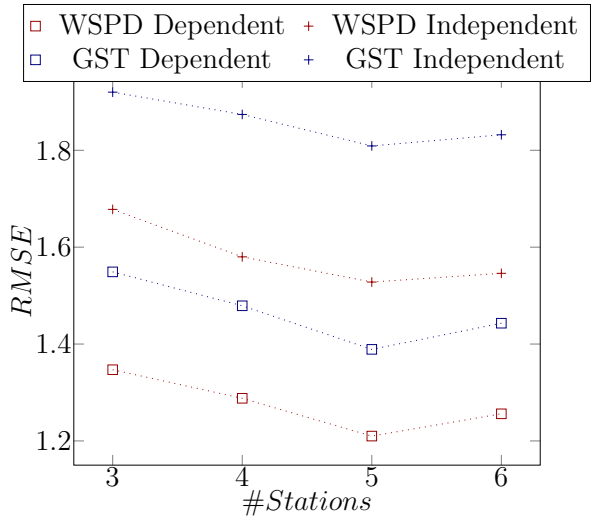


Figure 10.9: RMSE of PCAnEn used in a dependent and independent way.

Regarding the issue of *dependency*, Figure 10.9 shows the values of the RMSE obtained in the prediction of $WSPD$ and GST by the PCAnEn method used in a *dependent* and *independent* way, with three or more stations (for details on these two variants of the method see Subsection 10.1.1). The results show that an *independent* PCAnEn (PCs

used in an *independent* way) did not improve the results in any configuration or station, compared with the *dependent* version. It is also observed that increasing the number of stations up to 5 leads to a reduction in the RMSE error, but the increase to 5 stations no longer brings advantages, since the RMSE increases.

10.3.3 Comparing Performance

Figure 10.10 shows the processing times obtained by our MATLAB and R implementations of the PCAnEn and AnEN methods, with different station quantities. The processing times were measured when using 14 CPU-cores (above that number, the decrease in overall execution time was negligible – see Figure 10.11). As previously mentioned, for the AnEn method, only two predictor stations were used, so Figure 10.10 only provides two execution times (one for each AnEn implementation).

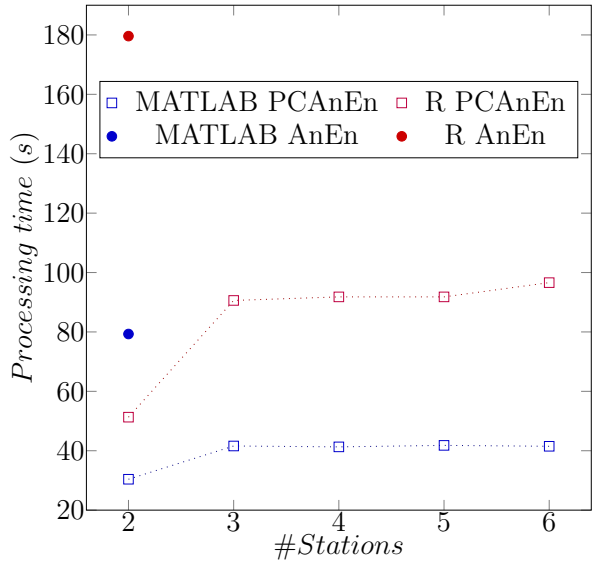


Figure 10.10: Processing time for different numbers of stations and different methods.

The PCAnEn method significantly reduces the total processing time for both implementations compared with the classical AnEn method. This is evident in the two-station scenario: using MATLAB, the PCAnEn method consumes 38% (30.4/79.3) of the time spent by the AnEn method (a speedup of 2.6x); in turn, using the R implementations, the PCAnEn method runs in 28% (51.3/179.6) of the time needed by the AnEn method (a speedup of 3.5x).

Focusing only on the PCAnEn method, the processing times vary little with the number of stations, in both implementations (the exception is the 2-station scenario, where the processing time is visibly smaller than with more stations).

The MATLAB implementation was always faster than the R implementation for any number of stations used. For instance, with six stations, PCAnEn in MATLAB was 2.3x (96.6/41.5) faster than in R, though with two stations the speedup was only 1.7x (51.3/30.4). In turn, also with two stations, AnEn in MATLAB was 2.3x (179.6/79.3) faster than its implementation in R.

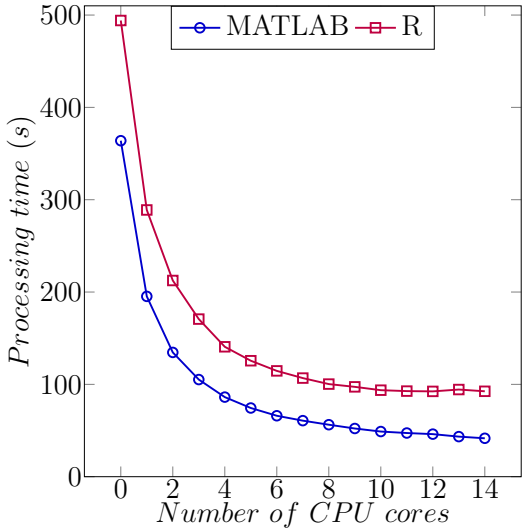


Figure 10.11: Processing time for different numbers of CPU cores used by PCAnEn.

Finally, the processing times as a function of the number of CPU cores used were also evaluated. Figure 10.11 shows the processing times for the PCAnEn method with six stations (using PCs generated from the variables $WSPD$ and GST), when varying the number of CPU-cores from 1 up to 14. It can be observed that both implementations scale reasonably well, though with diminishing returns past 8 CPU cores. Again, the MATLAB implementation offers superior performance and slightly better scalability. It should be noted, however, that MATLAB is known to be remarkably optimised to take advantage of Intel CPUs (as the one where this evaluation was performed), once it relies on the Intel MKL library.

10.4 Concluding Remarks

This chapter introduces a methodology for applying PCA to the classical AnEn method. Hence, information from several meteorological stations is combined in a reduced number of time series, corresponding to the principal components, which are then submitted to the AnEn method. In general, the location of the stations is of great importance because it promotes more correlation between variables and stations, thereby increasing the effectiveness of the PCA technique.

Combining the PCA technique with the AnEn method results in a new methodology (PCAnEn) that proved, in our experiments, to offer better hindcasting accuracy than the classical AnEn method. In the present study, the data reconstruction of the WDSV2 station using the five nearest stations seems optimal. However, the choice of predictor stations must consider the proximity and correlation between them (which needs to be assessed before determining the PCs).

Regarding computational performance, the PCAnEn method allows for a considerable reduction in processing time compared to the classical AnEn method. It was also verified that the implementation in MATLAB is faster (and by which magnitude) than the implementation in R. This information may be considered in the choice between a proprietary non-free platform and an open-source free one, to solve the same kind of hindcasting problems.

The introductory study presented in this chapter shows that the combination of PCA and EnAn has the potential to improve the reconstruction (or prediction) of time series based on a moderate number of predictor time series. However, due to the nature of the AnEn method, which requires sweeping the entire training period to determine each of the analogues, its application requires significant computational effort. Therefore, it is interesting to analyse the potential of the variant of the AnEn method with prior clustering of analogues (ClustAnEn) when applied to principal components. This issue will be addressed in the following chapter.

Chapter 11

Cluster-Based AnEn method Combined with PCA

The results from this chapter have already been published as:
Breve, M., Balsa, C., Rufino, J. (2024). *Reconstruction of meteorological records with PCA-based Analog Ensembles methods*. In: Rocha, A., Adeli, H., Dzemyda, G., Moreira, F., Colla, V. (eds) Information Systems and Technologies. WorldCIST 2023. Lecture Notes in Networks and Systems, vol. 799. Springer, Cham. ISBN 978-3-031-45641-1. ISSN 2367-3370. https://doi.org/10.1007/978-3-031-45642-8_8.

Information about past weather states is crucial to many scientific domains and practical applications. For instance, it is vital to know the historical weather data and meteorological patterns in the renewable energy field to estimate a given site's productive potential before making substantial financial investments [47]. However, complete meteorological data may not always be available or absent altogether. In this scenario, data reconstruction techniques come into play. These should be numerically accurate and computationally efficient.

As we have seen in the previous chapter, a well-known approach for meteorological data reconstruction is the Analogue Ensembles (AnEn) method. Initially, it was used as a post-processing technique to improve the accuracy of deterministic numerical forecast models [85]: past observations that are similar to the forecast are used to enhance the

accuracy of the estimates. The AnEn method can also be used directly for weather forecasting [132, 30]. AnEn method is also used to reconstruct a meteorological variable's data from other variables at the same site, or based on data from the same or different variables from neighbour locations [26].

Large training datasets (historical observations from which missing values are derived) are advantageous for the AnEn methods: the more data is available, the easier it is to capture the variation tendency of the variable(s) to be reconstructed [37, 105]. At the same time, more data entails more processing time. Hence, there is a lot of interest in improving the computational efficiency of these methods, while preserving (and ideally improving) their numerical accuracy. To this end, several variants of the AnEn methods have been investigated.

The AnEn method combined with K-means clustering (ClustAnEn), described in Subsection 10.1.2 of Chapter 10, has been successfully implemented in the reconstruction of missing gaps in time series using the values of one or two predictor time series of correlated meteorological variables [26, 7, 24]. In the study [26], the clustering of analogues with the K-means was compared with other metrics used to determine the analogues, showing that this method performs better in terms of reconstruction accuracy. In another work [25], it was also verified that the ClustAnEn method is considerably faster than the classic AnEn method. Nevertheless, reconstruction with more than two predictor time series of meteorological variables was not explored, mainly because it increases the computational costs of clustering all possible analogues and does not benefit the accuracy of the results.

In the previous chapter, we explored another approach that can be used to leverage the AnEn method, the Principal Components Analysis (PCA) technique. PCA proved to be effective in the context of the reconstruction of meteorological data when combined with the AnEn method, originating the PCAnEn hybrid method (see also [13, 12]).

The present chapter explores combining the PCAnEn method with the K-means clustering method (which results in the new PCClustAnEn variant). Thus, data from multiple predictor stations is reduced to one or two principal components (PC) that are then used (instead of the original variable records) when applying the original AnEn method or its ClustAnEn variant, to reconstruct the missing data. The PCAnEn and PCClustAnEn techniques are also compared regarding numerical accuracy and computational efficiency.

This chapter is based on the publication [35] and is organized as follows. Section 11.1

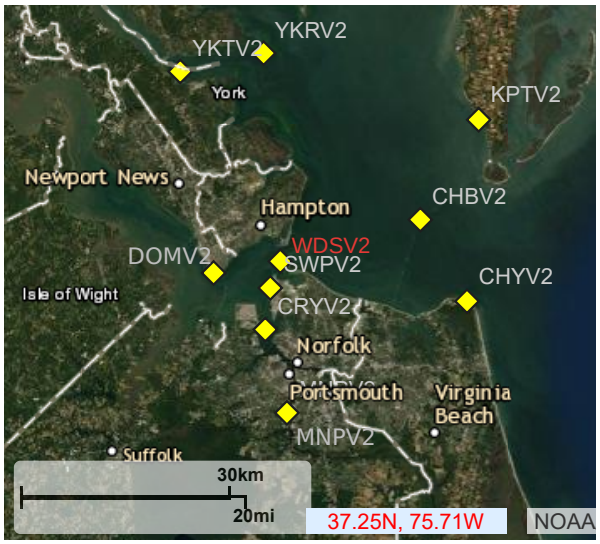


Figure 11.1: Geolocation of the meteorological stations [90].

describes the dataset used and points out the correlations between meteorological variables and stations. Sections 11.2 and 11.3 introduce the PCA technique and combine it with the AnEn and K-means methods. Section 11.4 applies the new approaches to reconstruct meteorological variables. Section 11.5 lays out final remarks.

11.1 Meteorological Dataset

The data used in this paper for the reconstruction experiments originates from the US government National Data Buoy Center (NDBC) [90]. NBDC operates a network of data-gathering buoys and coastal stations dispersed across various regions of the globe. This work uses data from multiple stations from the south of the Chesapeake Bay and surroundings, as illustrated in Figure 11.1. The predicted station is WDSV2 (red name in Figure 11.1), and the predictor stations (white name in Figure 11.1) are located within a 30 km radius from it (note: in the remainder of the chapter, the suffix 'V2' is omitted for simplicity). These stations are either at buoys or close to coastal stations at sea level, and share similar climatological conditions.

Several measurements of meteorological variables are taken at each station. These variables differ, depending on whether the station is in a buoy. For this work, a standard set of variables, available at all stations, was considered: atmospheric pressure ($PRES$)

Table 11.1: Meteorological dataset characterisation.

Station	**WSPD**		**GST**		**PRES**		**ATMP**	
	Mean	Avail.(%)	Mean	Avail.(%)	Mean	Avail.(%)	Mean	Avail.(%)
WDS	5.7	**97.5**	6.6	**97.5**	1017.4	**93.6**	16.5	**87.9**
YKR	5.9	**98.0**	6.9	**98.0**	1017.4	**98.6**	15.9	**98.5**
YKT	4.3	**97.7**	5.4	**97.7**	1017.3	**98.4**	16.0	**98.2**
MNP	2.6	**96.4**	4.1	**96.5**	1017.5	**97.9**	16.8	**97.7**
CHY	5.4	**95.5**	6.9	**95.5**	1017.0	31.1	16.1	**97.0**
DOM	3.9	**97.5**	5.3	**97.5**	1017.8	**98.3**	16.1	**98.2**
KPT	4.7	**97.4**	6.0	**97.5**	NA	0	NA	0
SWP	NA	0	NA	0	1017.7	**96.1**	NA	0
CRY	4.1	82.5	15.6	80.5	1017.6	82.8	16.5	34.3

[mbar], air temperature ($ATMP$) [°C], wind speed ($WSPD$) [m/s] and peak gust speed (GST) [m/s]. Depending on the variable, the measurements are taken every 6 minutes or correspond to averages over 6 minutes.

Some basic properties of these variables may be seen in Table 11.1, namely the global average value in the dataset and the availability, for each station, between 2010 and the end of 2019. Only variables with at least 85% of availability (in bold) were used in this study. Moreover, because some stations cannot comply with this degree of availability, the combination of stations used for each variable may differ.

To decide which variables will later be combined in the experiments, a preliminary study must be performed on the correlation between variables and stations. This is important because when variables are sufficiently correlated, the PCA method may be used to retain more data in fewer dimensions.

The correlations between the stations for each variable, relative to the same period of Table 11.1, are shown in Figure 11.2. All stations are generally correlated across all variables, with the $ATMP$ and $PRES$ variables showing the strongest correlations (1 or nearly 1).

Figure 11.3 presents the correlations between different meteorological variables within the same station. Due to the percentage of data available in the KPT and CHY stations, only records of two ($WSPD$ and GST) and three ($WSPD$, GST and $ATMP$) variables, respectively, are used for these stations. In the other stations, the records of all variables are available (see Table 11.1). The correlation between $WSPD$ and GST is strong across

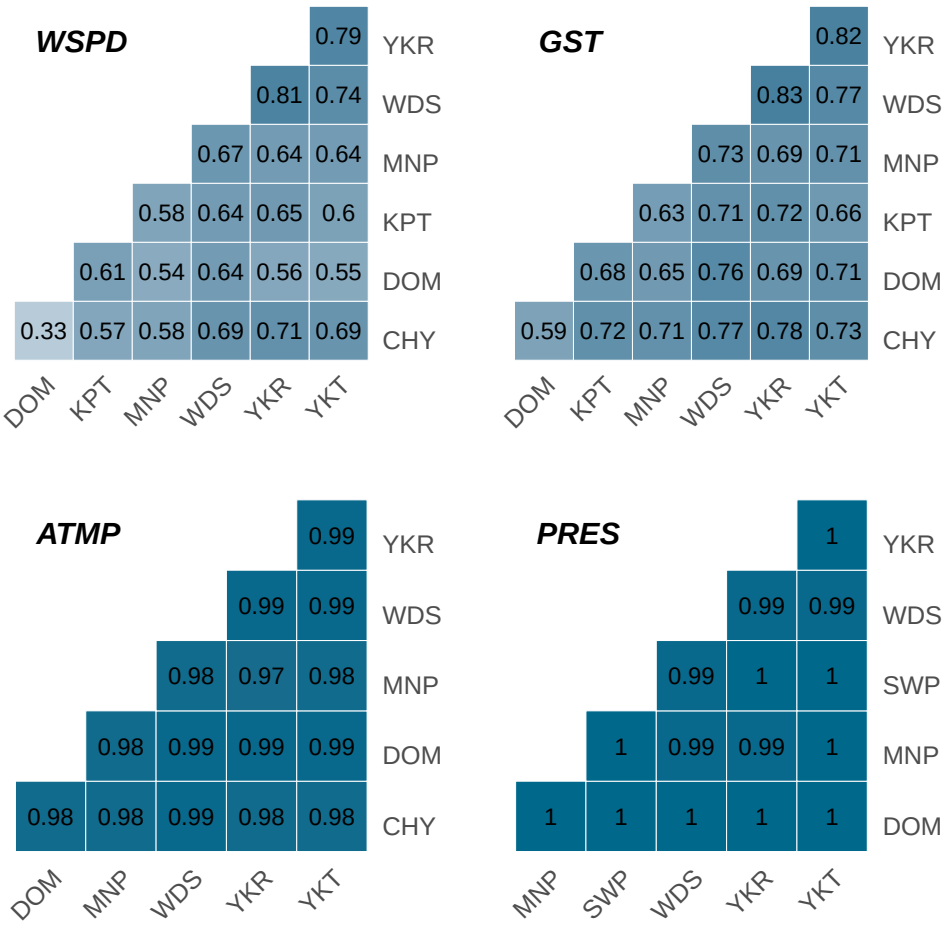


Figure 11.2: Correlation between stations for each variable.

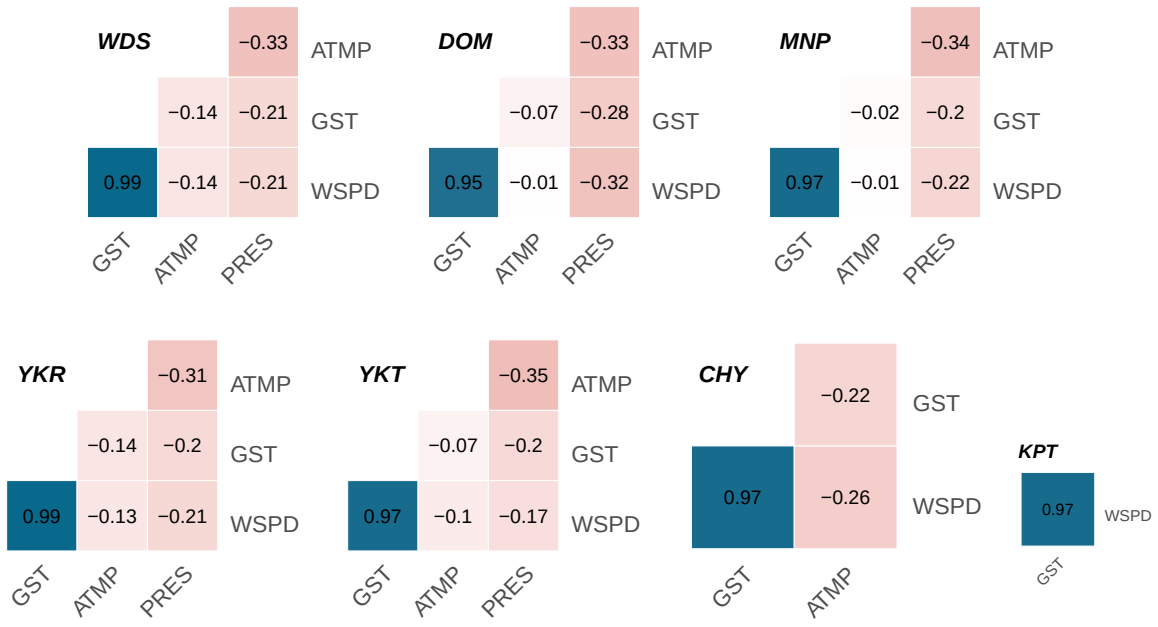


Figure 11.3: Correlation between variables at each station.

all stations. Between $ATMP$ and $PRES$, there is a minor inverse correlation. The other variable at different stations showed a weak correlation. Thus, only $WSPD$ and GST are used together in the experiments, since they are the only strongly correlated variables.

11.2 Determination of the Principal Components

Following the study on the correlation between variables and stations, the PCA approach may be applied to reduce the dimensionality of the datasets.

Firstly, the dimensions with the most data dispersion are identified. This enables identifying the principal components (PCs) that best distinguish the dataset under study. As in Section 10.2 of the previous chapter, we consider that the dataset corresponding to the multiple predictor stations is represented by the data matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, where each column $\mathbf{h}_i$, with $i = 1, \ldots, n$, includes the scaled and normalised records of a single variable. Then, the thin singular value decomposition (SVD) of $\mathbf{H}$ gives $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$. The diagonal matrix $\mathbf{\Sigma}$ contains the singular values σ_i of $\mathbf{H}$, for $i = 1, \ldots, n$, where $\sigma_1 > \sigma_2 > \ldots > \sigma_n$.

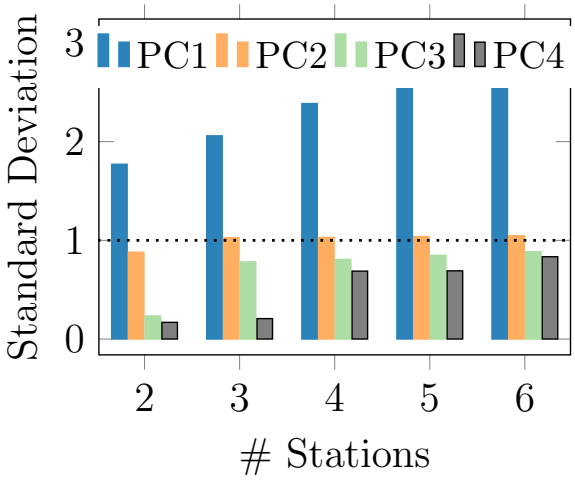


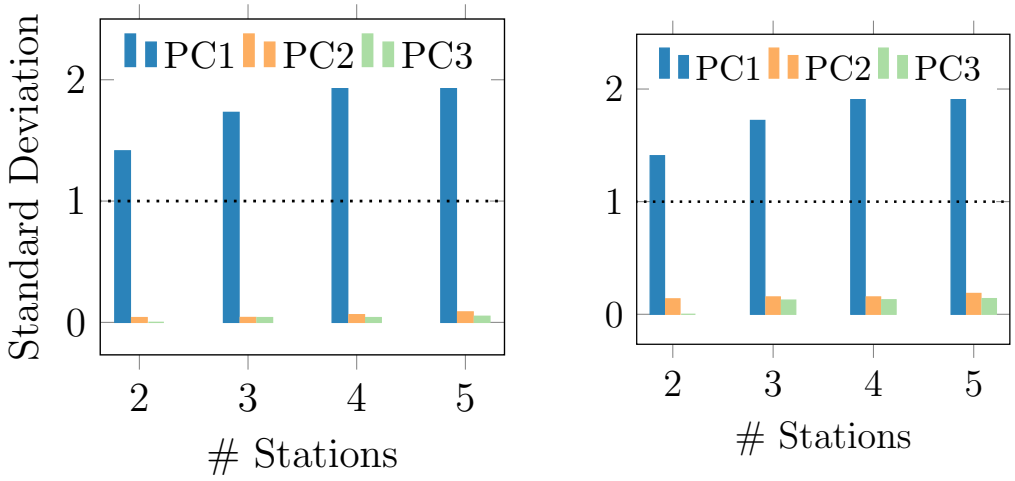
Figure 11.4: Standard deviation of the PCs from the variables $WSPD$ and GST.

The columns of matrix $\mathbf{V}$ are the right singular vectors $\mathbf{v}_j$, $j = 1, \ldots, n$, and are also called the *principal component directions* of $\mathbf{H}$. The vector $\mathbf{z}_1 = \mathbf{H}\mathbf{v}_1$ has the largest sample variance (σ_1^2/m) among all linear combinations of the columns of $\mathbf{H}$, and so $\mathbf{z}_1$ is the first principal component (PC_1). The second principal component (PC_2) is $\mathbf{z}_2 = \mathbf{H}\mathbf{v}_2$, once v_2 corresponds to the second largest variance (σ_2^2/m). The remaining principal components are defined similarly. The new variables are linear combinations of the columns of $\mathbf{H}$, that is, they are linear combinations of the normalised original variables $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$, given by

$$\mathbf{z}_j = v_{1j}\mathbf{h}_1 + v_{2j}\mathbf{h}_2 + \ldots + v_{nj}\mathbf{h}_n \quad \text{for} \quad j = 1, 2, \ldots, n \tag{11.1}$$

where the coefficients v_{ij} (called *loadings*), with $i = 1, 2, \ldots, n$, are the elements of the vector $\mathbf{v}_j$. The value of a coefficient is proportional to how significant a particular variable is in the principal component. A few of the first principal components are expected to accurately reflect the original dataset, since they are likely to account for a significant proportion of the overall variation [112].

Figures 11.4 and 11.5 show the standard deviations of each PC for different amounts of input stations (# Stations). A dotted line shows the standard deviation threshold of 1. PCs with standard deviation values above this line have more variance and, consequently, more information than the original normalised variables, whose standard deviation equals 1. Note that the variables from the WDS station were not included in the original predictor variables because WDS was used only as the predicted station.



a) Variable $PRES$. b) Variable $ATMP$.

Figure 11.5: Standard deviation of the PCs from the variables $PRES$ and $ATMP$.

In Figure 11.4, the Principal Component Analysis is conducted from a data matrix that includes two meteorological variables, $WSPD$ and GST, from distinct stations. Except for the 2-station arrangement, both $PC1$ and $PC2$ exhibited standard deviation values higher than one. These PCs with a standard deviation greater than one were selected for the training phase.

The standard deviations of the PCs computed from the $PRES$ and $ATMP$ variables are shown in Figure 11.5. It is crucial to highlight that, unlike the variables $WSPD$ and GST (see Figure 11.4), the $PRES$ and $ATMP$ variables were examined individually because they do not correlate sufficiently (see Figure 11.3). Both variables showed the same pattern of standard deviation values, indicating that the $PC1$ was enough to capture most of the data information in all station configurations. Furthermore, in contrast to the $WSPD$ and GST analysis, in the case of $PRES$ and $ATMP$ more information was contained in the $PC1$, since the relative concentration of standard deviation in this PC was significantly larger.

11.3 Analog Ensembles Combined with PCA

The AnEn method can reconstruct missing data in a time series. In this work, the time series builds on meteorological data and the reconstruction at a *predicted* station is carried

out using data from neighbouring *predictor* stations. The process starts by identifying, in each predictor station, the predictor value for the same moment in time for which the predicted value must be reconstructed. Then, past historical values in the predictor stations dataset are similar to the predictor value. These similar past values are called *analogues*. In the next step, the analogues are matched in time with corresponding observations (also historical measurements) of the predicted station. Finally, the missing value is predicted (reconstructed) by averaging the matching observations. The reconstruction error can then be evaluated if the real observed value for that instant is indeed available (as is the case in this work).

The previous simplified description considers single numerical values for the predictor and analogues. These are vectors of $2k + 1$ values (measurements), recorded at successive instants of the same time window, and $k > 0$ is an integer representing the width of each half-window (past and future) around the central instant. Therefore, analogues are established based on the similarity of vectors, not single values, which enables the selection of analogues based on similar weather trends rather than single similar values [26]. The mapping of analogues (vectors) into observations (single values) is then based on the analogues' central values.

In addition, when using multiple predictor stations, the analogues identified for each station may be required to overlap in time (dependent approach) or not (independent approach). In the current work, the first alternative was used (for details on the AnEn method and variants see Section 10.1 of Chapter 10).

When the PCA technique is combined with the AnEn method (PCAnEn), the principal components (PCs) generated from the datasets of the predictor stations are used instead of the original datasets. This allows the use of data from a larger number of predictor stations, without increasing the computational effort.

The same idea may be applied to the K-means variant of the AnEn method (see Subsection 10.1.2 of Chapter 10). This variant reduces the number of operations needed to determine the analogues of a given predictor. This is accomplished by replacing the comparison with all possible analogues with a comparison with the clusters produced using the K-means clustering method. More precisely, the comparison is made with the centroids of the clusters, whose number is much less than the total number of possible analogue vectors (see [25] for details). Thus, the PCClustAnEn method involves replacing the original predictor datasets with corresponding PCs before the clustering.

11.4 Experimental Evaluation

Both the PCAnEn and PCClustAnEn methods were put to the test for the reconstruction in the WDS station of the four variables selected for this study ($WSPD$, GST, $ATMP$ and $PRES$), every 6 minutes, from 10 am to 6 pm period, during the whole year of 2019 (prediction period). The remaining stations (within a 30 km radius around WDS, with more than 85% of data available) were used as predictor stations, considering the training period of 2011 to 2018.

The reconstruction was performed by two separate implementations of the methods, one in R [98] and another in MATLAB [78]. This allowed for the mutual verification of the numerical results and provided an opportunity to compare the implementations performance-wise. The computing system used to execute the methods was a KVM-based virtual machine (with 16 virtual cores of an Intel Xeon W-2195 CPU, 64 GB of RAM and 256 GB of SSD) hosted on the CeDRI-IPB cluster, running Linux Ubuntu 20.04 LTS, R 4.2.2 and MATLAB R2021a.

Besides testing the PCAnEn and PClustAnEn methods, the corresponding non-PCA variants (AnEn and ClustAnEn) applied to the original datasets were also tested. This way, the specific impact of the PCA technique may also be accessed. To make the comparison fair, AnEn and ClustAnEn were tested using as predictors the same two stations used to test the PCAnEn and PClustAnEn with a 2-station configuration (#Stations = 2). This means the variable is predicted from the same variable in the two closest stations, thus ensuring the most favourable configuration to the AnEn and ClustAnEn methods.

The accuracy of the predicted/reconstructed values is assessed by comparing them to the exact values recorded at station WDS during the prediction period. The comparison is done using the Root Mean Square (RMSE) error that measures simultaneously the systematic and the random error (see Subsection 10.3.1 of Chapter 10).

Table 11.2 presents the RMSE values for all tests performed. For each test, the number of PCs used was 1 or 2, after the values of the respective standard deviation, as explained in Section 11.2. There were no noteworthy changes in accuracy between PCAnEn and PCClustAnEn. The 5-station setup demonstrated a lower RMSE than the non-PCA approaches for most variables. The higher errors are obtained with the 2-station configurations, in which case, there's no sensible advantage in using the PCA variants over the

non-PCA ones. The reductions in error rates from the PCA implementations ranged from ≈18% to ≈30%, for the best setting of each variable, compared to the non-PCA methods. These considerations apply to both implementations (R and MATLAB).

Table 11.2: RMSE of the reconstruction with different methods.

Method	# Stations	**MATLAB**				**R**			
		WSPD	*GST*	*ATMP*	*PRES*	*WSPD*	*GST*	*ATMP*	*PRES*
PCAnEn	2	1.65	1.95	1.01	0.51	1.65	1.97	1.01	0.51
	3	1.32	1.52	0.84	0.48	1.32	1.52	0.84	0.48
	4	1.27	1.46	0.78	**0.45**	1.27	1.45	0.78	**0.45**
	5	**1.19**	**1.36**	**0.71**	0.61	**1.19**	**1.36**	**0.71**	0.61
	6	1.24	1.42	—	—	1.24	1.44	—	—
AnEn	2	1.68	1.77	0.86	0.59	1.67	1.76	0.86	0.59
PCClustAnEn	2	1.65	1.95	1.01	0.52	1.65	1.95	1.01	0.52
	3	1.32	1.50	0.84	0.48	1.32	1.51	0.84	0.48
	4	1.27	1.45	0.78	**0.45**	1.28	1.45	0.78	**0.45**
	5	**1.20**	**1.35**	**0.72**	0.61	**1.20**	**1.36**	**0.72**	0.60
	6	1.27	1.40	—	—	1.25	1.42	—	—
ClustAnEn	2	1.69	1.73	0.88	0.60	1.69	1.74	0.87	0.56

Regarding the computational performance, Figure 11.6 shows the processing times of the MATLAB (M) and R (R) codes, with different numbers of stations, for the reconstruction of the $WSPD$ variable, using all the CPU cores (16) available in the test bed computational system. The $WSPD$ variable was chosen for the performance evaluation because a) it is available for more stations (recall Table 11.2), and b) it requires 2 PCs to represent the original variables when using 3 or more stations. The same is also valid for the GST variable, whose processing times are either the same (PCA-based approaches) or similar (other approaches).

When using clustering (ClustAnEn and PCClustAnEn), the reconstruction times are the lowest, and the variations are minor for different numbers of stations, whether using PCA or not; also, for the only scenario where it makes sense to use the non-PCA approaches (2 stations), using clustering alone (ClustAnEn) is slower than combining it with PCA (PCClustAnEn).

The processing times are noticeably higher without clustering (AnEn and PCAnEn). When applying PCA (PCAnEn), the highest times are obtained with 3 or more stations (using 2 PCs), and they are similar; these times roughly double the time with 2 stations (using 1 PC); thus, without clustering, the number of PCs used has a noticeable influence

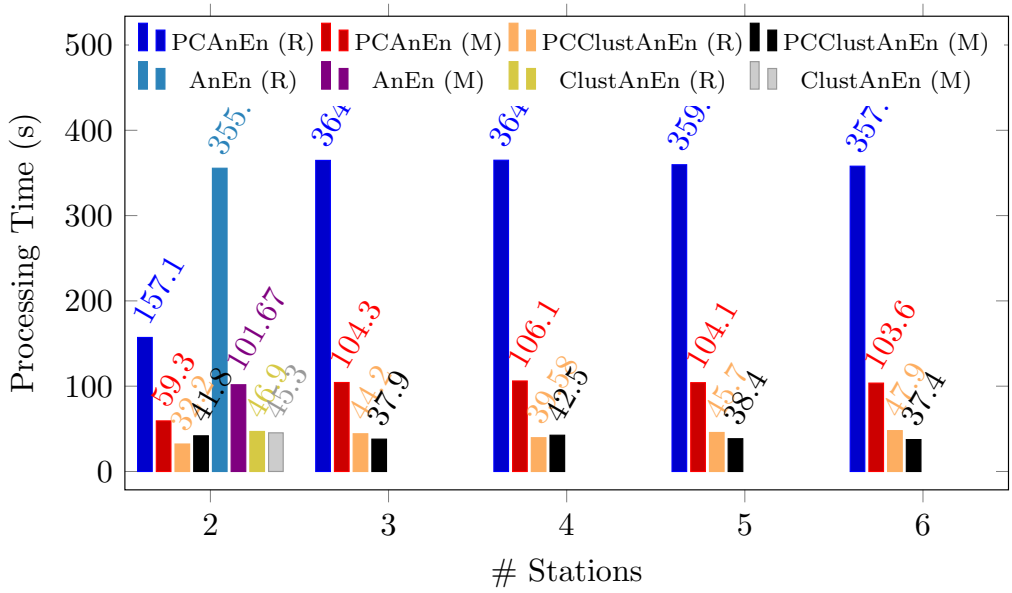


Figure 11.6: Reconstruction time of $WSPD$ with 16 cores (2 to 6 stations).

(direct proportionality) on the processing times. For the 2-stations scenario, not using PCA (AnEn) doubles the processing times compared to using PCA (PCAnEn), which is equivalent to using PCA with more than 2 stations, once it uses two time series.

Lowering the processing times is essential, but it shouldn't be at the expense of higher reconstruction errors. Ideally, the reconstruction should be faster and also more accurate. The smallest RMSE errors for the $WSPD$ variable are obtained with PCA-based methods using 5 stations, whether clustering is used (PCClustAnEn) or not (PCAnEn) – recall Table 11.2. However, clustering ensures much lower processing times, with a speedup between $\gtrapprox$2,7 (MATLAB code) and $\gtrapprox$7,8 (R code). Comparing the processing times of PCClustAnEn with 5 stations, with the ones of ClustAnEn with 2 stations (the best provided by not using PCA) yields almost no speedup (46,9/45,7=1,03 and 45,3/38,4=1,18); however, the RMSE error of PCClustAnEn with 5 stations is only 1,2/1,69≈70% of the error of ClustAnEn with 2 stations, thus favouring the first approach.

The impact on performance of using or not using the PCA method is perceivable in the 2-stations scenario. Here, PCA provides speedups ranging from 2,26 to 1,08, for comparable methods (AnEn vs PCAnEn, and ClustAnEn vs PCClustAnEn).

Another advantage of adding PCA emerges when two variables, like $WPSD$ and GST, are used together in the analysis. Once they share the same time series, PCA-based

methods can predict both variables in a single run, unlike the non-PCA approaches, which would require two runs of the reconstruction code.

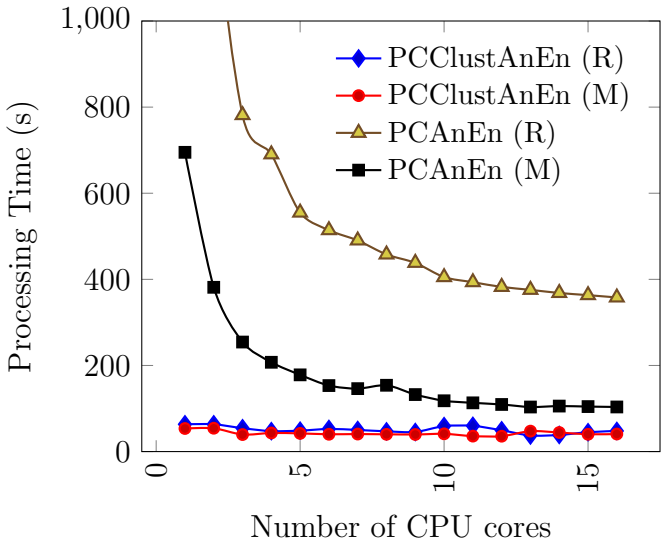


Figure 11.7: Reconstruction time of $WSPD$ with 1 to 16 cores (6 stations).

The MATLAB code was found consistently faster than the R code. This is visible in Figure 11.6 for 16 cores, and can also be seen in Figure 11.7 for a variable number of cores. However, under PCClustAnEn, the differences were minor, meaning both implementations are equally efficient when applying the K-means clustering. More importantly, PCClustAnEn requires much less processing time in all configurations than PCAnEn. Also, PCClustAnEn primarily doesn't benefit from the extra cores, contrary to the PCAnEn method, where searching for analogues is the most significant code hotspot and is easily parallelisable.

It should also be stressed that R and MATLAB were used with default configurations, without any extra performance tuning to optimise their behaviour.

11.5 Concluding Remarks

This chapter describes the combination of Analog Ensembles (AnEn) and the clustering methods with Principal Component Analysis (PCA), allowing for data from several stations to be reduced to a smaller number of time series, corresponding to the Principal

Components (PCs). Instead of the original variable records, these are then used to reconstruct data missing in the records of a meteorological site.

In our experiments, the PCA technique improved prediction accuracy without compromising the computational performance, since it is possible to increase the number of stations without increasing the quantity of input time series. It was also shown that the efficacy of PCA is heavily influenced by the correlation between the time series of several predictors, as higher correlation allows for a high proportion of information/variance in the first components.

Furthermore, two different implementations of the methods studied were used and compared, one in MATLAB and the other in R. This allowed us to double-check the numerical results and gain insight into the potential performance impact of choosing either implementation. The scalability of both codes was also studied in a medium-scale multicore system. The performance evaluation showed the superiority of the AnEn methods, where PCA is combined with clustering.

Chapter 12

AnEn method with Dimension Reduction of the Predictor Dataset

The results from this chapter have already been published as:
Balsa, C., Breve, M.M., Rodrigues, C.V., Rufino, J. (2023). *Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset.* Computation, 11(5): 98. https://doi.org/10.3390/computation11050098.

The study presented in this chapter, based on one [14], enriches the AnEn method with techniques that take advantage of many predictor variables through dimension reduction. We include reducing the original predictor dataset to a few new predictor variables without losing essential information. This approach improves the quality of the reconstructions as well as their computational efficiency. We explore the dimension reduction through two alternative methods: principal component analysis (PCA) and partial least squares (PLS).

The PCA technique identifies the data's most dispersed dimensions, i.e., with the most significant variance (see for instance [63]). In this way, we can identify the dimensions that best differentiate the dataset under analysis, i.e., its principal components, which are used as the new predictor variables. The PLS technique extracts a set of latent (not directly observed or measured) variables from the set of predictor variables with the best predictive power. These new predictor variables are obtained by maximising the covariance between the predictors and the predicted variable (see, for instance, [133]).

We present the combination of the AnEn method with PCA (PCAnEn) and PLS (PLSAnEn), to take advantage of the potentialities of the AnEn method and the dimension reduction provided by the PCA and PLS techniques. Furthermore, PCA and PLS are usually combined with multivariate linear regression, giving rise to the principal components regression (PCR) and partial least squares regression (PLSR) methods. We present a comparative study of the performance of all these methods in a hindcasting problem, corresponding to the reconstruction of missing data in a given meteorological station using data from a set of predictor stations with different geographical locations.

The chapter is organised as follows. Section 12.1 presents the various reconstruction methods employed in this study. Section 12.2 introduces the meteorological datasets used. Sections 12.3 and 12.4 focus on selecting principal components and latent variables. Section 12.5 presents the numerical results of the tests performed with the various reconstruction methods. Section 12.6 analyses the computational performance of the same techniques. Finally, Section 12.7 concludes the chapter, summarising the main findings and their implications for solving hindcasting and forecasting problems with a high number of predictors.

12.1 Reconstruction Methods

This section briefly describes all the reconstruction methods used in this work. It begins by presenting the analog ensemble method, the foundation for the other hindcasting methods. This is followed by introducing dimension reduction methods, using PCA and PLS, and their use to reconstruct missing values.

12.1.1 Analog Ensemble Method

In this study, the AnEn method is used to reconstruct meteorological records missing in a time series Reconstruction of missing values in a time series is a problem equivalent to hindcasting, i.e., to predict past events recurring to a forecast method and historical data (e.g., measurements at some other location or from another variable). The data are reconstructed based on one or more predictor time series that present some correlation with the incomplete series to be reconstructed/predicted.

A practical application scenario consists of reconstructing data from a meteorological station using data from neighbouring stations. In this context, several time series are used as predictors, being represented by the column vectors

$$\mathbf{x}_j \in \mathbb{R}^n, \quad \text{with} \quad j = 1, 2, \ldots, q, \tag{12.1}$$

each one contains n records of the values of certain meteorological variables. For simplicity, these vectors will often be referred to as predictor variables.

The predictor variables can be used in a *dependent* or *independent* way. In the dependent variant, the analogues selected in different predictor variables must be concomitant (overlapping) in time. In the independent version, such is not mandatory. Previous work [26] verified that the dependent version of the AnEn method is more accurate. Therefore, from now on, that is the version assumed to be used, unless otherwise stated.

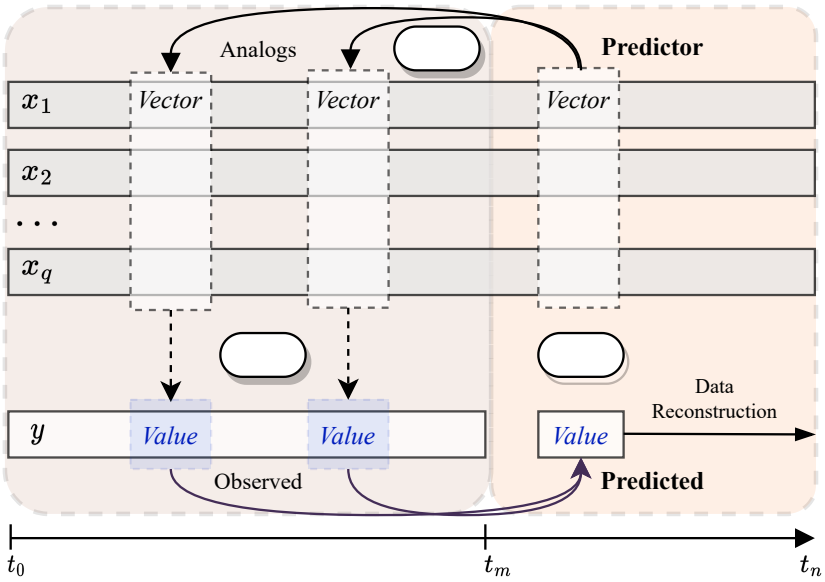


Figure 12.1: Reconstruction of missing meteorological records with the AnEn method.

Figure 12.1 illustrates the AnEn method with q predictor variables. The historical data is complete in the predictor variables and incomplete in the reconstructed/predicted one ($\mathbf{y} \in \mathbb{R}^m$). The period of missing records is denoted as the *reconstruction period*, but, often, it is also designated as *prediction period*. This designation originates from applying the AnEn method to post-processing meteorological forecasts, in which the predictor series contains the history of forecasts. In this work, the reconstruction period corresponds to the

part of the time series in which the records are reconstructed (or, by analogy, predicted). The period for which all series contain complete data is known as the *training period*. The longer the training period (compared to the prediction period), the better the AnEn method is expected to perform (the more comparison data, the more likely it is to find meteorological conditions similar to those sought).

As depicted in Figure 12.1, firstly (step 1), a certain number of analogs are selected in the training period of the predictor variables, due to being the past records most similar to the predictor record at instant $t_P \in \{t_m, ..., t_n\}$. To evaluate the analogs, a time window is defined that encompasses the predictor record at time t_P and its k consecutive neighbors in the recent past ($t_P - \Delta t$, $t_P - 2\Delta t$, ..., $t_P - k\Delta t$) and immediate future ($t_P + \Delta t$, $t_P + 2\Delta t$, ..., $t_P + k\Delta t$); these $2k + 1$ records make up a predictor vector. Next, the same kind of time window is defined for every instant in the training period, $t_T \in \{t_0, ..., t_m\}$, with a corresponding training vector; the comparison of all the training vectors with the predictor vector makes it possible to assess their similarity (see [26] for similarity metrics); the training vectors most similar to the predictor vector form the *analog ensemble* (AnEn). Note that comparing vectors, instead of single values, accounts for the evolutionary trend of the meteorological variable around the central instant of the time window, allowing for the selection of analogues to consider weather patterns (instead of single isolated values). A range of $5 \leq k \leq 10$ was reported to enhance prediction accuracy [7]; thus, this study employed $k = 5$ to reduce computational demands while attaining optimised predictions. For the datasets under analysis in this work, the resulting time window corresponded to one hour, as the time series had a sampling period of 6 minutes.

In step 2, the analogues are mapped onto observations of the predicted station by selecting the simultaneous records in the observed time series. This mapping is conducted only for the central time of each analogue time window, i.e., for each analogue vector, a single observed value is selected in the training period.

Finally, in step 3, the observed values chosen are used to predict (reconstruct) the missing values in the predicted variable $\mathbf{y}$, through its average (weighted or not). When this target value is available as real observational data (as it happens in this work), it becomes possible to compute the error of the reconstruction/prediction and, consequently, to validate the method.

12.1.2 ClustAnEn Method

The extension of the training period influences the performance of the AnEn method. The longer the training period, the more accurate the predictions/reconstructions are expected. On the other hand, more extended training periods imply greater computational effort to identify the analogues in each reconstruction. To alleviate this problem, an alternative version of the AnEn method was developed in which all possible analogs (all of the time windows that one can make from the predictor variables in the training period) are previously classified into a predefined number of clusters (see Subsection 10.1.2, from Chapter 10 and [24, 12]), with the number of clusters set to the square root of the total number of possible analogs. This heuristic is based on the empirical results previously obtained [7]. In this way, a predictor vector is compared only with the centroid (median of the analogues in that cluster) of each cluster to identify the *analog cluster* that contains the analogues selected in step 1 (see Figure 12.1). This AnEn variant, denoted ClustAnEn, is much faster than the classic AnEn method in identifying the analog ensemble [12].

Another critical data reconstruction/prediction issue is the number of predictor variables. It is expected that the more predictors there are, the more helpful information for the reconstruction is available. Hence, knowing how reconstruction/prediction methods can use all available predictive information is essential. As the AnEn and ClustAnEn methods lose computational efficiency with the increase in the number of predictor variables, using these methods is not suited for a large number (q) of predictor variables. Therefore, the main objective of this work is to reduce the dimension of the predictor dataset while minimising the loss of information by using the AnEn and ClustAnEn methods in hindcasting and forecasting problems with a high number of predictor variables.

12.1.3 Principal Components Analysis

The Principal Components Analysis (PCA) technique makes it possible to reduce the dimensionality of a dataset consisting of many interrelated variables, while retaining as much of the information of variation present in the dataset as possible. This is achieved by transforming to a new set of uncorrelated variables, called the principal components (PCs), which are ordered so that the first few contain most of the variation information in the original variables dataset [63]. We briefly describe the application of PCA to dimension reduction in predictor variables.

Let the original dataset of predictor variables be represented by the matrix

$$\mathbf{X} = [\mathbf{x}_1\, \mathbf{x}_2\, \cdots\, \mathbf{x}_q] \in \mathbb{R}^{n\times q}, \tag{12.2}$$

where predictor variables are represented by the q column vectors $\mathbf{x}_j$, with $j = 1, \ldots, q$, each one with n records of the value of a given meteorological variable. The matrix $\mathbf{X}$ is assumed to be centred, i.e., the mean of each column is equal to zero, and standardized such that its variance equals unity. To identify the dimensions along which the data are most dispersed, i.e., the dimensions that best differentiate the predictor dataset, it is necessary to compute the principal component (PC) vectors. This can be achieved by the thin singular value decomposition (SVD) of the predictor matrix $\mathbf{X}$, given by

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{12.3}$$

where the columns of the matrix $\mathbf{U} \in \mathbb{R}^{n\times q}$ contain the left singular vectors, the diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{q\times q}$ contains the singular values σ_i, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q \geq 0$ and the matrix $\mathbf{V} \in \mathbb{R}^{q\times q}$ contains the right singular vectors $\mathbf{v}_j$, with $j = 1, \ldots, q$, which are the *principal components directions* of $\mathbf{X}$. The matrices of the left and right singular vectors are orthonormal, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{X}$ is assumed to be centred, i.e., the mean of each column is equal to zero, and standardised.

The vectors

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j, \quad \text{with} \quad j = 1, \ldots, q, \tag{12.4}$$

are the original dataset's principal components (PCs) and define new variables that will be used instead of the original predictor variables. The first principal component, $\mathbf{z}_1$, has the largest sample variance, equal to σ_1^2/n, among all normalised linear combinations of the columns of $\mathbf{X}$ (for details see [53]). The second principal component, given by $\mathbf{z}_2 = \mathbf{X}\mathbf{v}_2$, is the new variable with the second largest variance (σ_2^2/n). Likewise, the remaining principal components define new variables with decreasing variances.

The new variables $\mathbf{z}_j$ are linear combinations of the columns of $\mathbf{X}$, i.e., the original predictor variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$, being given by

$$\mathbf{z}_j = v_{1j}\mathbf{x}_1 + v_{2j}\mathbf{x}_2 + \ldots + v_{qj}\mathbf{x}_q, \quad \text{with} \quad j = 1, \ldots, q, \tag{12.5}$$

where the coefficients v_{ij}, with $i = 1, 2, \ldots, q$, designated as *loadings*, are the elements of

the vector $\mathbf{v}_j$. The magnitude of a coefficient is related to the relative importance of the corresponding original variable in the principal component.

The substitution criterion of the original predictor variables, $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_q$, by p PCs, $\mathbf{z}_1$, $\cdots$, $\mathbf{z}_p$, with $p < q$, in the AnEn or ClustAnEn methods, must take into account the influence of the new variables in the original dataset. This influence is directly proportional to the variances given by σ_i^2/n, with $i = 1, 2, \ldots, q$. The first few principal components, corresponding to the most significant singular values, are expected to account for a large proportion of the total variance, all needed to describe the original dataset [112]. Therefore, one of the possible criteria that can be used to choose how many PCs should be used is the magnitude of the respective singular values. If the original variables are previously scaled, by dividing each variable by the respective standard deviation, each of them will have a standard deviation equal to one. Suppose a PC has a standard deviation greater than one. In that case, it means that it contains more information than any of the original variables and, as such, should be chosen to represent the original dataset.

In the exploratory study on the use of the AnEn method based on the principal components of a meteorological dataset presented in Chapter 10 (see also [13]), it was verified that the efficiency of this combination strongly depends on the correlation between the predictor variables. It is impossible to reduce them to a few components without losing significant information if they are poorly correlated. On the contrary, if the predictor variables are correlated, they can reduce their dimension to a few components and improve the prediction quality with the AnEn method.

Combining the AnEn and ClustAnEn methods with PCA gives rise to two new methods designated in this work by PCAnEn and PCClustAnEn, respectively.

12.1.4 Principal Component Regression

As an alternative to the reduction in the size of the predictor dataset, the reconstruction of missing data can be accomplished using multivariate linear regression. Unlike the AnEn method, this method allows for the direct use of all the original predictor variables.

The goal of the multivariate regression is to predict $\mathbf{y}$ from $\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{m \times q}$ contains in the columns the values of all the predictor variables recorded during the training period, and $\mathbf{y} \in \mathbb{R}^m$ the corresponding values of the predicted variable. This problem involves the

determination of the vector $\mathbf{b} \in \mathbb{R}^q$, that is, the approximated solution of the linear system of equations

$$\mathbf{X}\mathbf{b} \approx \mathbf{y}. \tag{12.6}$$

Such is equivalent to solving the linear least squares problem

$$\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|, \tag{12.7}$$

where $\|.\|$ is the usual 2-norm (see [53] for details). If $\mathbf{X}$ is a full rank column matrix, then the solution of the problem (12.7) is given by

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{12.8}$$

The expectation is that the solution vector $\mathbf{b}$ can be used to predict values in the reconstruction period based on the predictor variables for that same period, that is:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b}, \tag{12.9}$$

where $\tilde{\mathbf{y}} \in \mathbb{R}^{(n-m)}$ represents the reconstructed/predicted variable during the reconstruction/prediction period and $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-m)\times q}$ contain the values of the predictor variables along the same period.

The multivariate regression model given by Equation (12.9) can be implemented only if the matrix $\mathbf{X}$ has full column rank (its column vectors are linearly independent). The near-collinearity of columns can occur if there are highly correlated predictor variables. In this case, the least squares problem (12.7) becomes ill-conditioned and challenging to solve.

The principal component regression (PCR) [76] method circumvents the rank deficiency by replacing the original predictor variables $\mathbf{X}$ by their principal components (PCs) in the regression model. Once the principal components, $\mathbf{Z} = \mathbf{X}\mathbf{V}$, are obtained from matrix $\mathbf{X}$ in the same way as described in Section 12.1.3, a few of them (p) are used in the regression model to estimate $\mathbf{y}$.

Therefore, the PCR method consists of regressing $\mathbf{y}$ not on $\mathbf{X}$ itself but on the principal components matrix $\mathbf{Z} \in \mathbb{R}^{m\times p}$, assuming that p PCs have been previously selected. This

implies that linear least squares can solve it.

$$\mathbf{Z}\mathbf{c} \approx \mathbf{y}, \tag{12.10}$$

whose solution is the parameter vector

$$\mathbf{c} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}. \tag{12.11}$$

The PCR regression model on $\mathbf{Z}$ is then given by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\mathbf{c}, \tag{12.12}$$

were $\tilde{\mathbf{y}} \in \mathbb{R}^{(n-m)}$ is, as before, the vector of the reconstructed/predicted values of $\mathbf{y}$ during the reconstruction/prediction period, $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{V} \in \mathbb{R}^{(n-m)\times p}$ contains the values of the p selected PCs along the reconstruction/prediction period and $\mathbf{c} \in \mathbb{R}^p$ is the parameter vector of the PCR.

The regression model (12.12) can be expressed in function of $\tilde{\mathbf{X}}$ instead of $\tilde{\mathbf{Z}}$ by replacing $\tilde{\mathbf{Z}}$ by $\tilde{\mathbf{X}}\mathbf{V}$, with $\mathbf{V} \in \mathbb{R}^{q\times p}$, thus originating

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{V}\mathbf{c}. \tag{12.13}$$

PCR is an alternative to the AnEn-based methods that combines the size reduction provided by PCA with linear regression. This combination prevents collinearity problems between vectors of predictor variables. Another advantage of PCR is reducing the number q of original predictor variables to a lower number p of principal components containing most original information. Thus, the good performance of this method depends strongly on the choice of the PCs.

It is expected that a few of the PCs, which have a higher variance, are enough to describe the evolution of the original predictor dataset. However, these components were chosen to explain the evolution of the original predictor variables, contained in $\mathbf{X}$, and, as such, there is no guarantee that these PCs will be relevant for the prediction of $\mathbf{y}$.

12.1.5 Partial Least Squares Regression

In contrast with the PCR method, the partial least squares regression (PLSR) method uses the components from $\mathbf{X}$ that best predict $\mathbf{y}$. These components, also called the latent variables (because they are not directly observed or measured), are coming from the joint decomposition of $\mathbf{X}$ and $\mathbf{y}$, taking into account the obligation of the components to explain the covariance between $\mathbf{X}$ and $\mathbf{y}$ as best as possible [1, 83]. Since the predictor and predicted/reconstructed time series have the same size, the predictor variables include just records from the training period, whose size is m.

PLSR computes the latent variables that model $\mathbf{X} \in \mathbb{R}^{m\times q}$ and $\mathbf{y} \in \mathbb{R}^{m}$ and best predict $\mathbf{y}$, resulting in the variable decompositions

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \text{and} \qquad \mathbf{y} = \mathbf{R}\mathbf{q}^T + \mathbf{f}, \tag{12.14}$$

where $\mathbf{T} \in \mathbb{R}^{m\times p}$ and $\mathbf{R} \in \mathbb{R}^{m\times p}$ are the matrix with p latent vectors (also known as scores) extracted from $\mathbf{X}$ and $\mathbf{y}$, respectively; $\mathbf{P} \in \mathbb{R}^{q\times p}$ and $\mathbf{q} \in \mathbb{R}^{p}$ represent the loading vectors; the matrix $\mathbf{E} \in \mathbb{R}^{m\times q}$ and vector $\mathbf{f} \in \mathbb{R}^{m}$ represent the residuals, whose norms are minimized. Additionally, the scores vectors are orthogonal and, consequently, $\mathbf{T}^T\mathbf{T}$ is a diagonal matrix. The decompositions (12.14) can be achieved by different procedures, such as the nonlinear iterative partial least squares (NIPALS) algorithm [124] or the statistically inspired modification of PLS (SIMPLS) algorithm [64].

The decompositions (12.14) are performed to minimise the norm of the residual matrices, $\mathbf{E}$ and $\mathbf{f}$, and to maximise the covariance between the latent vectors, columns of $\mathbf{T}$ and $\mathbf{R}$. Consequentially, there is a linear relation between $\mathbf{T}$ and $\mathbf{R}$, expressed as

$$\mathbf{R} = \mathbf{T}\mathbf{D} + \mathbf{H}, \tag{12.15}$$

where $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with the regression weights and $\mathbf{H}$ denotes the matrix of the residuals. Combining (12.15) with the decomposition of $\mathbf{y}$, given by (12.14), leads to

$$\mathbf{y} = \mathbf{T}\mathbf{D}\mathbf{q}^T + \left(\mathbf{H}\mathbf{q}^T + \mathbf{f}\right), \tag{12.16}$$

or simply

$$\mathbf{y} = \mathbf{T}\mathbf{c}^T + \mathbf{f}^*, \tag{12.17}$$

where $\mathbf{c}^T = \mathbf{D}\mathbf{q}^T \in \mathbb{R}^p$ denotes the regression vector and $\mathbf{f}^* = \mathbf{H}\mathbf{q}^T + \mathbf{f}$ is the residual vector, so that $\mathbf{y}$ can be estimated as

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{c}^T. \tag{12.18}$$

The regression model (12.18) makes it possible to estimate $\mathbf{y}$ based on the latent variables $\mathbf{T}$. Still, regressing $\mathbf{y}$ on the original predictor variables $\mathbf{X}$ is useful. To accomplish this, the matrix $\mathbf{W} = \mathbf{X}^T\mathbf{R}$, of the PLS weights, computed such that $\mathbf{EW} = \mathbf{0}$, is post-multiplied by the decomposition of $\mathbf{X}$ in (12.14):

$$\begin{aligned} \mathbf{XW} &= \mathbf{T}\mathbf{P}^T\mathbf{W} + \mathbf{EW} \\ \Leftrightarrow \mathbf{T} &= \mathbf{XW}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}. \end{aligned} \tag{12.19}$$

Replacing (12.19) by (12.18) returns an expression that can be applied to the training data:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{XW}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{c}^T \\ &= \mathbf{X}\mathbf{X}^T\mathbf{R}\left(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R}\right)^{-1}\mathbf{c}^T \\ &= \mathbf{X}\mathbf{d} \end{aligned} \tag{12.20}$$

where

$$\mathbf{d} = \mathbf{X}^T\mathbf{R}\left(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R}\right)^{-1}\mathbf{c}^T \tag{12.21}$$

is the parameter vector of the PLSR regression model. Since the solution of (12.17) by linear least squares, with orthogonal latent predictors $\mathbf{T}$, leads to $\mathbf{c} = \mathbf{y}^T\mathbf{T}$, the parameter vector (12.21) can be written as

$$\mathbf{d} = \mathbf{X}^T\mathbf{R}\left(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R}\right)^{-1}\mathbf{T}^T\mathbf{y}. \tag{12.22}$$

For the reconstruction/prediction period, the PLS regression model will be given by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{d} = \tilde{\mathbf{T}}\mathbf{c}^T = \tilde{\mathbf{T}}\mathbf{T}^T\mathbf{y}, \tag{12.23}$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-m)\times q}$ is the matrix of the predictor variables along the reconstructions/prediction

period and

$$\tilde{\mathbf{T}} = \tilde{\mathbf{X}}\mathbf{X}^T\mathbf{R}\left(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R}\right)^{-1} \in \mathbb{R}^{(n-m)\times p} \tag{12.24}$$

represents the matrix of latent variables in the same period. Other formulations of the PLSR model can be derived based on the properties and identities between the vectors resulting from the algorithm used to obtain the decompositions (12.14) (see, for instance, [133, 103, 83]).

As Eq. (12.24) makes it possible to extend the latent variables along the prediction period ($\tilde{\mathbf{T}} = \left[\tilde{\mathbf{t}}_1\ \tilde{\mathbf{t}}_2 \dots \tilde{\mathbf{t}}_p\right]$), it is possible to use them as predictors in the AnEn-based method. In this work, we also explore combining the AnEn and ClustAnEn methods with the PLS decomposition, using the latent variables as predictors instead of the original variables. The resulting techniques are denoted PLSAnEn and PLSClustAnEn, respectively.

The PLS regression method is also, by itself, an alternative method to PCR and AnEn-based methods for the reconstruction/prediction of the missing records, by estimating them via the regression model (12.23) and, therefore, is also included in the present study.

12.2 Meteorological Datasets

The US National Data Buoy Center (NDBC) [90] is the source for the data used in this study. NDBC manages a network of coastal stations and buoys for data collection. Part of the National Oceanic and Atmospheric Administration (NOAA), the most representative weather centre in America [129], NDBC provides open-source, credible and updated data.

The NDBC database covers regions from almost the entire US coast and some areas across the globe. For this work, we focused on an area with a station density as large as possible. The choice fell on a region centred on the north of the San Francisco Bay (California, USA), for which records are produced by 16 meteorological stations relatively close. Figure 12.2 shows the selected region and its 16 NDBC stations.

Records of various meteorological variables are available for each station, with a sampling period of 6 min. The variables used in this study are atmospheric pressure (PRES, mbar), air temperature (ATMP, °C), wind speed (WSPD, m/s) and peak gust speed (GST, m/s). WSPD and GST can vary significantly over short intervals, and sporadic data gathering incompletely describes their real behaviour. NDBC solved this problem by sampling

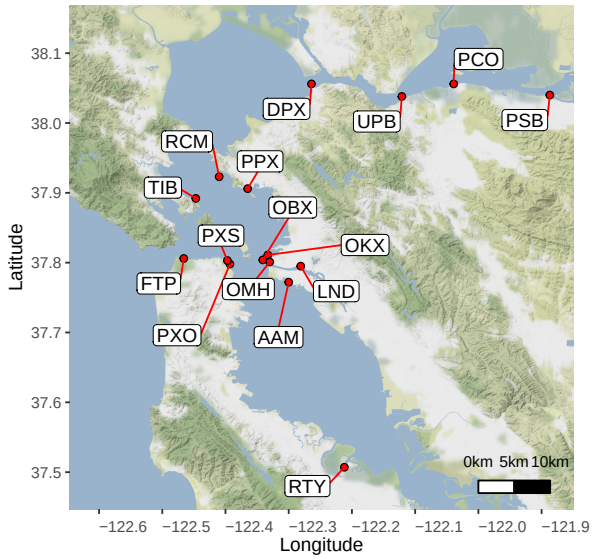


Figure 12.2: Geolocation of the selected NDBC meteorological stations.

the wind speed (WSPD) every 6/8 s and averaging the readings across 6 min; additionally, it considers the maximum wind speed on the same interval as the peak gust speed (GST). In turn, the collection of ATMP and PRES is straightforward: the instantaneous value every 6 min is recorded in the NDBC database.

The records start on 1 January 2016 and end on 31 December 2021. Due to sensor failures and station maintenance operations, variables often present time series with incomplete data. Table 12.1 presents the mean, standard deviation (SD) and availability (in percentage) for the considered variables (WSPD, GST, ATMP and PRES) in the 16 stations.

Only variables with more than 85% of availability (in bold in Table 12.1) were chosen for this study to maximise the amount of training data. The presence of NA in Table 12.1 means that the corresponding variable is not available at the corresponding station. Consequently, the working dataset consists of a total of 41 time series: 10 series with records of the WSPD variable, 10 of the GST variable, 12 of the ATMP variable and 9 of the PRES variable.

Figures 12.3–12.6 show the correlations between the stations with more records for each variable (10 stations for the WSPD, GST and ATPM variables, and 9 stations for the PRES variable). In each figure, a heat map is shown for each station; the polygonal format of this map matches (approximately) a four-sided polygon that includes the 16 stations,

Table 12.1: Characterization of the dataset.

Station	WSPD			GST			ATMP			PRES		
	Mean	SD	Avail.	Mean	SD	Avail.	Mean	SD	Avail.	Mean	SD	Avail.
AAM	2.2	1.6	**93.9**	3.1	2.1	**93.7**	12.7	2.9	**98.8**	1016.9	4.8	**98.9**
DPX	3.6	2.1	**97.6**	4.9	2.7	**97.6**	12.9	3.3	**98.7**	1016.1	5.1	**98.7**
FTP	2.5	1.6	**97.8**	4.2	2.2	**97.8**	14.3	2.5	**98.9**	1016.8	4.8	**99**
LND	2.1	1.4	**92.5**	3	1.9	**92.5**	12.8	2.9	**93.6**	1016.6	4.8	**93.6**
OBX	NA	NA	NA	NA	NA	NA	12.8	2.6	**95.7**	NA	NA	NA
OKX	2.6	1.6	81.7	3.6	2.1	81.7	NA	NA	NA	NA	NA	NA
OMH	3.0	2.0	**85.5**	4.0	2.5	**85.5**	NA	NA	NA	NA	NA	NA
PCO	4.3	2.4	**88.5**	5.6	2.8	**88.5**	12.3	4.0	**93.3**	1016.2	5.1	**93.4**
PPX	3.8	2.2	**95.2**	5.1	2.8	**95.2**	13.1	2.9	**96.6**	1016.8	4.9	**96.6**
PSB	3.9	2.4	**95.1**	5.4	3.2	**95.1**	13.7	4.4	**96.2**	1015.8	5.3	**96.2**
PXO	2.2	1.5	84.7	3.4	2.0	84.7	12.6	2.6	**87.8**	1015.6	4.9	76.1
PXS	NA	NA	NA	NA	NA	NA	13.2	2.5	**96.9**	NA	NA	NA
RCM	2.6	1.6	**93.9**	3.9	2.1	**93.9**	12.8	2.7	**98**	1016.4	4.9	**98**
RTY	1.7	1.5	**96.1**	2.5	1.9	**95.7**	12.5	3.6	**97.6**	1017.0	4.9	**97.7**
TIB	1.7	1.5	3.3	2.7	2.2	3.3	NA	NA	NA	1016.3	3.5	3.3
UPB	4.2	2.3	41.2	5.6	2.9	41.2	NA	NA	NA	NA	NA	NA

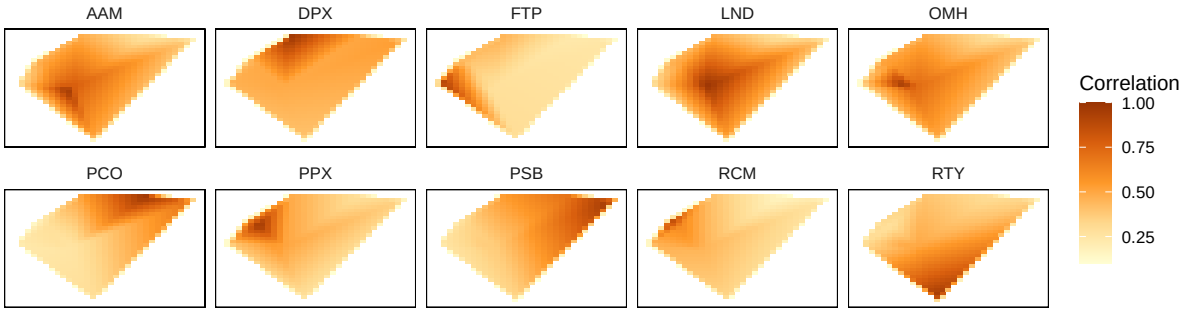


Figure 12.3: Correlation between stations for the WSPD variable. Note: the horizontal and vertical axes refer to longitude and latitude, respectively (labels removed for display purposes).

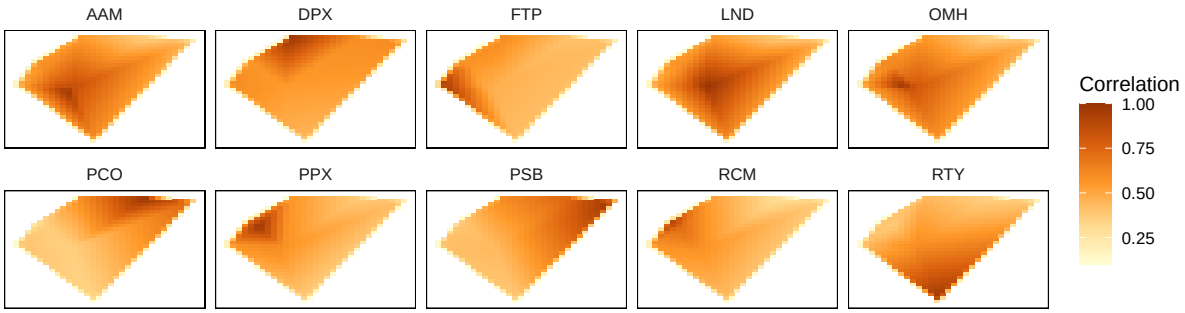


Figure 12.4: Correlation between stations for the GST variable. For further details refer to Figure 12.3.

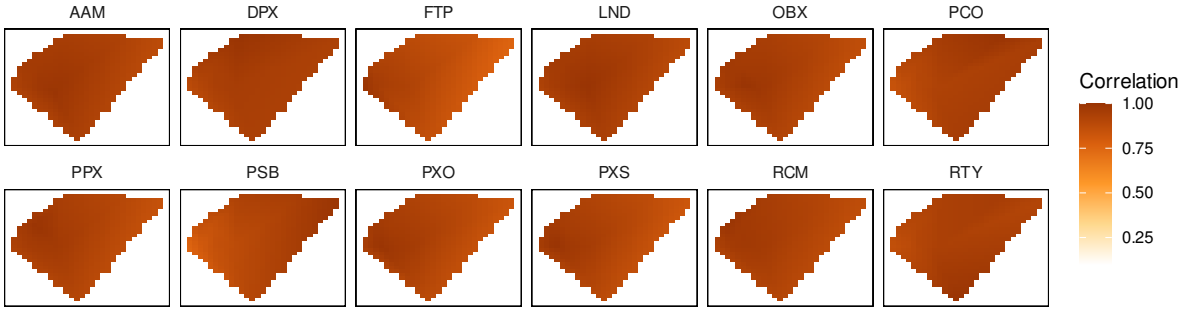


Figure 12.5: Correlation between stations for the ATMP variable. For further details refer to Figure 12.3.

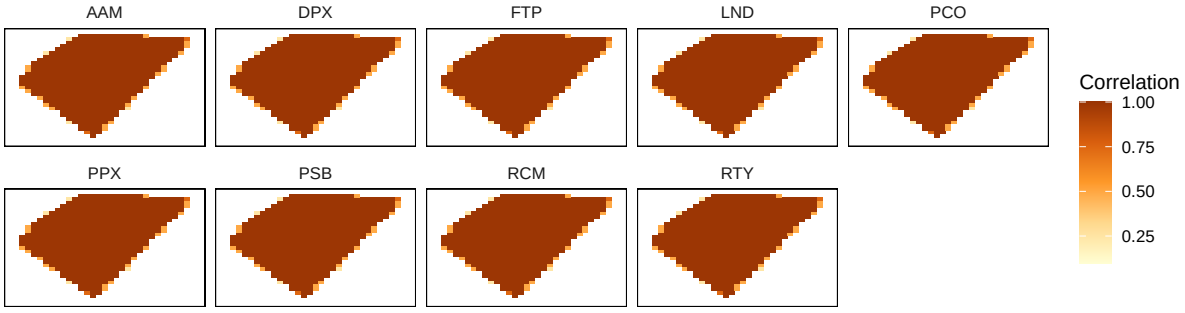


Figure 12.6: Correlation between stations for the PRES variable. For further details refer to Figure 12.3.

preserving their geographical positions and distances, considering the layout of Figure 12.2. In each heat map, only 10 (or 9) points represent the station correlations — the points corresponding to their locations; the correlations for the other points were produced by interpolation.

As may be observed in Figures 12.3 and 12.4, for the WSPD and GST variables, only the closest stations present a strong correlation. Furthermore, the GST variable presents correlations between stations slightly higher than those observed for the WSPD variable. These observations confirm that the wind-dependent variables have a local (and not regional) variation, significantly depending on the morphology of the terrain where the station is implemented.

Figures 12.5 and 12.6 show that the ATMP and PRES variables have a different behaviour from that of the WSPD and GST variables (Figures 12.3 and 12.4), as ATMP and PRES present a very high correlation between stations, even among the most distant ones. This shows that the ATMP and PRES variables have a regional character, varying little or not locally.

12.3 Selecting the Principal Components

In this section, we show how the selection of the principal components (PCs) that are used in the PCAnEn, PCClustAnEn and PCR methods is conducted. The number of PCs to be included in these methods is very important. An insufficient number of PCs translates into the loss of information necessary for data reconstruction, whilst a high number translates into redundant information and increased computational costs.

Identifying the dimensions with most data dispersion makes it possible to identify the principal components $\mathbf{z}_j$, with $j = 1, \ldots, p$, that best distinguish the dataset under study. The dataset corresponding to the multiple predictor time-series, from the various meteorological stations described in Section 12.2, is represented by the data matrix $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_q]$, where each column vector $\mathbf{x}_j$, with $j = 1, \ldots, q$, includes the centered and scaled records of a single variable. The thin singular value decomposition of $\mathbf{X}$ makes it possible to obtain the principal components $\mathbf{z}_j$, $j = 1, \ldots, q$, each one corresponding to a singular value σ_j, $j = 1, \ldots, q$, where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q$ (see Section 12.1.3). The PC vector z_1 has the largest sample variance (σ_1^2/m), z_2 has the second largest variance (σ_2^2/m)

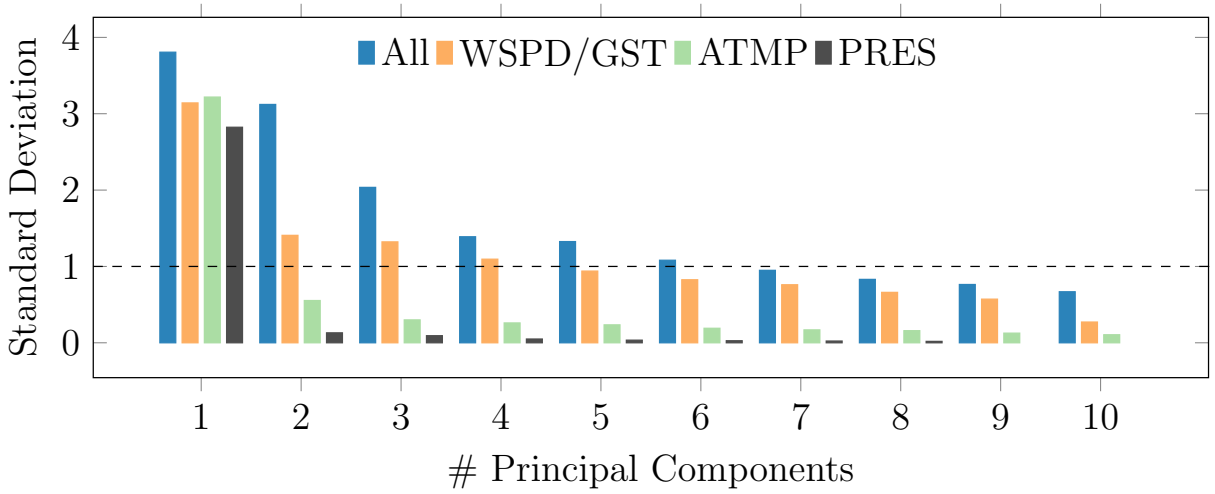


Figure 12.7: Standard deviation of the first PCs for different predictor variables.

and so on. On the other hand, as the original predictor time series are previously scaled by the respective standard deviation, if a PC has a standard deviation greater than 1, this PC defines a dimension with more dispersion, i.e., it contains more information than the original variables. This will be the criteria for selecting the PCs, as employed in [47].

Figure 12.7 shows the standard deviations of the first 10 PCs obtained from 37 predictor variables of the dataset. These predictor variables do not include the time series from the PPX station because these series are used only as predicted/reconstructed variables.

Based on the same figure, the principal component analysis is performed from different predictor matrices $\mathbf{X}$: in the first case (blue bars), the PCs are obtained from all predictor variables ($q = 37$); in the second case (orange bars), the PCs are obtained using the predictor variables WSPD and GST ($q = 18$); in the third case (green bars), only the ATMP time-series were used ($q = 11$); finally, in the fourth case (gray bars), only the PRES series where considered as predictors ($q = 8$). This analysis aims to verify whether there is any advantage in using predictor variables that are different from the predicted ones. As in Chapter 10 and Chapter 11, the variables WSPD and GST were merged as they are highly correlated; hence, it makes no sense to use them separately.

As shown in Figure 12.7, where a dashed horizontal line indicates the PC choice threshold, if all variables are used as predictors, the first six PCs ($p = 6$) must be selected to represent the predictor dataset. For wind-related predictor variables (WSPD and GST), the first four PCs ($p = 4$) must be used. Finally, for a single predictor variable (ATMP or

Table 12.2: Errors of the prediction of PPX station for different principal components.

Predicted	Predictor	# PCs (p)	PCR			PCClustAnEn			PCAnEn		
			BIAS	**RMSE**	**SDE**	**BIAS**	**RMSE**	**SDE**	**BIAS**	**RMSE**	**SDE**
WSPD	All	6	0.47	1.99	1.93	0.55	1.91	1.83	0.55	1.8	1.71
	WSPD/GST	4	**0.39**	**1.85**	**1.81**	**0.38**	**1.79**	**1.75**	**0.42**	**1.75**	**1.7**
ATMP	All	6	**−0.01**	**0.54**	**0.54**	−0.05	1.00	1.00	−0.07	0.8	0.8
	ATMP	1	−0.04	0.64	0.64	**−0.03**	**0.66**	**0.66**	**−0.03**	**0.66**	**0.66**
PRES	All	6	0.04	0.29	0.29	0.08	1.59	1.58	0.05	0.97	0.97
	PRES	1	**0.04**	**0.20**	**0.19**	**0.04**	**0.32**	**0.31**	**0.04**	**0.32**	**0.32**

PRES), only the first PC ($p = 1$) should be chosen.

Table 12.2 shows the errors for the prediction/reconstruction of the variables WSPD, ATMP and PRES, of the PPX station, using the PCR, PCClustAnEn and PCAnEn methods, with the number of PCs previously defined (6, 4 or 1). The reconstruction period was 2021 (the last one of the dataset), and the training period spanned from 2016 to 2020. Due to computing resource constraints, the daily records were reconstructed only from 10 am to 4 pm; moreover, the analogues were searched only in the same period. The reconstruction errors shown are the bias (BIAS), which is an indicator of the systematic error, the standard deviation of the error (SDE), an indicator of the random error and the root mean square error (RMSE), including both the systematic and the random error (for details see Subsection 10.3.1 of Chapter 10). The minor errors are highlighted in bold.

As may be observed in Table 12.2, except for the ATMP variable, there seems to be no advantage in using all available predictor variables, and, in most cases, it is preferable to use the same variable to be predicted/reconstructed as a predictor. This indicates that the proposed methods are not adequate to use all available information without scrutiny. Some signals (time series) are uncorrelated with others; if these are used, they will introduce noise. Hence, the tests with limited variables, based on signals that share the same physical significance, have better results. Moreover, comparing the results obtained by the three methods, we see that they result in very close errors. At most, errors are in the order of tenths of a unit. Notably, the consistent spatial correlation of ATMP and PRES, as seen in Figures 12.5 and 12.6, gives PCRs a slight advantage in their reconstruction, which suggests that a regression model uses this correlation more effectively. In contrast, WSPD has lower spatial correlation and more frequent temporal variations, making PCAnEn the best method for its reconstruction. Finally, PCClusAnEn yields comparable or nearly identical results to PCAnEn for all three variables.

12.4 Selecting the Latent Variables

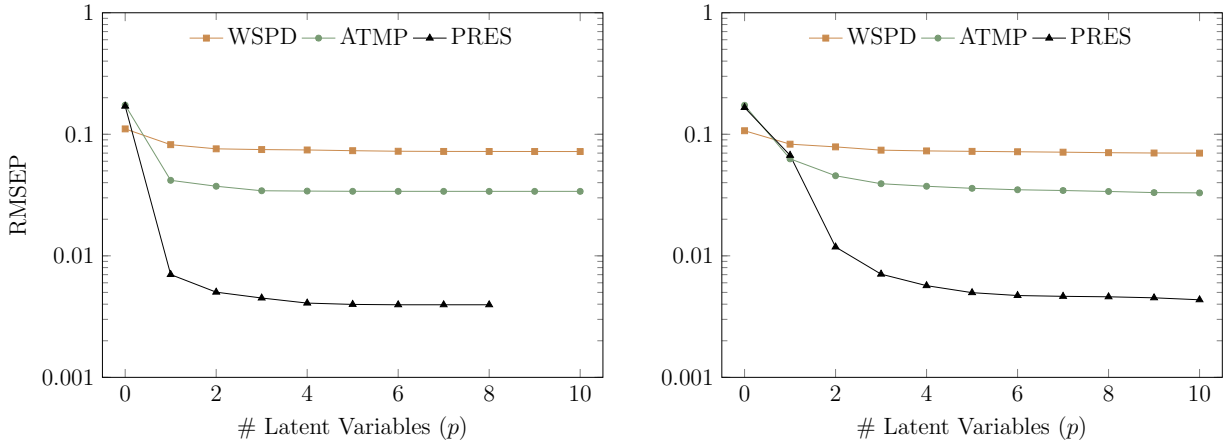
Choosing the number of components (latent variables) is an essential step in applying the PLSR method or variants. As a latent variable is relevant only if it improves the prediction of $\mathbf{y}$, it is first necessary to solve the problem of which and how many latent variables should be kept in the PLSR model to achieve optimal predictions.

In this section, we propose two approaches that can be used to determine the number of latent variables (p). To achieve this, the variables of the PPX station were predicted/reconstructed using the variables of the neighbouring stations. Thus, $q = 37$ original predictor variables can be used to obtain latent variables.

The performance of a PLSR model can be evaluated with computer-based re-sampling techniques, such as cross-validation (c.f. [113, 56]). In this technique, the data of the training period (see Figure 12.1) are split into a *learning* set (used to build a PLSR model) and a *testing* set (used to test the model). In particular, in the Leave-One-Out Cross-Validation (LOOCV) approach, the initial training dataset is partitioned into exactly k subsets (k-fold). Each subset is then used to test the PLSR model built using the data in the $k-1$ learning subsets (for details, refer to [83, 70]). The predicted observations for each testing set are stored in the vector $\hat{\mathbf{y}}^{[p]}$, which is used to determine the overall quality of the PLSR model using p latent variables. The quality of the PLSR model is evaluated by measuring the discrepancy between $\mathbf{y}$ and $\hat{\mathbf{y}}^{[p]}$, using the root mean-squared error predicted (RMSEP):

$$\text{RMSEP} = \frac{\sqrt{m}}{m} \left\| \mathbf{y} - \hat{\mathbf{y}}^{[p]} \right\|. \tag{12.25}$$

Figure 12.8 (a) shows the values of the normalised RMSEP as a function of the number of latent variables (p), originating from the predictions of the three meteorological variables (WSPD, ATMP and PRES), from the PPX stations, using the 10-fold LOOCV technique. Each variable was predicted only by the time-series corresponding to the same meteorological variable ($\mathbf{X}$ and $\mathbf{y}$ contain data from the same meteorological variable), except for WSPD and GST, which are used together to predict WSPD, based on the analysis in Section 12.2. In the case of WSPD, it can be observed that the first four components are responsible for the highest decrease in RMSEP. For several latent variables greater than four, the reduction in RMSEP is not significant. For the ATMP and PRES meteorological variables, the most significant decrease in RMSEP happens for the first three components.



(a) The same variable as predictor

(b) All variables as predictor

Figure 12.8: Normalized RMSEP values for different latent variables on a logarithmic scale. The normalization factor used was the difference between the maximum and minimum values [110].

Figure 12.8 (b) shows the values of the RMSEP generated in the same conditions as in Figure 12.8 (a), with the exception that all meteorological variables in the neighbouring stations were used for each variable. In the case of WSPD, the first two or three latent variables are responsible for the highest decrease in RMSEP. In the case of ATMP and PRES, the most significant reduction in RMSEP occurs in the first three or four components.

An alternative approach to determining the optimal number of latent variables is based on the metric

$$Q_p^2 = 1 - \frac{\text{PRESS}_p}{\text{RESS}_{p-1}}, \tag{12.26}$$

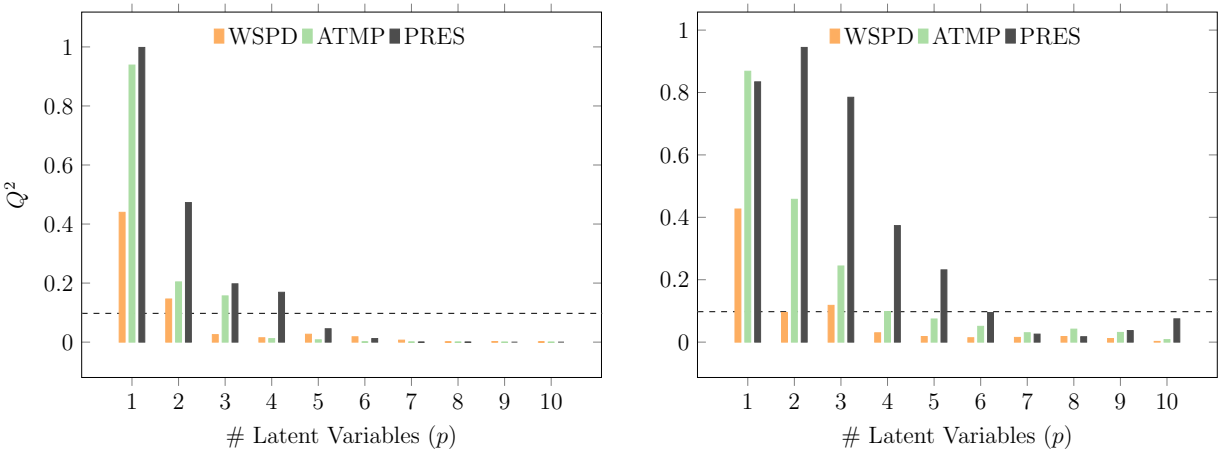
where PRESS_p is the predicted residual sum of squares originated by the LOOCV technique, with the p latent variable, being computed through

$$\text{PRESS}_p = \left\| \mathbf{y} - \tilde{\mathbf{y}}^{[p]} \right\|^2, \tag{12.27}$$

and

$$\text{RESS}_{p-1} = \left\| \mathbf{y} - \hat{\mathbf{y}}^{[p-1]} \right\|^2 \tag{12.28}$$

The residual sum of squares originates from the PLSR model, obtained with the $p-1$ latent variables and built with all data from the training period. The idea of this criterion, proposed in [114], is that a latent variable is kept if the value of the metric (12.26) is larger



(a) The same variable as predictor

(b) All variables as predictor

Figure 12.9: Q^2 metric for a different number of latent variables.

than a certain threshold (ϵ), generally set to $\epsilon = 0.0975$, i.e.,

$$Q_p^2 \geq 0.0975. \tag{12.29}$$

Figure 12.9 (a) shows the values of Q^2 as a function of the number of latent variables, in the case of predictions with the same meteorological variables. It can be observed that the number of latent variables that verify the criterion (12.29) is two for the WSPD meteorological variable, three for the ATMP variable, and four for the PRESS variable.

Similarly, Figure 12.9 (b) represents the values of Q^2 as a function of the number of latent variables, in the case of predictions with all the meteorological variables. It can be observed that the number of latent variables that verify the criterion (12.29) is now three for the variable WSPD, three or four for the ATMP variable, and six for the PRESS variable.

Table 12.3 shows the errors obtained in the prediction of the variables WSPD, ATMP and PRES, of the PPX station, with the PLSAnEn, PLSClustAnEn and PLSR methods, for the cases where a different number of latent variables (LVs) are used as predictors, chosen according to the previously discussed criteria. As in Section 12.3, the prediction/reconstruction period is the year of 2021 (restricted, every day, to the same period—10 a.m. to 4 p.m.) and the remaining years, from 2016 to 2020, make up the training period. For each number (p) of predictor LVs, and each method, the smallest errors are highlighted in bold.

Table 12.3: Errors of the prediction of PPX station for a different number of latent variables (LVs).

Predicted	Predictor	# LVs	PLSR			PLSClustAnEn			PLSAnEN		
			BIAS	**RMSE**	**SDE**	**BIAS**	**RMSE**	**SDE**	**BIAS**	**RMSE**	**SDE**
WSPD	All	2	0.41	1.91	1.86	0.46	1.95	1.89	0.39	1.90	1.85
		3	0.47	1.81	1.75	0.50	1.84	1.77	0.50	1.80	1.73
	WSPD/GST	2	0.44	1.86	1.81	0.43	1.85	1.80	0.35	1.84	1.81
		3	**0.41**	**1.80**	**1.75**	**0.41**	**1.81**	**1.76**	**0.37**	**1.76**	**1.72**
ATMP	All	3	−0.02	0.60	0.60	−0.02	0.82	0.82	−0.03	0.71	0.7
		4	0.02	0.58	0.58	0	0.85	0.85	−0.01	0.7	0.7
	ATMP	3	**−0.01**	**0.53**	**0.53**	**−0.04**	**0.57**	**0.57**	**−0.04**	**0.56**	**0.56**
		4	−0.01	0.53	0.53	−0.03	0.59	0.59	−0.03	0.57	0.57
PRES	All	3	0.03	0.20	0.20	−0.03	1.04	1.04	−0.01	0.75	0.75
		6	0.01	0.13	0.13	−0.01	1.27	1.27	−0.02	0.85	0.85
	PRES	3	0.04	0.12	0.12	0.03	0.33	0.33	0.03	0.31	0.31
		4	**0.01**	**0.12**	**0.12**	**0.03**	**0.31**	**0.31**	**0.02**	**0.31**	**0.31**

In agreement with what was also observed in Section 12.3, using all variables as predictors did not bring any advantages, and the errors obtained were smaller for predictor variables corresponding to the same meteorological variables being predicted. It may also be observed that the increase in the number of LVs does not always translate into a smaller error in the reconstructed/predicted values; for instance, the errors obtained for ATMP were higher with four LVs, rather than three. The PLSAnEn method exhibited the best results in the reconstruction/prediction of WSPD. The PLSR method showed the fewest errors in the reconstruction/prediction of PRES and ATMP, with results not far from those produced by the PLSAnEn method (for WSPD and ATMP).

12.5 Comparison of Methods

In this section, we examine the performance of the different reconstruction methods using data from neighbouring stations. The results are divided into two subsections: Section 12.5.1, featuring the prediction of the station PPX, and Section 12.5.2, which focuses on the reconstruction of each station contained in the dataset. As in previous sections, the reconstruction period is 2021 (only the daily period from 10 a.m. to 4 p.m.), and the remaining years, from the beginning of 2016 to the end of 2020, constitute the training period. All reconstruction/prediction methods were applied for each meteorological variable,

Table 12.4: Number of PCs or LVs used by each method.

Predicted	**Predictor**	**# Orig. Pred.** q	**PCR** p	**PCClustAnEn** p	**PCAnEn** p	**PLSR** p	**PCClustAnEn** p	**PCAnEn** p
WSPD	WSPD/GST	18	4	4	4	3	3	3
ATMP	All/ATMP	37/11	6	1	1	3	3	3
PRES	PRES	8	1	1	1	4	4	4

and the optimal number of PCs or LVs was determined in the two previous sections.

Table 12.4 contains the number of PCs or LVs used by each method. As shown in Table 12.2, ATMP reconstruction using the PCR method obtains better results when all meteorological variables are used as predictors. For this reason, this is the only case where the PCs ($p = 6$) are obtained from all $q = 37$ original predictor variables. In the remaining cases, both PCs and LVs are obtained from predictor variables corresponding to the same meteorological variable that is predicted, i.e., PRES is predicted from PRES and WSPD is predicted from WSPD and GST, as shown in Table 12.4.

12.5.1 Reconstruction of Meteorological Variable in One Station

Figure 12.10 makes it possible to compare the real/observed values with the reconstructed (or predicted) ones, for the three meteorological variables, at the PPX station, on 9 January 2021, from 10 a.m. to 4 p.m. Only the values obtained by the methods with the fewest errors in Table 12.2 and Table 12.3 are presented. It can be seen that the WSPD variable varies much more in time than ATMP, and especially in comparison to the PRES variable. Visually, WSPD exhibits more variance at higher frequencies, i.e., more fluctuations between consecutive records.

For the WSPD variable, the reconstructed values cannot reproduce specific variations. However, they reproduce the general trend of variation of the variable. It should be stressed that this meteorological variable undergoes permanent changes over time. These changes are greatly influenced by factors inherent to the location, such as the orientation and topography of the site. It can also be seen that there are no significant differences between the values reconstructed by the PCAnEn and PLSAnEn methods.

For the ATMP variable, the values obtained by the PLSR method are closer to the observed values than those produced by the PCR method. Both sets of values exhibit some distance from the real values. In this geographic region (San Francisco Bay Area), there are

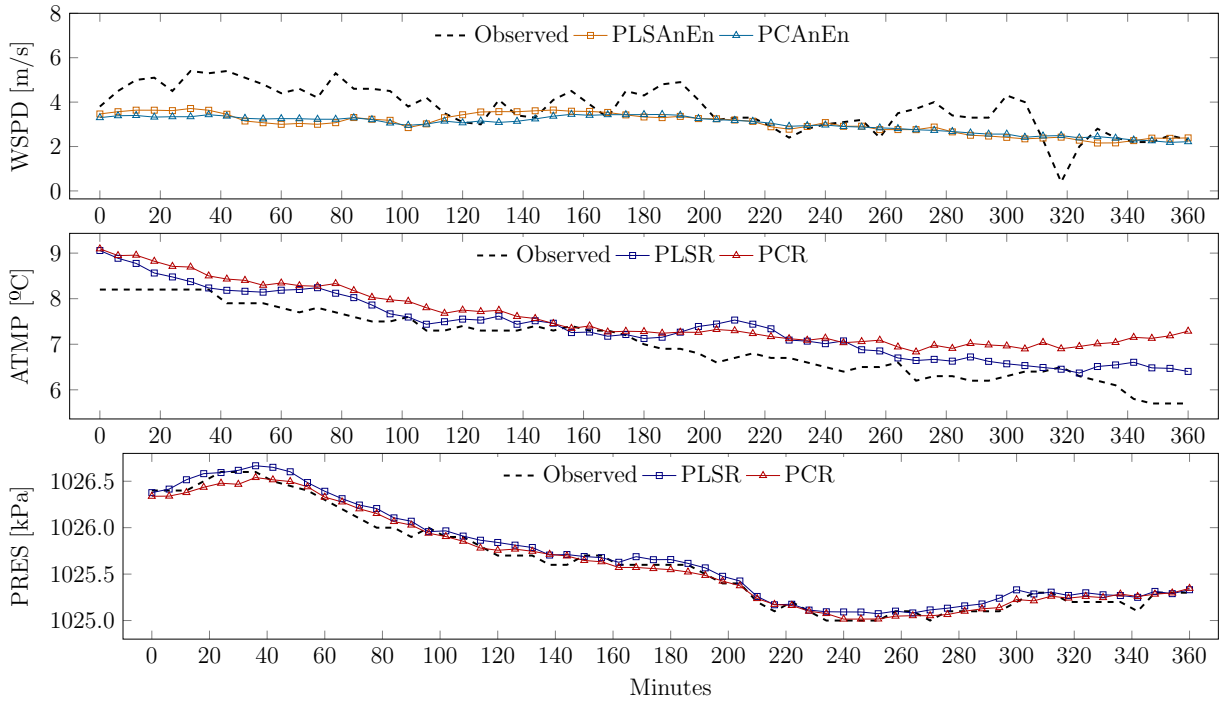
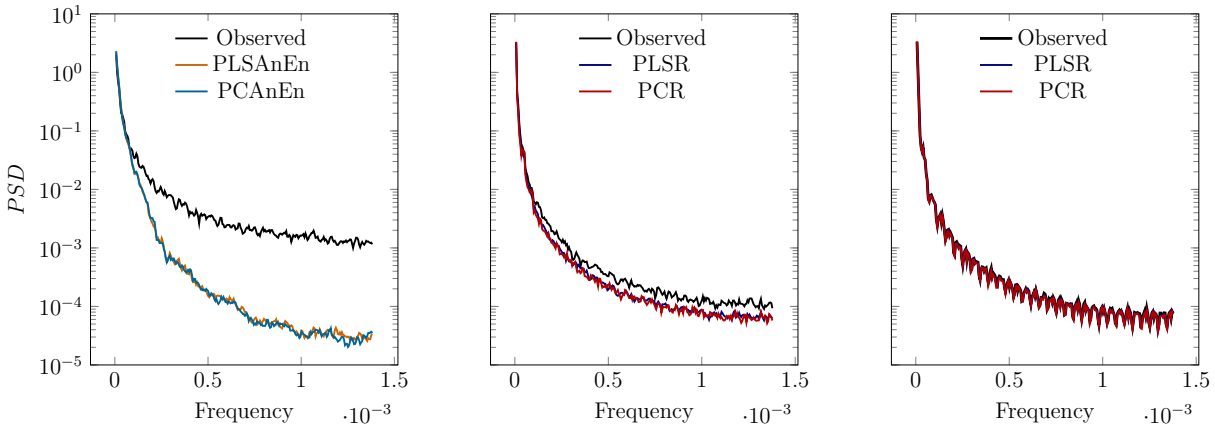


Figure 12.10: Comparison between reconstructed and observed values of the meteorological variables WSPD, ATMP and PRES, from the PPX station, at 9 January 2021, from 10 a.m. to 4 p.m



(a) WSPD (b) ATMP (c) PRES

Figure 12.11: Power spectral densities of the reconstructed/observed time series from the PPX station.

persistent meteorological effects, namely the northern Oregon–California coastal jet [36], coupled with the summer sea breeze [54]. The temperature's spatial distribution depends on how inland the meteorological stations are located. Despite the spatial correlation shown in Figure 12.5, the PCR method is worse than PLSR in reproducing the tendencies in temperature evolution.

The values reconstructed by the PLSR and PCR methods for PRES are very close to the observed values. Firstly, this meteorological variable is highly correlated over the spatial region of interest (c.f. Figure 12.6); hence, the tendencies of the signals are similar. This is because the stations are located at similar altitudes (near sea-level), and as pressure spatial variations are smooth. Secondly, its signal exhibits lower high-frequency fluctuations when compared to the wind speed signals. Both of these characteristics contribute to the performance of the reconstruction methods.

Although Figure 12.10 presents a one-day comparison of predicted and observed values, it is insufficient to conclude that these patterns remain constant throughout the entire reconstructed period (1 year). Thus, an analysis is necessary to verify the consistency of these patterns in the complete dataset. To evaluate the prediction methods' ability to capture high-frequency patterns, the normalised power spectrum densities (PSDs) [27] of both reconstructed and observed series are compared in Figure 12.11. Upon initial observation, the WSPD data display a higher frequency density than ATMP and PRES. Generally, at low frequencies, all the methods align with the observed data across all

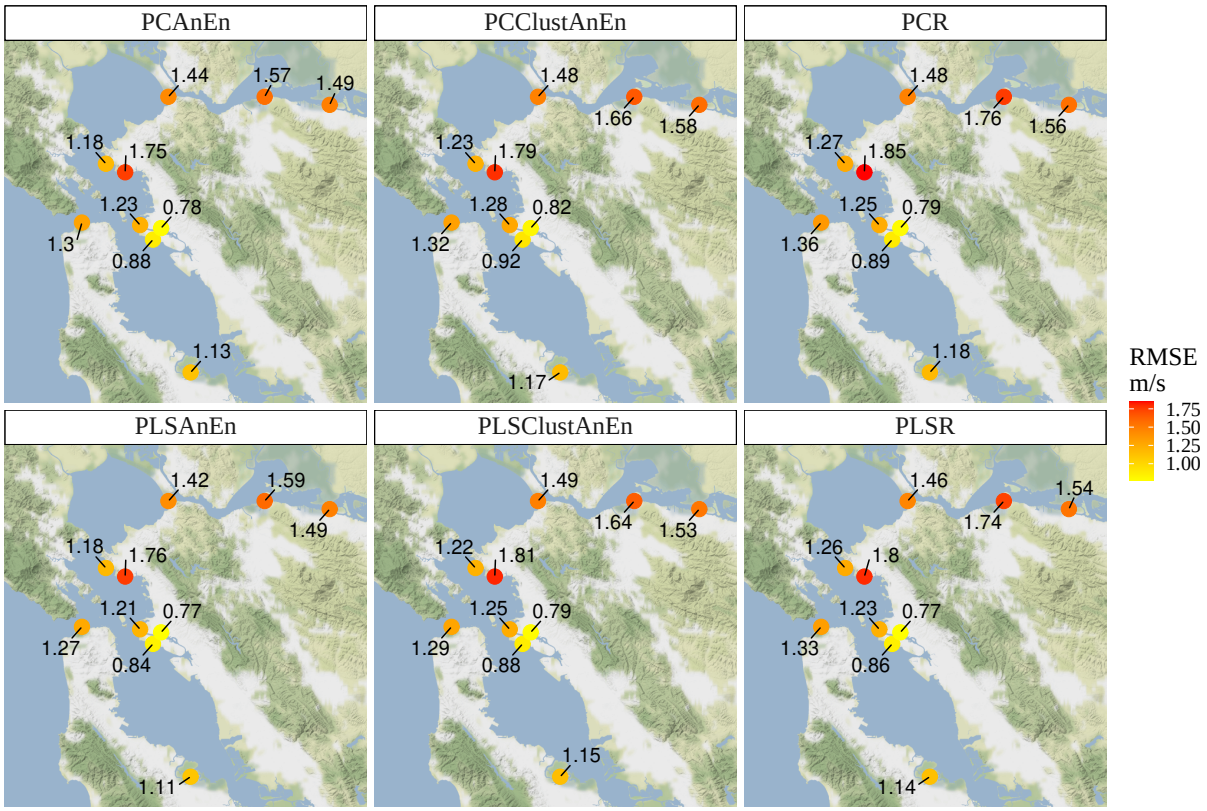


Figure 12.12: RMSE for the reconstruction of WSPD variable across all available stations.

variables. However, for WSPD, the predictions fail to accurately capture the observed high-frequency variance. In contrast, for ATMP and PRES, the reconstructions exhibit similar high-frequency variance to the observed data. These results corroborate the previous one-day analysis.

12.5.2 Reconstruction of Meteorological Variable in All Stations

Figure 12.12 shows the values of the RMSE errors per station, resulting from the reconstruction of the WSPD meteorological variable by the different methods. For all methods, it is observed, in general, that the lowest errors are obtained for the stations that occupy central positions in relation to the others. In general, it is also observed that PLS-based methods obtain lower errors than PC-based methods, although the differences are minor (on the order of a tenth or a hundredth of a unit). The PLSAnEn method presents the best results in reconstructing the WSPD in all meteorological stations.

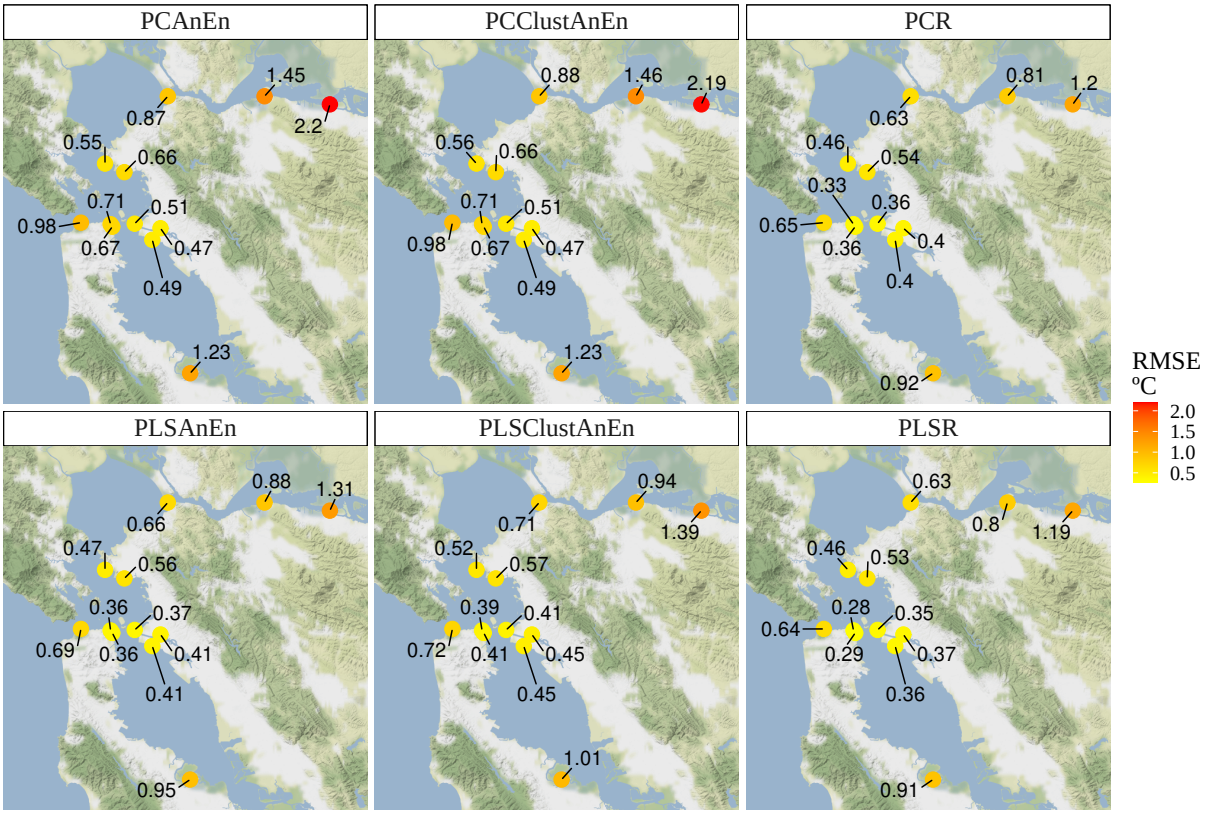


Figure 12.13: RMSE for the reconstruction of ATMP variable across all available stations.

In Figure 12.13, it is possible to observe the values of the same error for the reconstruction of the ATMP variable. In general, the errors obtained are minorer than in the case of WSPD. Here, the highest errors are obtained in the most peripheral stations, which correlate less with the remaining stations. For this meteorological variable, the supremacy of methods based on the PLS method is also verified. The best results at all stations are obtained by the PLSR method, followed closely by the results obtained by the PLSAnEn method.

Finally, Figure 12.14 also shows the RMSE errors during the reconstruction of the meteorological variable PRES. The low error values show that PRES is the easiest variable to reconstruct. In this case, PLS-based methods do not always obtain the best score, which the PLSR method achieves. Following PLSR, PCR obtains the best results, especially in the most central stations. The PLSAnEn and PLSClustAnEn methods get the same results at all stations.

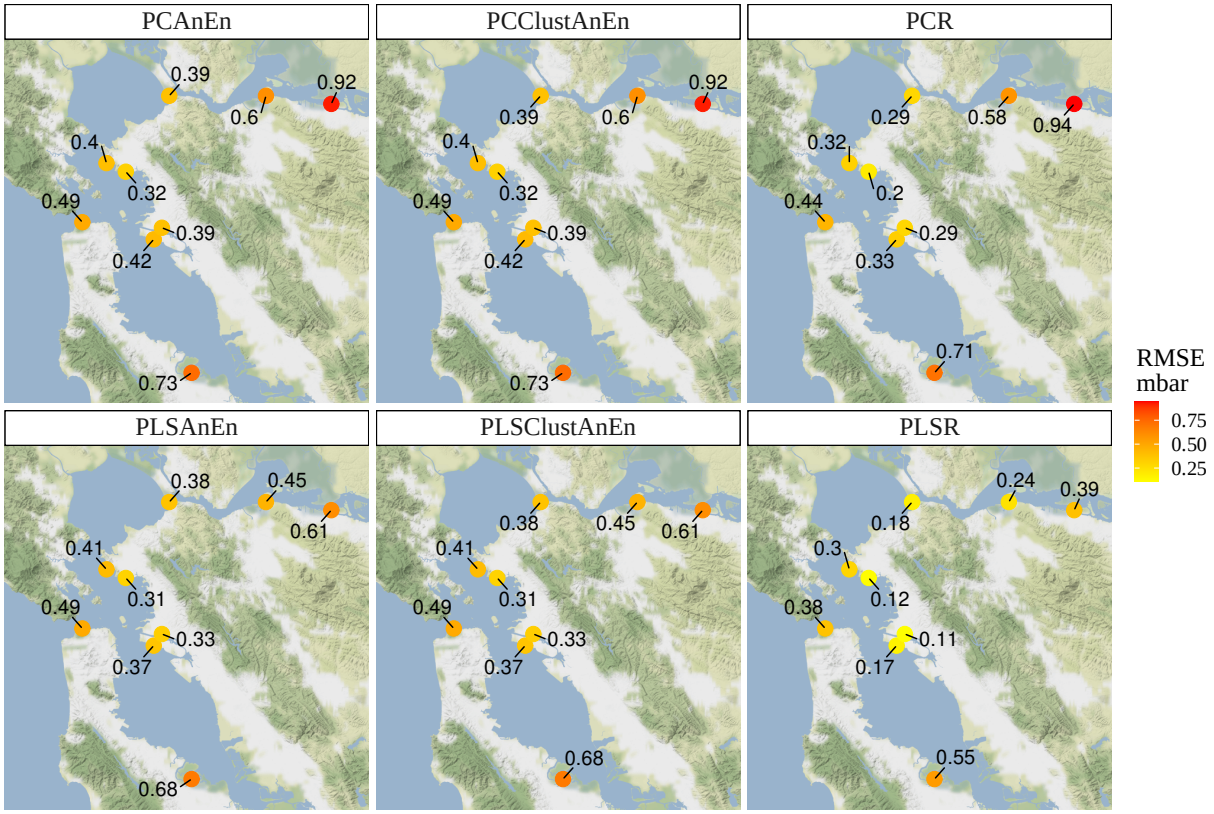


Figure 12.14: RMSE for the reconstruction of PRES variable across all available stations.

12.6 Computational Performance

This section analyses the computational performance of the reconstruction methods considered in this study. The computational system used for the evaluation was a virtual machine hosted in a KVM-based virtualisation cluster, with the following characteristics: 16 cores of an Intel Xeon W-2195 CPU, 64 GB of RAM, SSD-based local storage, Ubuntu 20.04 operating system, R version 4.2.2 (all tests were implemented in the R language [98]).

Table 12.5 presents the mean execution times, in seconds, needed by each method in the reconstruction of the meteorological variables for all the different stations analysed in the previous section (each execution time presented in the table corresponds to the average of the times of all stations). These execution times concern the execution in a parallel regime (by instructing the R platform to exploit, whenever possible, all the available CPU cores).

Table 12.5: Mean execution times across methods, variables, and steps in seconds.

Method	**WSPD**			**ATMP**			**PRES**		
	Loading	**PCA/PLS**	**Prediction**	**Loading**	**PCA/PLS**	**Prediction**	**Loading**	**PCA/PLS**	**Prediction**
PCAnEn	37.2	0.5	341.4	24.1	0.3	72.1	24.6	0.3	72.1
PLSAnEn	37.2	0.7	289.9	24.1	0.4	287.1	24.6	0.3	340.3
PCClustAnEn	37.2	0.5	13.2	24.1	0.3	4.2	24.6	0.3	5.5
PLSClustAnEn	37.2	0.7	11.8	24.1	0.4	6.0	24.6	0.3	7.1
PCR	37.2	3.7	0	61.2	6.8	0	24.6	2.0	0
PLSR	**37.2**	**3.4**	**0**	**24.1**	**2.1**	**0**	**24.6**	**1.7**	**0**

The execution times are divided into three consecutive stages: loading, decomposition (PCA or PLS), and prediction/reconstruction. The loading stage corresponds to reading the data set from CSV files and interpolating missing values (when they are not more than four consecutive values). In the case of the reconstruction of WSPD, the loading step involves the reading of both the WSPD and GST dataset files, which takes ≈37.2 s; even longer, the loading of the ATMP dataset, when using the PCR method, takes ≈61.2 s because this method uses all 37 predictor variables (recall Table 12.2).

The PCA or PLS decomposition step corresponds to calculating PCs or LVs. This step is conducted by internal R functions that are already highly optimised from a computational point of view. As such, the execution time of this step is very fast (however, this step can also include the choice of the number of components through the LOOCV method, which may have some computational costs).

The final step consists of the reconstruction/prediction of the missing values. In the

PCR and PLSR methods, this stage is executed quickly through linear regression with previously determined PCs or LVs. However, for PCAnEn and PLSAnEn, this stage is very demanding because, for each value to be reconstructed/predicted, the training period is swept in search of analogues. In turn, for the PCClustAnEn and PLSClustAnEn methods, this step is not as demanding, once all possible analogues are previously clustered, and the sweeps are reduced to a single operation in which the predictor value is compared with the cluster centroid. Overall, the slowest methods are thus the PCAnEn method (WSPD) or the PLSAnEn method (ATMP and PRES), and the fastest (for all variables) is the PLSR method.

The execution times presented in Table 12.5 were obtained using all CPU cores available. The impact of using various CPU cores may be apprehended by inspecting Figure 12.15. This figure represents the execution time for the reconstruction of WSPD in station PPX as a function of the number of CPU cores, without including the loading time (more IO-sensitive). PLSR and PCR cross-validation were parallelised in these experiments to accelerate the 10-fold LOOCV process. It can be seen that the PCR and PLSR methods perform similarly for any number of CPUs. In turn, the PLSClustAnEn and PCClustAnEn methods benefit from increasing the number of CPU-cores, especially up to 6/8; however, as these methods depend heavily on the clustering phase of possible analogues, which is not always performed in the same number of iterations, their performance does not continually improve with the increased calculation capacity.

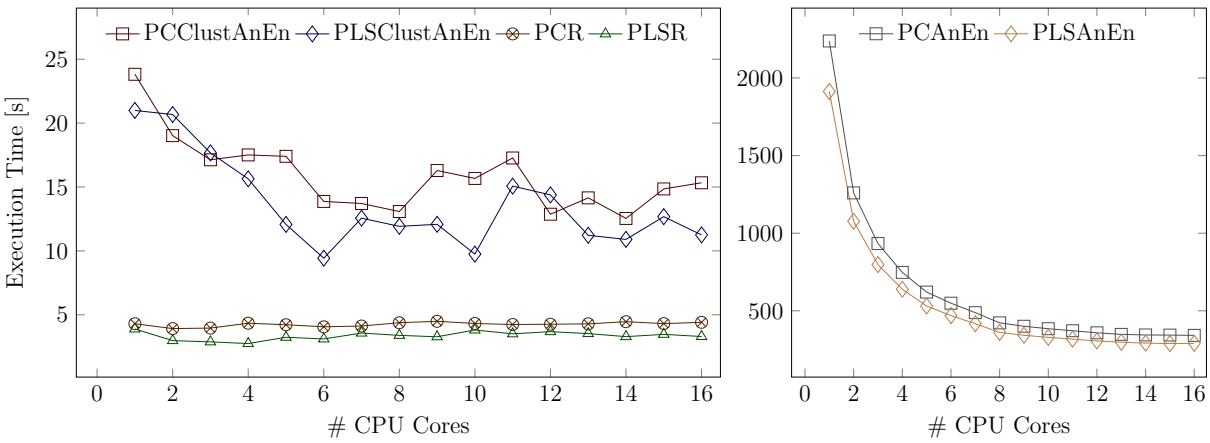


Figure 12.15: CPU time for the reconstruction of WSPD in station PPX in function of the number of cores (excluding loading time).

Regarding the PCAnEn and PLSAnEn methods (right of Figure 12.15), it may be

observed that they are more sensitive to the increase in the CPU-cores employed, with their computational efficiency improving when using up to about ≈10 cores. It turns out that these methods are highly parallelisable: many searches for analogues may be carried out simultaneously once they are inherently independent from each other. However, despite the performance gains brought by the parallel execution, the PCAnEn and PLSAnEn methods are still considerably slower than the others.

12.7 Concluding Remarks

This study presents methods that allow for solving hindcasting and forecasting problems with a high number of predictors. Solving these problems with a large number of predictors using the classical analog ensemble methodology, though feasible, is very inefficient, due to the magnitude of the computational load involved.

The methods presented here combine the robustness of the AnEn method (with or without clustering) and the PCA and PLS techniques for dimension reduction of the predictor dataset. Using these techniques, the predictor variables are reduced to a small number of new variables that mostly retain (PCA) and may even enhance (PLS) the meteorological information used by the AnEn method to reconstruct or forecast the records sought.

The results produced by the PLS-based techniques were slightly more accurate than those obtained with the PCA-based ones, especially in reconstructing or forecasting meteorological variables with significant oscillation, such as wind speed (WSPD). This happens because PLS builds the latent variables so that they simultaneously explain the variation of the predictor and the predicted variable. In contrast, the main components, obtained by PCA, only explain the variation of the predictor variables.

Combining the AnEn method with PLS results in a hybrid method, PLSAnEn, which makes it possible to accurately reconstruct or predict the wind speed. It is therefore a highly suitable forecasting method for meteorological time series with potential applications in wind resource assessment and wind energy. At the same time, PLSAnEn is very demanding from a computational point of view, which benefits from a parallel implementation.

PLSClustAnEn, which combines the AnEn methods with the prior clustering of analogues, could be an alternative to the PLSAnEn method, as it is much more computatio-

nally efficient. However, it is a method that depends on many parameters that have to be chosen appropriately to improve the accuracy of the results.

Simultaneously with the AnEnbased methods, regression methods were also tested on the new variables determined by PCA or PLS. The resulting methods, PCR and PLSR, are speedy and allow accurate reconstructions. In particular, the PLSR method is the most suitable for reconstructing or forecasting highly correlated predictor variables.

Chapter 13

Optimal Latent Variables Number for PLSR

The results from this chapter have already been published as:
Balsa, C., Dupuis, H., Breve, MM., Guivarch, R., Rufino, J. (2024). *Optimal Latent Variables Number for the Reconstruction of Time Series with PLSR.* In: Garcia, M.V., Gordón-Gallegos, C., Salazar-Ramírez, A., Nuñez, C. (eds) Proceedings of the International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2023). CSEI 2023. Lecture Notes in Networks and Systems, vol. 775. Springer, Cham. ISBN 978-3-031-69227-7. ISSN 2367-3370. https://doi.org/10.1007/978-3-031-69228-4_13.

As we have seen in the previous chapter, the Partial Least Squares Regression (PLSR) is a method that combines the partial least squares decomposition (PLS) with multilinear regression [125]. It is particularly used when the number of predictors is greater than the number of predicted variables, which is the case in reconstructing meteorological time series.

PLSR is upgrading linear multi-regression methods by maximising the covariance between the latent (new predictor variables) and predicted variable space. It reduces the dimension of the predictor dataset without losing valuable information. Consequently, it reduces the computational time involved, a significant issue in reconstruction problems

with large amounts of data. Therefore, PLSR is an appropriate method for hindcasting problems where gaps of meteorological time series are filled based on the values of other time series coming from neighbouring stations (see Chapter 12 and [19]).

Nevertheless, the main difficulty when using PLSR is finding how many latent variables to choose. Indeed, one wants to include all the valuable information in the prediction, without adding unnecessary information. For instance, it is conjectured that the lower-order latent variables are associated with process noise, so they should not be added to the model. It is thus necessary to use some form of criterion to select which latent variables should be chosen or not.

This chapter is a complementary study to the one presented in Section 12.4 of the previous chapter and is based on the publication [15]. It is focused on estimating the optimal number of latent variables based on different criteria. Several criteria are applied, such as cross-validation or information criteria, and the efficiency of each one is discussed based on the results obtained with a meteorological dataset.

The remaining of the paper is organised as follows: Section 13.1 revisits the fundamentals of the PLSR method; Section 13.2 introduces the criteria to estimate the optimal number of latent variables; Section 13.3 is devoted to the results obtained with the application and validation of the different tested criteria; Section 13.4 sums up the main concluding remarks of the study presented in this chapter.

13.1 Partial Least Squares Regression

The problem that frames this research is the prediction or reconstruction of the values of a defined meteorological variable ($\mathbf{y}$ vector) at one particular station, based on meteorological variables from the nearby stations, called predictors ($\mathbf{X}$ matrix). The Partial Least Squares Regression (PLSR) method can tackle this problem.

In this context, PLSR is used for single- and multivariate regression that uses the components (called latent variables) of $\mathbf{X}$ that best predict $\mathbf{y}$. A PLS model will try to find the multidimensional direction in the $\mathbf{X}$ space that explains the maximum multidimensional variance direction in the $\mathbf{y}$ space [1, 83]. PLSR describes the changes in both the $\mathbf{X} \in \mathbb{R}^{n \times q}$ space and the $\mathbf{y} \in \mathbb{R}^n$ space by determining the latent variables, resulting in the variable

decomposition

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{13.1}$$

and

$$\mathbf{y} = \mathbf{R}\mathbf{q}^T + \mathbf{f} \tag{13.2}$$

where $\mathbf{T} \in \mathbb{R}^{m\times p}$ and $\mathbf{R} \in \mathbb{R}^{m\times p}$ are the matrix with p latent vectors (also known as scores), extracted from $\mathbf{X}$ and $\mathbf{y}$, respectively; $\mathbf{P} \in \mathbb{R}^{q\times p}$ and $\mathbf{u} \in \mathbb{R}^p$ are the loading matrices; the matrix $\mathbf{E} \in \mathbb{R}^{m\times q}$ and vector $\mathbf{f} \in \mathbb{R}^m$ are the residuals terms. The decompositions are performed to minimise the norm of the residual matrices, $\mathbf{E}$ and $\mathbf{f}$, and to maximise the covariance between the latent vectors, columns of $\mathbf{T}$ and $\mathbf{R}$, through some algorithms such as the Nonlinear Iterative Partial Least Squares (NIPALS) [125] or the Statistically Inspired Modification of PLS (SIMPLS) [64].

As we have seen in Section 12.1.5 of Chapter 12, the latent vectors (variables) can be used to build the Partial Least Square Regression regression model that enables to estimate the values of the meteorological variable, $\mathbf{y}$, along the reconstruction period, denoted by $\tilde{\mathbf{y}}$. The expression of the PLSR model is

$$\tilde{\mathbf{y}} = \tilde{\mathbf{T}}\mathbf{T}^T\mathbf{y}, \tag{13.3}$$

where

$$\tilde{\mathbf{T}} = \tilde{\mathbf{X}}\mathbf{X}^T\mathbf{R}\left(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R}\right)^{-1} \in \mathbb{R}^{(n-m)\times p}$$

represents the matrix of latent variables in the reconstruction period, and $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-m)\times q}$ is the matrix of the predictor variables along the same period.

13.2 Criteria for Optimal Number of Latent Variables

One of the main issues concerning the reconstruction of time series with the PLSR method is determining the optimal number, p, of latent variables (LVs). Choosing a number too low may result in losing information about the data to predict, while selecting a number too high results in over-fitting the model.

This chapter compares different criteria to determine the optimal number of LVs, namely, RMSECV, Repeated RMSECV, Q^2, Wold's R, Osten's F, VIP, AIC and BIC.

13.2.1 Cross-Validation Methods

One of the most popular methods used to determine the efficiency of a regression model is the cross-validation (CV) [113, 56]. This method is used when the dataset is large. The dataset is divided into k subsets of equal size (k-Fold Cross-Validation). The $k-1$ subsets form a training set, and the kth subset is the testing set. When there are a lot of observations in the data set, it is helpful to compare the replicate error variation to the fit variation.

The criteria for estimating the optimal number of latent variables based on CV techniques, and presented in these Subsection, are based on the seminal work of Svante Wold [126].

One of the ways to compare the quality of the prediction consists of the Residual Estimated Sum of Squares (RESS), given by

$$\text{RESS} = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 \tag{13.4}$$

where y is the vector that contains the observations and $\tilde{y}$ is the vector of the estimates values.

The Predicted Residual Estimated Sum of Squares (PRESS) measures the quality of the prediction by predicting the model with the PLS regression, and it is given by

$$\text{PRESS} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \tag{13.5}$$

where $\hat{y}$ contains the predictions of the PLS regression model.

As we have seen in Section 12.4, a first approach to determine the optimal number of latent variables is based on the metric

$$Q_p^2 = 1 - \frac{\text{PRESS}_p}{\text{RESS}_{p-1}}, \tag{13.6}$$

where PRESS_p is the predicted residual sum of squares computed with p latent variable.

Another criterion is the RMSECV (Root Mean Square Estimation of Cross-Validation), given by

$$\text{RMSECV} = \sqrt{\frac{\text{PRESS}}{N}} \tag{13.7}$$

where PRESS is the Predicted Residual Estimated Sum of Squares given by Eq. (13.5), and N is the number of observations in the set. The optimal number of LVs is then given by the minimisation of Eq. (13.7).

Nevertheless, when applying some permutations on the rows of the initial dataset, the scaling and the centring of the values differ. Consequently, other values for the residuals are observed. A classic solution is to perform repeated cross-validation to indicate the variability better. Here, the algorithm was repeated 10 times to a set divided into two subsets, randomly selecting each iteration's rows in the training subset.

13.2.2 Wold's R

Another of the criteria tested in this work, based on CV, to determine the optimal number of latent variables is Wold's R criterion [126]. It consists of comparing the rate of two consecutive values of PRESS:

$$\mathrm{R} = \frac{\mathrm{PRESS}_{p+1}}{\mathrm{PRESS}_{p}} \tag{13.8}$$

A threshold is set to choose the number of latent variables (typically, 0.90 or 0.95). If the R value exceeds the threshold, no additional latent variables will be used in the model.

13.2.3 Osten's F

Donald W. Osten [92] suggested another criterion, based on CV, called the F-test-based criterion, given by

$$\mathrm{F} = \frac{\mathrm{PRESS}_{p} - \mathrm{PRESS}_{p+1}}{\mathrm{PRESS}_{p+1}} \times (N - p - 1) \tag{13.9}$$

where PRESS_p is the PRESS value for the model with p latent variables and N is the sample size.

13.2.4 Variable Importance in Projection (VIP)

Variable Importance in Projection (VIP) value explains the dependent variable's changes, reflecting the relative importance of each latent variable in the model. It is given by [107]

$$\text{VIP} = \sqrt{\frac{q \sum_{i=1}^{p} SS_i \left(\frac{w}{\|w\|^2} \right)}{\sum_{i=1}^{p} SS_i}} \tag{13.10}$$

where q is the number of predictors, p is the number of LVs, SS_i is the sum of squares for the model with i latent variables (given by $\mathbf{T}^T\mathbf{T} \times \mathbf{q}\mathbf{q}^T$, following the notations from Eq. (13.1) and Eq. (13.2)), and w is a loading weights vector. A variable is often considered appropriate if the VIP value exceeds 1.

13.2.5 Information Criteria (AIC and BIC)

Two criteria created by Hirotugu Akaike [2] and Gideon E. Schwartz [109] are, respectively, the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion). When a latent variable is added to the model, the likelihood increases. These criteria penalise the models in terms of the number of variables that satisfy the parsimony criterion. The following formulas compute the AIC and BIC values [31]:

$$\text{AIC} = \frac{\text{RSS}}{N} + \frac{2\text{DoF}\hat{\sigma}^2}{N} \tag{13.11}$$

$$\text{BIC} = \frac{\text{RSS}}{N} + \frac{\log(n)\text{DoF}\hat{\sigma}^2}{N} \tag{13.12}$$

where N is the number of observations, RSS is the Residual Sum of Squares:

$$\text{RSS} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2, \tag{13.13}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the vector of the observed and predicted values, respectively, $\hat{\sigma}$ is the estimate of the error variance, given by (c.f. [122])

$$\hat{\sigma} = \text{RSS}/(\text{DoF} - 1),$$

and DoF is the number of degrees of freedom of the observation vector. For the PLS regression (c.f. [119])

$$\text{DoF} = n(1 - \text{RMSE}/\text{RMSECV}).$$

13.2.6 Other Criteria

This study focuses on the most common criteria currently used in the targeted research domain. Still, other studies exploited other criteria from specific domains (see, for instance, [106, 81]).

It was also pondered another type of criterion, that uses the residuals to evaluate the quality of a model (the difference between the observations and the predictions), such as the Durbin-Watson (DW) criterion or the Van der Voet t-test [118]. They measure the correlation between consecutive values of the residuals vector and conclude, by a Student test, whether one should add the latent variable to the model. Unfortunately, these criteria can not be used in our study because our observations are relatively constant over time, while the DW test or the t-test help detect huge variations and huge decorrelations in a short period. That is why they are primarily used in chemometrics or spectroscopy, where the values can change quickly.

13.3 Application and Validation of the Criteria

The results obtained with the criteria introduced in the previous section are now presented. An estimation of the number of latent variables is done using each criterion. The section begins by explaining the data used to implement the criteria.

13.3.1 Dataset Description

The dataset is from the US National Data Buoy Center (NDBC) [90]. It contains information about 16 stations relatively close to each other, in the north of the San Francisco Bay, CA, USA (see Section 12.2 for details). The dataset has records of various meteorological variables sampled every 6 minutes, starting on the 1st of January 2016 and ending on the 31st of December 2021, but the time series is incomplete. Therefore, it was decided

to work only with a subset of variables that have more than 85% of availability: wind speed (WSPD) [m/s], peak gust speed (GST) [m/s], air temperature (ATMP) [ºC] and atmospheric pressure (PRES) [bar].

Overall, the study involved 41 different time series: 10 of the WSPD variable, 10 of the GST, 12 of the ATMP and 9 of the PRES. As seen in Chapter 12, there is a strong correlation between the WSPD and GST variables. As a result, the WSPD and GST variables are used together for the reconstruction of the WSPD variable.

In this study, the year 2021 of the PPX station is reconstructed for the three variables WSPD/GST, ATMP and PRES. Since the observed data of 2021 are available, it is possible to validate the reconstruction done with PLSR.

13.3.2 Cross-Validation Methods

Figure 13.1 shows the values of Q^2 as a function of the number of latent variables, for the station PPX. The threshold to determine whether a latent variable should be used or not is essential, and Tenenhaus [114] chooses to set it at 0.0975. This means that for the reconstruction of the PPX station, one should use 2 LVs for the WSPD/GST variable, 3 for the ATMP variable, and 4 for the PRES variable.

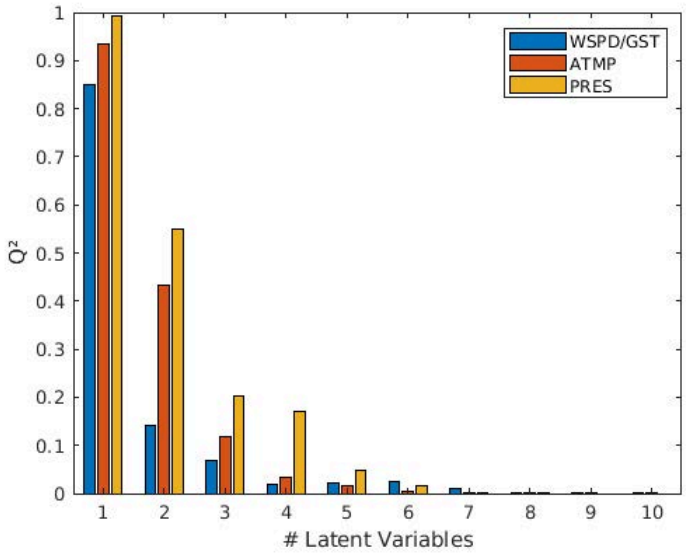


Figure 13.1: Values of Q^2 as a function of the number of latent variables

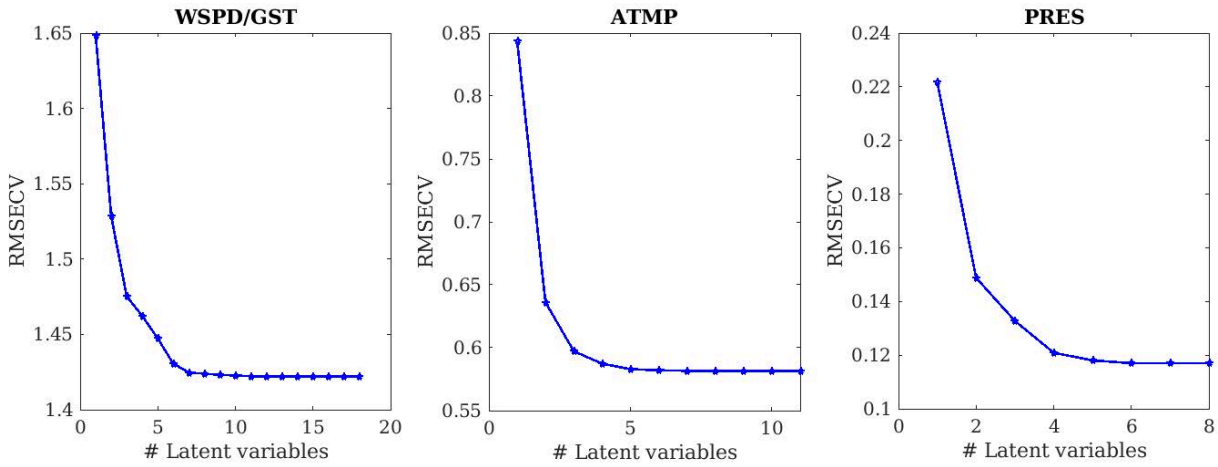


Figure 13.2: Values of RMSECV in function of the number of latent variables

Figure 13.2 shows the values of RMSECV as a function of the number of latent variables for each meteorological variable. It is not easy to determine the optimal number of latent variables. One can see a minimum for each variable, but should we keep adding latent variables until this minimum is reached? For cases like this, Wold [126] proposed further criteria such as "The First Minimum" or "Start of the Plateau". This last criterion is based only on graphic observation. The "plateau" is considered to have been reached when fluctuations in the value of RMSECV are no longer noticeable. According to these criteria, one could choose either 3 or 6 latent variables for the WSPD variable, 3 for the ATMP variable and 4 for the PRES variable.

The values of the repeated CV are shown in blue in Figure 13.3, along with the mean of these 10 values. This other criterion considers the variability of the RMSECV, and by taking the mean of all its values, one can estimate the error better. The data from the dataset used does not vary much, so the gaps between the observations are close to 0.

13.3.3 Wold's R

Figure 13.4 shows the values of R as a function of the number of LVs for each predicted variable, for the station PPX. According to this criterion, with a threshold of 0.9, one should use only 1 latent variable to reconstruct the WSPD variable, 2 for ATMP and 3 for PRES. If the threshold equals 0.95, one will use 2 LVs for WSPD/GST, 2 for ATMP and 4 for PRES.

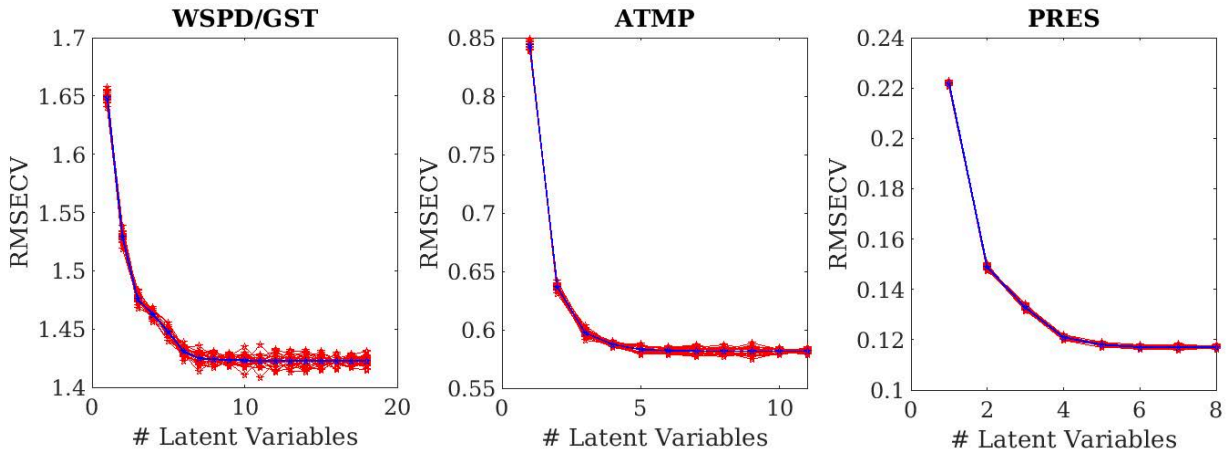


Figure 13.3: Repeated RMSECV for 10 cross-validation in function of the number of LVs

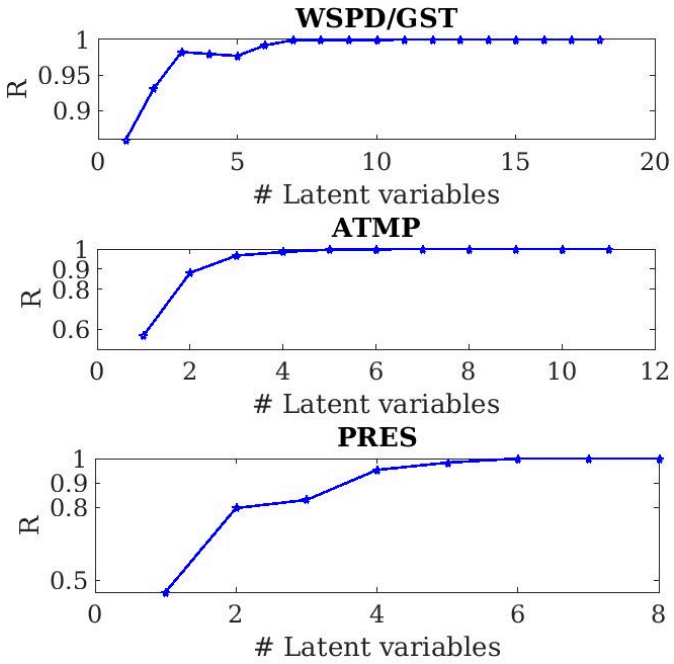


Figure 13.4: Values of Wold's R criterion in function of the number of LVs

13.3.4 Osten's F

Figure 13.5 shows the Osten's F test values for different number of LVs. If one wants a very accurate model, the criterion used to determine the number of latent variables should be when the Osten's F value is close to 0. In this case, one should use 7 variables for

WSPD/GST, 4 variables for ATMP and 5 for PRES.

It can be observed for WSPD/GST that after 3 LVs, the F value increases, which means that the 4th and the 5th latent variables don't influence the model, so they should not be used. Following the same rationale for these variables, one can use 3 LV for ATMP and 4 for PRES.

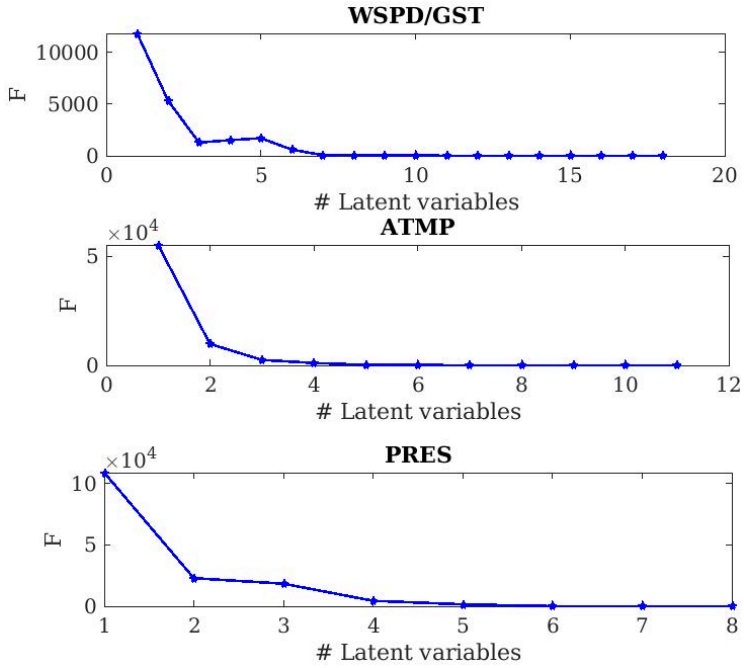


Figure 13.5: Values of Osten's F criterion in function of the number of LVs

13.3.5 Variable Importance in Projection (VIP)

Table 13.1 shows the values of the VIP metric for each variable. According to these values, one should use 7 LVs for the WSPD/GST variable, 4 for the ATMP, and 3 for the PRES (bold values).

Table 13.1: Values of the VIP variables for the prediction of each variable

LV	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
WSPD/GST	0.84	**1.06**	0.88	**1.13**	0.91	**1.09**	0.67	0.9	0.97	**1.20**	0.82	0.98	0.89	**1.18**	**1.09**	**1.55**	0.65
ATMP	0.98	**1.07**	0.79	0.95	0.85	**1.23**	**1.33**	0.84	0.82	0.92	**1.10**						
PRES	0.97	**1.02**	0.98	0.98	**1.03**	**1.06**	0.99	0.98									

13.3.6 Information Criteria (AIC and BIC)

The values of AIC and BIC for each variable as a function of the number of latent variables are depicted in Figure 13.6.

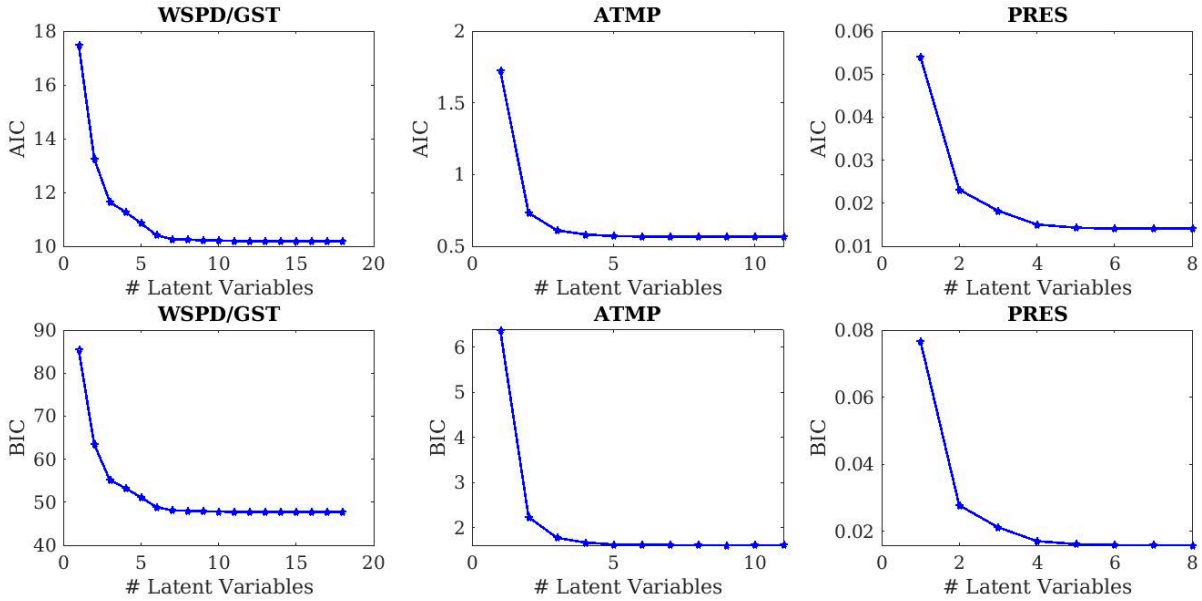


Figure 13.6: Values of AIC and BIC in function of the number of latent variables

The optimal number of latent variables varies between 4 and 7, depending on the meteorological variable. As with the other criteria, this criterion also indicates that for the WSPD/GST variable, the number of LVs is greater than for the other meteorological variables.

13.3.7 Comparison of the Selection Criteria

Table 13.2 recapitulates the optimal number of latent variables obtained from each criterion evaluated in this study and the time required for the computational execution of the corresponding criterion.

From a computational point of view, one can notice that it takes more time to use criteria based on cross-validation, especially for the repeated variant (in this case, up to two orders of magnitude). The other criteria consume similar computational time.

Table 13.2: Optimal number of LVs and computational time for the different criteria

Criterion	WSPD/GST	ATMP	PRES	Computational Time (s)
RMSECV	3 or 6	3	4	0.94
Repeated RMSECV	4	3	4	36.2
Q^2	2	3	4	0.87
Wold's R	3	4	4	0.89
Osten's F	3	3	4	0.92
VIP	7	4	3	0.92
AIC	3 or 6	3	4	0.89
BIC	3 or 6	3	4	0.89

13.3.8 Discussion on the Best Criteria

To compare the reconstruction of missing data with PLSR with different numbers of LVs, corresponding to various criteria, the meteorological variables WSPD, ATMP and PRES of the PPX station were reconstructed for 2021. Since the real values for that year are available, it is possible to validate the results. For this purpose, Table 13.3 presents three different types of errors. The Bias is a measure of the systematic error, SDE (Standard Deviation of the Error) represents the random error, and RMSE (Root Mean Square Error) corresponds to the sum of the two previous types of errors (for errors description see 10.3.1 of Chapter 10).

Table 13.3: BIAS, RMSE and SDE for different numbers of latent variables.

	WSPD			ATMP			PRES		
Number of LVs	BIAS	RMSE	SDE	BIAS	RMSE	SDE	BIAS	RMSE	SDE
1	-0.40	1.96	1.92	0.03	0.64	0.64	-0.04	0.20	0.19
2	-0.44	1.86	1.81	-0.02	0.60	0.60	0.01	0.13	0.13
3	**-0.41**	**1.80**	**1.76**	**0.01**	**0.53**	**0.53**	-0.04	0.12	0.11
4	-0.43	1.80	1.75	0.01	0.53	0.53	**-0.01**	**0.12**	**0.12**
5	-0.44	1.78	1.73	0.01	0.53	0.53	0.01	0.13	0.12
6	**-0.42**	**1.75**	**1.70**	0.01	0.53	0.53	0.05	0.13	0.12

Based on Table 13.3, one can notice that the reconstruction errors generally decrease as the number of latent variables increases, until a local minimum is reached, as one of the previous criteria hinted, so the following latent variables are adding more noise than actual useful information.

According to the results of the reconstruction of the PPX station with the PLS regression method, the best performance is achieved with 3 latent variables for the ATMP variable and 4 for PRES. For the WSPD variable, the experiments prove that one can choose RMSECV with the "First Minimum" criterion, which is reached at 3 LVs, or the "Start of the Plateau" criterion, at 6 LV.

The experiments also confirm that criteria such as RMSECV, Osten's F, AIC, and BIC give useful information for determining the optimal number of latent variables.

For the present study with meteorological data, the most indicated criteria are based on the RMSECV. Still, the main problem of cross-validation is its high computational time, especially when the number of observations is large, which is the case of the Leave One-Out Cross Validation (LOOCV) method with a large number of folds (k). For such situations, criteria such as the AIC or the BIC can be interesting, since they are very fast.

13.4 Concluding Remarks

The PLS regression method is fast and accurate for reconstructing missing data, but the crucial part is optimising its efficiency by choosing the optimal number of latent variables.

In this chapter, several criteria were presented and compared to find the optimal number of latent variables. After some experiments on meteorological data, some conclusions may be derived: the cross-validation techniques such as the RMSECV or the Osten's F criterion are efficient and give a relatively close amount of latent variables to the experimental optimal number; that is why these criteria are currently the most common in recent studies.

The information criteria, such as the AIC and the BIC, also give interesting results and are often used for prediction. However, the threshold is an essential factor to consider, and, at the moment, no studies have obtained a theoretical universal value for the threshold, meaning it is set experimentally.

Part III

Closure

Chapter 14

Conclusion

The concluding remarks for each research domain explored in the two parts of the manuscript are presented below.

Part 1 - Two-Dimensional Vortex Dynamics

This part of the thesis focuses on determining the optimal trajectory of a passive particle advected by a flow generated by N point vortices – ranging from one to three – between two specified points on the surface of a sphere. The formulation of the corresponding optimal control problem is presented. The control is formulated to minimize the squared L^2-norm of the control function, representing the energy expended over the prescribed displacement time interval.

The strategy adopted to solve the problem offers an alternative to classical methods in optimal control theory. It follows a direct approach in which the optimal control problem is first discretised and then numerically optimised. Specifically, the total time allocated for displacement is divided into a fixed number of subintervals, during each of which the control variables are assumed to be constant.

Through this discretisation, the original optimal control problem is reformulated as a constrained nonlinear optimisation problem (NOP), which is then solved using numerical algorithms such as the Interior Point or Active Set methods. Both approaches successfully identify optimal trajectories, independently of the discretisation level or the number of vor-

tices involved. Although both methods exhibit comparable computational costs, the Active Set method tends to require more time as the number of control parameters increases.

The equations governing the dynamics of the passive particle are presented in Cartesian and spherical coordinates. For each coordinate system, determining the particle's trajectory between two points is accomplished by the same direct approach to minimise overall control. The results obtained suggest that each formulation produces a feasible solution. In simpler terms, each model based on a different coordinate system allows the passive particle to attain the desired displacement within the allocated time.

These two formulations of the same problem confirm the robustness and effectiveness of the proposed direct method. The numerical results consistently produce optimal solutions, regardless of the coordinate system employed.

Generally, as the number of controls increases, there is greater freedom in the trajectory choice and the objective function's value, which corresponds to the energy used for the displacement, decreases. However, beyond a certain threshold, the objective function ceases to decrease and remains constant. However, the computation time increases with the number of controls.

Computation times generally rise when the problem is addressed using spherical coordinates and angular controls. Typically, the most efficient computational times are achieved through either the Cartesian formulation or the spherical formulation with Cartesian controls.

The direct approach to the optimal control problem allows for the incorporation of dynamics induced by multiple point vortices. In the present thesis, the dynamics are induced by one ($N = 1$), two ($N = 2$), three ($N = 3$), or four ($N = 4$) point vortices. In each case, the results show near/quasi-optimal control. It is also verified that as the number of vortices increases, the computation time also increases. This is because more calculations are required to solve the corresponding ordinary differential equations systems that govern the dynamics of the passive particle and point vortices.

It was also observed that variations in the sphere's rotation speed alter the nature of the solution, enabling displacement with either lower or higher energy expenditure. In general, higher rotation speeds require the particle to complete several revolutions around the sphere before reaching its target.

The proposed control method was also used to simulate the collection of ocean waste using autonomous technology. Passive particles model the autonomous technology responsible for trash transport, and the ocean current is generated by the motion of point vortices on a sphere. The findings focus on optimising the displacement of a group of passive particles (representing autonomous vehicles) interacting with vortices on a sphere's surface. The results indicate the existence of multiple trajectories for the transport of garbage particles. However, energy costs are lower for the most direct trajectories between the starting and destination points. This simplified conceptual model can be used for educational purposes and as an introduction to research activity since it blends concepts from physics and mathematics. It is also a low-cost alternative to experimental studies with real prototypes.

For future work, we would like to apply the methodology presented above to simulate the behaviour of passive particles subject to two competing effects: transport by a field generated by moving point vortices and a consensus mechanism that drives the particles to approach each other globally. The goal is to analyse how these competing influences affect particle distribution and dynamics.

Dynamical systems subject to consensus mechanisms arise in diverse real-world situations, both in technological and biological and medical contexts. For instance, ocean currents create vortices that unintentionally move the vehicles, such as underwater drones, used for ocean exploration and surveillance. The application of consensus and control strategies allows drones to maintain formation or follow specific trajectories, even in adverse conditions. Thus, studying dynamical systems subject to consensus mechanisms is of interdisciplinary interest, combining fluid dynamics, control and practical applications in technology and biology, from the ocean to the human body.

Part 2 - Meteorological Data Reconstruction

A common challenge in data analysis involves gaps – i.e., sequences of missing values – in time series, which can hinder or even prevent meaningful analysis. When the gaps are short, interpolation is typically employed to estimate the missing data. However, this approach becomes ineffective when the length of the gap exceeds the characteristic time scale that governs the predictability of the time series. In meteorology, one of the techniques used to reconstruct missing data is based on the analog ensembles method (AnEn).

The AnEn method plays an essential role in meteorological hindcasting. It involves

applying a forecast model to a past starting point to validate forecast models by comparing their predictions with available observations (reanalysis). If some observations are unavailable, their time series can be further complemented (reconstructed or predicted) using the corresponding forecasts. For example, a variable at a weather station can be reconstructed based on data from correlated variables from the same station and/or other nearby stations (predictors).

There is a consensus on the idea that the efficiency of the AnEn method improves with the increase in the size of the available dataset. However, the large amount of data that must be processed to determine analogues sometimes makes the computational cost prohibitive. Therefore, there is great interest in improving the computational efficiency of the AnEn method.

Part 2 of this thesis focuses on developing methodologies that allow using the AnEn method with large volumes of data. To this end, the robustness of the AnEn method (or AnEn-based methods) is combined with the Principal Component Analysis (PCA) or with the Partial Least Squares (PLS) for reducing the dimension of the predictor dataset. This approach enables the reduction of the predictor variables to a small number of new variables that mostly retain and may even enhance the original meteorological information.

When PCA is applied to the classical AnEn method, the information from several meteorological stations is combined in a reduced number of time series, corresponding to the principal components (PCs), which are then submitted to the AnEn method to reconstruct/predict the missing values. The combination of the PCA technique with the AnEn method results in a new methodology (PCAnEn) that proved, in our experiments, to offer reconstructions more accurate than the classical AnEn method, and allows for a considerable reduction in processing time.

In general, the location of the stations is of great importance because it promotes more correlation between variables and stations, thereby enhancing the effectiveness of the PCA technique. Thus, the choice of predictor stations must consider the proximity and correlation between them (which needs to be assessed before determining the PCs).

The results produced by the PLS-based techniques were slightly more accurate than those obtained with the PCA-based ones, especially in reconstructing (or predicting) meteorological variables with significant oscillation, such as wind speed. This happens because PLS builds the latent variables so that they simultaneously explain the variation of the

predictor variables and the predicted variable. In contrast, the main components, obtained by PCA, only explain the variation of the predictor variables.

Combining the AnEn method with PLS results in a hybrid method, PLSAnEn, which makes it possible to accurately reconstruct (or predict) the wind speed. It is therefore a highly suitable forecasting method for meteorological time series with potential applications in wind resource assessment and wind energy. At the same time, PLSAnEn is very demanding from a computational point of view, which benefits from a parallel implementation.

PLSClustAnEn, which combines the AnEn methods with the prior clustering of analogues, could be an alternative to the PLSAnEn method, as it is much more computationally efficient. However, it is a method that depends on many parameters that have to be chosen appropriately to improve the accuracy of the results.

Simultaneously with the AnEn-based methods, regression methods were also tested on the new variables determined by PCA or PLS. The resulting methods, PCR and PLSR, are fast and allow for accurate reconstructions. In particular, the PLSR method is the most suitable for reconstructing or forecasting highly correlated predictor variables such as atmospheric temperature or pressure.

Partial Least Squares (PLS) regression is a fast and reliable method for reconstructing missing data. However, its success depends mainly on the choice of the number of latent variables to be used. Several criteria were studied to find the optimal number of latent variables. After some experiments, it was observed that cross-validation techniques, such as the RMSECV or the Osten's F criterion, are efficient and give a relatively close amount of latent variables to the experimental optimal number. These criteria are, however, quite demanding from a computational point of view. Therefore, information-based criteria, such as AIC and BIC, are faster alternatives.

Furthermore, two different implementations of the reconstruction (prediction) methods studied were used and compared, one in MATLAB and the other in R. This allowed us to double-check the numerical results and gain insight into the potential performance impact of choosing either implementation. The scalability of both codes was also studied in a medium-scale multicore system. It is verified that the implementation in MATLAB is faster than the implementation in R.

As future work, we plan to apply the developed methodology to scientific domains

beyond meteorology, such as econophysics and bioinformatics. Econophysics is of particular interest, as it applies methods from physics to address problems in economics. In this field, time series of economic indicators—such as the CAC-40, AEX, BEL-20, and PSI-20 stock indices—are commonly used to estimate or forecast key economic variables.

Another avenue for future research involves adapting the previously developed methods to self-forecasting scenarios. This challenge arises when no external time series is available that differs from the one being reconstructed or predicted. In such cases, it is essential to identify a suitable training period from the historical data and analyse it to extract relevant features and patterns. These patterns can then be used to define analogous segments, which serve as the basis for prediction or reconstruction.

Bibliography

[1] H. Abdi. Partial least squares regression and projection on latent structure regression (PLS regression). *WIREs Computational Statistics*, 2(1):97–106, jan 2010.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 6:716 – 723, 1974.

[3] S. Alessandrini. Predicting rare events of solar power production with the analog ensemble. *Solar Energy*, 231:72–77, Jan. 2022.

[4] S. Alessandrini, L. D. Monache, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, 157:95–110, 2015.

[5] S. Alessandrini, L. D. Monache, S. Sperati, and J. N. Nissen. A novel application of an analog ensemble for short-term wind power forecasting. *Renewable Energy*, 76:768–781, 2015.

[6] J. Andrade, S. Boatto, and C. Vidal. Dynamics of restricted three and four vortices problem on the plane. *Journal of Mathematical Physics*, 57(4):042901, 2016.

[7] L. Araújo, C. Balsa, C. V. Rodrigues, and J. Rufino. Parametric Study of the Analog Ensembles Algorithm with Clustering Methods for Hindcasting with Multistations. In *Advances in Intelligent Systems and Computing*, pages 544–559. Springer International Publishing, 2021.

[8] H. Aref. Point vortex dynamics: A classical mathematics playground. *Journal of Mathematical Physics*, 48(6):065401, jun 2007.

[9] H. Aref. 150 years of vortex dynamics. *Theoretical and Computational Fluid Dynamics*, 24(1–4):1—7, Dec. 2009.

[10] H. Aref. Relative equilibria of point vortices and the fundamental theorem of algebra. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2132):2168–2184, 2011.

[11] A. Babiano, G. Boffetta, A. Provenzale, and A. Vulpiani. Chaotic advection in point vortex models and two-dimensional turbulence. *Physics of Fluids*, 6(7):2465–2474, 1994.

[12] C. Balsa, M. M. Breve, B. André, C. V. Rodrigues, and J. Rufino. PCAnEn - hindcasting with analogue ensembles of principal components. In *CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI)*, pages 169–183. Springer Nature Switzerland, 2023.

[13] C. Balsa, M. M. Breve, C. V. Rodrigues, L. S. Costa, and J. Rufino. *An Exploratory Study on Hindcasting with Analogue Ensembles of Principal Components*, pages 488–499. Springer Nature Switzerland, 2022.

[14] C. Balsa, M. M. Breve, C. V. Rodrigues, and J. Rufino. Reconstruction of meteorological records by methods based on dimension reduction of the predictor dataset. *Computation*, 11(5):98, May 2023.

[15] C. Balsa, H. Dupuis, M.-M. Breve, R. Guivarch, and J. Rufino. *Optimal Latent Variables Number for the Reconstruction of the Series with PLSR*, pages 193–205. Springer Nature Switzerland, 2024.

[16] C. Balsa and S. Gama. A control problem with passive particles driven by point vortices on the sphere. In *Communications in Computer and Information Science*, pages 139–150. Springer Nature Switzerland, 2022.

[17] C. Balsa and S. M. Gama. The control of the displacement of a passive particle in a point vortex flow. *Journal of Computational Methods in Sciences and Engineering*, 21(5):1215–1229, Nov. 2021.

[18] C. Balsa and S. M. A. Gama. A numerical algorithm for optimal control problems with a viscous point vortex. In *CONTROLO 2022*, pages 726–734. Springer International Publishing, 2022.

[19] C. Balsa, R. Monville-Letu, and S. Gama. Optimization of vortex dynamics on a sphere. In *CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI)*, pages 201–213. Springer Nature Switzerland, 2023.

[20] C. Balsa, M. Otero-Espinar, and S. Gama. Particle movement on a rotating sphere. In A. P. Aguiar, P. Rocha Malonek, V. H. Pinto, F. A. C. C. Fontes, and R. Chertovskih, editors, *CONTROLO 2024*, pages 320–328, Cham, 2025. Springer Nature Switzerland.

[21] C. Balsa, M. V. Otero-Espinar, and S. Gama. *A Simple Mathematical Model to Steering Oceanic Debris to a Targeted Region*, pages 58–70. Springer Nature Switzerland, Dec. 2023.

[22] C. Balsa, M. V. Otero-Espinar, and S. Gama. Exploring controlled passive particle motion driven by point vortices on a sphere. *Computation*, 12(2):23, Jan. 2024.

[23] C. Balsa, V. Otero-Espinar, and S. Gama. *An Approach to Environmental Cleanup Through Conceptual Autonomous Technology*, pages 582–591. Springer Nature Singapore, 2024.

[24] C. Balsa, C. V. Rodrigues, L. Araújo, and J. Rufino. Hindcasting with cluster-based analogues. In *Comm. in Computer and Inf. Science*, pages 346–360. Springer International Publishing, 2021.

[25] C. Balsa, C. V. Rodrigues, L. Araújo, and J. Rufino. Cluster-based analogue ensembles for hindcasting with multistations. *Computation*, 10(6):91, jun 2022.

[26] C. Balsa, C. V. Rodrigues, I. Lopes, and J. Rufino. Using analog ensembles with alternative metrics for hindcasting with multistations. *ParadigmPlus*, 1(2):1–17, Jun. 2020.

[27] A. J. Barbour and R. L. Parker. psd: Adaptive, sine multitaper power spectral density estimation for r. *Computers & Geosciences*, 63:1–8, 2014.

[28] J. Betts and I. Kolmanovsky. Practical methods for optimal control using nonlinear programming. *Applied Mechanics Reviews*, 55:B68, 2002.

[29] L. T. Biegler. *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. CAMBRIDGE UNIV PR, 2010.

[30] Y. Birkelund, S. Alessandrini, Ø. Byrkjedal, and L. D. Monache. Wind power prediction in complex terrain using analog ensembles. *Journal of Physics: Conference Series*, 1102(1):012008, oct 2018.

[31] A. C. Bissett. *Improvements to PLS Methodology*. PhD thesis, University of Manchester, 2015.

[32] V. A. Bogomolov. Dynamics of vorticity at a sphere. *Fluid Dynamics*, 12:863–870, 1977.

[33] B. Bonnard, O. Cots, and B. Wembe. A Zermelo navigation problem with a vortex singularity. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:S10, 2021.

[34] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[35] M. Breve, C. Balsa, and J. Rufino. Reconstruction of meteorological records with PCA-based Analog Ensembles methods. In A. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and V. Colla, editors, *Information Systems and Technologies WorldCIST 2023*. Springer, 2024.

[36] S. D. Burk and W. T. Thompson. The summertime low-level jet and marine boundary layer structure along the california coast. *Monthly Weather Review*, 124(4):668 – 686, 1996.

[37] C. M. Castellano and A. T. DeGaetano. Downscaling extreme precipitation from cmip5 simulations using historical analogs. *Journal of Applied Meteorology and Climatology*, 56(9):2421 – 2439, 2017.

[38] G. Cervone, L. Clemente-Harding, S. Alessandrini, and L. Delle Monache. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renewable Energy*, 108:274–286, 2017.

[39] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[40] P.-H. Chavanis and M. Lemou. Kinetic theory of point vortices in two dimensions: analytical results and numerical simulations. *The European Physical Journal B-Condensed Matter and Complex Systems*, 59(2):217–247, 2007.

[41] Y. Chen, T. Kolokolnikov, and D. Zhirov. Collective behaviour of large number of vortices in the plane. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2156):20130085, Aug. 2013.

[42] R. Chertovskih, D. Karamzin, N. T. Khalil, and F. L. Pereira. An indirect numerical method for a time-optimal state-constrained control problem in a steady two-dimensional fluid flow. In *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*, pages 1–6. IEEE, Nov. 2018.

[43] R. Chertovskih, N. T. Khalil, and F. L. Pereira. *Time-Optimal Control Problem with State Constraints in a Time-Periodic Flow Field*, pages 340–354. Springer International Publishing, 2020.

[44] A. Chesneau, C. Balsa, C. V. Rodrigues, and I. M. Lopes. Hindcasting with multistations using analog ensembles. In *CEUR Workshop Proceedings*, volume 2486, pages 215–229. CEUR-WS, 2019.

[45] A. J. Chorin. *Vorticity and turbulence*, volume 103. Springer Science & Business Media, 2013.

[46] S. D. Conte and C. De Boor. *Elementary numerical analysis: an algorithmic approach*. SIAM, 2017.

[47] F. Davò, S. Alessandrini, S. Sperati, L. D. Monache, D. Airoldi, and M. T. Vespucci. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. *Solar Energy*, 134:327–338, sep 2016.

[48] R. F. de Mello and M. A. Ponti. *Machine Learning*. Springer International Publishing, 2018.

[49] J. G. Derraik. The pollution of the marine environment by plastic debris: a review. *Marine pollution bulletin*, 44(9):842–852, 2002.

[50] M. T. Dibattista and L. M. Polvani. Barotropic vortex pairs on a rotating sphere. *Journal of Fluid Mechanics*, 358:107–133, Mar. 1998.

[51] H. M. V. D. Dool. A New Look at Weather Forecasting through Analogues. *Monthly Weather Review*, 117(10):2230–2247, 1989.

[52] D. G. Dritschel and S. Boatto. The motion of point vortices on closed surfaces. *Proc. R. Soc. A.*, 471, 2015.

[53] L. Eldén. *Matrix methods in data mining and pattern recognition.* SIAM, Philadelphia, PA, USA, 2007.

[54] M. A. Fosberg and M. J. Schroeder. Marine air penetration in central california. *Journal of Applied Meteorology and Climatology*, 5(5):573 – 589, 1966.

[55] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., USA, 1990.

[56] S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, June 1975.

[57] H. Helmholtz. Über integrale der hydrodynamischen gleichungen, welche den wirbelbewegungen entsprechen. *Journal für die reine und angewandte Mathematik*, 55:25–55, 1858.

[58] W. Hu, G. Cervone, A. Merzky, M. Turilli, and S. Jha. A new hourly dataset for photovoltaic energy production for the continental USA. *Data in Brief*, 40:107824, feb 2022.

[59] W. Hu, G. Cervone, G. Young, and L. D. Monache. Weather analogs with a machine learning similarity metric for renewable resource forecasting. *ArXiv*, abs/2103.04530, 2021.

[60] S. Hwang and S.-C. Kim. Point vortices on hyperbolic sphere. *Journal of Geometry and Physics*, 59(4):475–488, 2009.

[61] H. Jain and R. Jain. Big data in weather forecasting: Applications and challenges. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).* IEEE, mar 2017.

[62] J. R. Jambeck, R. Geyer, C. Wilcox, T. R. Siegler, M. Perryman, A. Andrady, R. Narayan, and K. L. Law. Plastic waste inputs from land into the ocean. *Science*, 347(6223):768–771, 2015.

[63] I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, 2 edition, 2002.

[64] S. Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, mar 1993.

[65] S.-C. Kim and S.-I. Sohn. Interactions of three viscous point vortices. *Journal of Physics A: Mathematical and Theoretical*, 45(45):15, 2012.

[66] Y. Kimura and H. Okamoto. Vortex motion on a sphere. *Journal of the Physical Society of Japan*, 56(12):4203–4206, 1987.

[67] J. King, X. Xue, W. Yao, and Z. Jin. A fast analysis of pesticide spray dispersion by an agricultural aircraft very near the ground. *Agriculture*, 12(3), 2022.

[68] G. R. Kirchhoff. Vorlesungenbër mathematische physik. *Mechanik*, 1876.

[69] K. Knight, S. Klein, and J. Duffie. A methodology for the synthesis of hourly weather data. *Solar Energy*, 46(2):109–120, 1991.

[70] P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.

[71] J. B. Lamb, B. L. Willis, E. A. Fiorenza, C. S. Couch, R. Howard, D. N. Rader, J. D. True, L. A. Kelly, A. Ahmad, J. Jompa, et al. Plastic waste associated with disease on coral reefs. *Science*, 359(6374):460–462, 2018.

[72] J. Liu and H. Hu. Biological inspiration: From carangiform fish to multi-joint robotic fish. *Journal of Bionic Engineering*, 7(1):35–48, mar 2010.

[73] E. N. Lorenz. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969.

[74] G. Marques, S. Gama, and F. L. Pereira. Optimal control of a passive particle advected by a Lamb-Oseen (viscous) vortex. *Computation*, 10(6):87, 2022.

[75] G. Marques, T. Grilo, S. Gama, and F. Pereira. Optimal control of a passive particle advected by a point vortex. In *International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability*, pages 512–523. Springer, 2021.

[76] W. F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, mar 1965.

[77] MathWorks. *Matlab Optimization Toolbox: User's Guide (R2020a)*. The MathWorks, Inc., Natick, Massachusetts, United State, 2020.

[78] MATLAB. *version 9.10.0.1602886 (R2021a)*. The MathWorks Inc., Natick, Massachusetts, 2021.

[79] V. Mayer-Schnberger. *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*. John Murray Publishers, London, GBR, 2013.

[80] S. Meech, S. Alessandrini, W. Chapman, and L. Delle Monache. Post-processing rainfall in a high-resolution simulation of the 1994 piedmont flood. *Bulletin of Atmospheric Science and Technology*, 1(3):373–385, Dec 2020.

[81] M. Meloun, J. Capek, P. Miksik, and R. Brereton. Critical comparison of method predicting the number of components in spectroscopic data. *Analytica Chimica Acta*, 423(1):51–68, 2000.

[82] D. Mercer. Clustering large datasets. Technical report, Linacre College, 2003.

[83] B.-H. Mevik and R. Wehrens. The pls package: Principal component and partialleast squares regression in r. *Journal of Statistical Software*, 18(2), 2007.

[84] I. I. Mokhov, S. G. Chefranov, and A. G. Chefranov. Point vortices dynamics on a rotating sphere and modeling of global atmospheric vortices interaction. *Physics of Fluids*, 32(10):106605, 2020.

[85] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight. Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, 141(10):3498–3516, 2013.

[86] L. D. Monache, T. Nipen, Y. Liu, G. Roux, and R. Stull. Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139(11):3554–3570, 2011.

[87] C. Nava-Gaxiola and J. Montaldi. Point vortices on the hyperbolic plane. *Journal of Mathematical Physics*, 55, 03 2014.

[88] P. K. Newton. *The N-vortex problem: analytical techniques*, volume 145. Springer Science & Business Media, 2001.

[89] P. K. Newton and H. Shokraneh. The n-vortex problem on a rotating sphere. i multi-frequency configurations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 462(2065):149–169, 2006.

[90] NOAA. National Data Buoy Center. `https://www.ndbc.noaa.gov`. Accessed on 28/07/2022.

[91] NOAA. What is an ocean glider? `https://oceanservice.noaa.gov/facts/ocean-gliders.html`, Jan. 2023. Accessed on 01/07/2023.

[92] D. Osten. Selection of optimal regression models via cross-validation. *J.Chemometrics*, 2:39 – 48, 1988.

[93] F. L. Pereira, T. Grilo, and S. Gama. Optimal power consumption motion control of a fish-like vehicle in a vortices vector field. In *OCEANS 2017 - Aberdeen*. IEEE, jun 2017.

[94] L. M. Polvani and D. G. Dritschel. Wave and vortex dynamics on the surface of a sphere. *Journal of Fluid Mechanics*, 255(-1):35, oct 1993.

[95] L. Pontryagin. *Mathematical Theory of Optimal Processes (Classics of Soviet Mathematics) (Volume 4)*. Gordon and Breach Science Publishers, 1986.

[96] M. J. Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis*, pages 144–157. Springer, 1978.

[97] B. Protas. Vortex dynamics models in flow control problems. *Nonlinearity*, 21(9):R203, 2008.

[98] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[99] E. Redish. Using math in physics: 4. toy models. *The Physics Teacher*, 59(9):683–688, dec 2021.

[100] L. Ricciardulli, G. R. Foltz, A. Manaster, and T. Meissner. Assessment of saildrone extreme wind measurements in hurricane sam using MW satellite sensors. *Remote Sensing*, 14(12):2726, jun 2022.

[101] C. M. Rochman. The complex mixture, fate and toxicity of chemicals associated with plastic debris in the marine environment. *Marine anthropogenic litter*, pages 117–140, 2015.

[102] C. M. Rochman. Strategies for reducing ocean plastic debris should be diverse and guided by science. *Environmental Research Letters*, 11(4):041001, 2016.

[103] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer Berlin Heidelberg, 2006.

[104] S. M. Ross. *Simulation*. Academic Press, 2022.

[105] C. M. Rozoff and S. Alessandrini. A comparison between analog ensemble and convolutional neural network empirical-statistical downscaling techniques for reconstructing high-resolution near-surface wind. *Energies*, 15(5):1718, Feb. 2022.

[106] D. Rutledge, J.-M. Roger, and M. Lesnoff. Different methods for determining the dimensionality of multivariate models. *Front. Anal. Sci. 1:754447*, 2021.

[107] A. Sabziparvar and F. Khoshhal Jahromi. Evaluating the most effective climatic parameters affecting the monthly mean soil temperature estimates using the pls method. *Arab J Geosci 15, 1044*, 2022.

[108] C. Schmid and L. Biegler. Quadratic programming methods for reduced hessian SQP. *Computers & Chemical Engineering*, 18(9):817–832, sep 1994.

[109] G. Schwarz. Estimating the dimension of the model. *Annals of Statistics, 6*, 2:461 – 464, 1978.

[110] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, V. A. Kamaev, et al. A survey of forecast error measures. *World applied sciences journal*, 24(24):171–176, 2013.

[111] E. Solomou, A. Pappa, I. Kioutsioukis, A. Poupkou, N. Liora, S. Kontos, C. Giannaros, and D. Melas. Analog ensemble technique to post-process WRF-CAMx ozone and particulate matter forecasts. *Atmospheric Environment*, 256:118439, jul 2021.

[112] L. Spence, A. Insel, and S. Friedberg. *Elementary Linear Algebra: A matrix Approach*. Pearson Education Limited, July 2013.

[113] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, jan 1974.

[114] M. Tenenhaus. *La Regression PLS - Théorie et Pratique.* Editions Technip, Paris, 1998.

[115] W. Thomson (Lord Kelvin). On vortex motion. *Trans. R. Soc. Edin*, 25:217–260, 1869.

[116] TOC. The Ocean Cleanup. `https://theoceancleanup.com/technology/`. Accessed on 01/07/2023.

[117] D. Vainchtein and I. Mezić. Vortex-based control algorithms. In *Control of fluid flow.*, pages 189–212. Berlin: Springer, 2006.

[118] H. van der Voet. Comparing the predictive accuracy of models using a simple randomisation test. *Chemometrics and Intelligent Laboratory Systems*, 2:313 – 323, 1994.

[119] H. van der Voet. Pseudo-degrees of freedom for complex predictive models: The example of partial least squares. *Journal of Chemometrics, 13*, 3-4:195–208, 1999.

[120] J. Vankerschaver and M. Leok. A novel formulation of point vortex dynamics on the sphere: Geometrical and numerical aspects. *Journal of Nonlinear Science*, 24(1):1–37, jul 2013.

[121] S. Vannitsem, J. B. Bremnes, J. Demaeyer, G. R. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts, S. Theis, A. Atencia, Z. B. Bouallègue, J. Bhend, M. Dabernig, L. D. Cruz, L. Hieta, O. Mestre, L. Moret, I. O. Plenković, M. Schmeits, M. Taillardat, J. V. den Bergh, B. V. Schaeybroeck, K. Whan, and J. Ylhaisi. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, mar 2021.

[122] W. Venables and B. Ripley. Modern applied statistics with s.berlin. *Springer-Verlag*, 2002.

[123] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical programming*, 107(3):391–408, 2006.

[124] H. Wold. Nonlinear iterative partial least squares (NIPALS) modelling: Some current developments. In *Multivariate Analysis–III*, pages 383–407. Elsevier, 1973.

[125] H. Wold. Nonlinear iterative partial least squares (nipals) modelling: Some current developments. *Multivariate Analysis-III*, pages 383–407, 1973.

[126] S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics 20*, 4:397 – 405, 1978.

[127] D. Yang. Ultra-fast analog ensemble using kd-tree. *Journal of Renewable and Sustainable Energy*, 11(5):053703, sep 2019.

[128] D. Yang and D. van der Meer. Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, 140:110735, apr 2021.

[129] D. Yang, W. Wang, J. M. Bright, C. Voyant, G. Notton, G. Zhang, and C. Lyu. Verifying operational intra-day solar forecasts from ECMWF and NOAA. *Solar Energy*, 236:743–755, apr 2022.

[130] E. Zermelo. Hydrodynamishe untersuchungen über die wirbelbewegungen in einer kugelfläche. *Math. Phys.*, 47:201, 1902.

[131] B. Zhang, Q. Tang, L. ping Chen, and M. Xu. Numerical simulation of wake vortices of crop spraying aircraft close to the ground. *Biosystems Engineering*, 145:52–64, 2016.

[132] X. Zhang, Y. Li, S. Lu, H. F. Hamann, B.-M. Hodge, and B. Lehman. A solar time based analog ensemble method for regional solar power forecasting. *IEEE Transactions on Sustainable Energy*, 10(1):268–279, 2019.

[133] X.-D. Zhang. *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore, 2020.

Appendix A

Copyrights

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

May 27, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela. ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6036960257850
License date	May 27, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	A Control Problem with Passive Particles Driven by Point Vortices on the Sphere
Licensed Content Author	Carlos Balsa, Sílvio Gama
Licensed Content Date	Jan 1, 2022
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes

Title of new work	Teacher - Researcher
Institution name	Instituto Politécnico de Bragança
Expected presentation date	Jul 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela.
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	1
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink

Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author,

Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040670626238
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Optimization of Vortex Dynamics on a Sphere
Licensed Content Author	Carlos Balsa, Raphaelle Monville-Letu, Sílvio Gama
Licensed Content Date	Jan 1, 2023
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99

Author of this Springer Nature content	yes
Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	2
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the

same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF

PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

computation

MDPI

Article

Exploring Controlled Passive Particle Motion Driven by Point Vortices on a Sphere

Carlos Balsa [1,*], M. Victoria Otero-Espinar [2] and Sílvio Gama [3]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, 5300-253 Bragança, Portugal
[2] Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain; mvictoria.otero@usc.es
[3] Mathematics Center of the Porto University (CMUP), Mathematics Department, Faculty of Sciences, University of Porto, R. Campo Alegre s/n, 4169-007 Porto, Portugal; smgama@fc.up.pt
* Correspondence: balsa@ipb.pt

Abstract: This work focuses on optimizing the displacement of a passive particle interacting with vortices located on the surface of a sphere. The goal is to minimize the energy expended during the displacement within a fixed time. The modeling of particle dynamics, whether in Cartesian or spherical coordinates, gives rise to alternative formulations of the identical problem. Thanks to these two versions of the same problem, we can assert that the algorithm, employed to transform the optimal control problem into an optimization problem, is effective, as evidenced by the obtained controls. The numerical resolution of these formulations through a direct approach consistently produces optimal solutions, regardless of the selected coordinate system.

Keywords: vortex; passive particle; spherical motion; control problem; nonlinear optimization problem; Cartesian and spherical coordinates

Citation: Balsa, C.; Otero-Espinar, M.V.; Gama, S. Exploring Controlled Passive Particle Motion Driven by Point Vortices on a Sphere. *Computation* **2024**, *12*, 23. https://doi.org/10.3390/computation12020023

Academic Editor: Markus Kraft

Received: 13 December 2023
Revised: 22 January 2024
Accepted: 28 January 2024
Published: 31 January 2024

1. Introduction

This study focuses on analyzing the movement of a passive particle within a flow generated by a specific configuration of point vortices on the surface of a sphere. Point vortices, which provide finite-dimensional approximations to the vortex dynamics of two-dimensional, incompressible ideal fluids, can be traced back to Helmholtz [1] and subsequent contributions by Kelvin [2] and Kirchhoff [3]. More recently, Hassan Aref was one of the world's leading researchers in the dynamics of point vortices (see, for instance, [4–6]).

Ongoing research in this domain integrates various disciplines, such as theories of dynamical systems, differential geometry, numerical analysis, optimal control, and more. Notably, the exploration of point vortices extends beyond conventional planes to encompass diverse surfaces, including the sphere [7–9], plane [10], and hyperbolic sphere [11–13].

Point vortices on the sphere are relevant as they provide a simplified representation of the behavior observed in specific geophysical flows where the curvature of the Earth plays a crucial role, and these flows persist over extended periods [9]. In fact, vortex point models play a pivotal role in addressing numerous fundamental dynamics questions related to atmospheric flows [7]. Conceptual models of point vortices are also used to identify and evaluate physical phenomena affecting the structure and interaction of atmospheric and oceanic vortices [14].

This work focuses on optimizing the displacement of a passive particle interacting with vortices located on the surface of a sphere. More specifically, our focus lies in the optimal control of the passive particle's displacement between two fixed points. The objective is to minimize the energy expended during the displacement while considering a fixed time for its completion. This problem can be conceptualized as a simplified model of

Copyrights

Copyright and Licensing

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.

Reproducing Published Material from other Publishers

It is absolutely essential that authors obtain permission to reproduce any published material (figures, schemes, tables or any extract of a text) which does not fall into the public domain, or for which they do not hold the copyright. Permission should be requested by the authors from the copyright holder (usually the Publisher, please refer to the imprint of the individual publications to identify the copyright holder).

Permission is required for:

1. Your own works published by other Publishers and for which you did not retain copyright.
2. Substantial extracts from anyone's works or a series of works.

4. Photographs for which you do not hold copyright.

Permission is not required for:

1. Reconstruction of your own table with data already published elsewhere. Please notice that in this case you must cite the source of the data in the form of either "Data from..." or "Adapted from...".
2. Reasonably short quotes are considered fair use and therefore do not require permission.
3. Graphs, Charts, Schemes and Artworks that are completely redrawn by the authors and significantly changed beyond recognition do not require permission.

Obtaining Permission

In order to avoid unnecessary delays in the publication process, you should start obtaining permissions as early as possible. If in any doubt about the copyright, apply for permission. MDPI cannot publish material from other publications without permission.

The copyright holder may give you instructions on the form of acknowledgement to be followed; otherwise follow the style: "Reproduced with permission from [author], [book/journal title]; published by [publisher], [year].' at the end of the caption of the Table, Figure or Scheme.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040670335028
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	An Approach to Environmental Cleanup Through Conceptual Autonomous Technology
Licensed Content Author	Carlos Balsa, Victoria Otero-Espinar, Sílvio Gama
Licensed Content Date	Jan 1, 2024
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99

Author of this Springer Nature content	yes
Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	4
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the

same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to

the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF

THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040670924277
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	A Simple Mathematical Model to Steering Oceanic Debris to a Targeted Region
Licensed Content Author	Carlos Balsa, M. Victoria Otero-Espinar, Sílvio Gama
Licensed Content Date	Jan 1, 2024
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99

Author of this Springer Nature content	yes
Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	5
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the

same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to

the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF

THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040671111635
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Particle Movement on a Rotating Sphere
Licensed Content Author	Carlos Balsa, M. Victoria Otero-Espinar, Sílvio Gama
Licensed Content Date	Jan 1, 2025
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99
Author of this Springer Nature content	yes

Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	6
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate

to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been

reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION

APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

computation

MDPI

Article

Cluster-Based Analogue Ensembles for Hindcasting with Multistations

Carlos Balsa [1,*,†], Carlos Veiga Rodrigues [2,†], Leonardo Araújo [3,†] and José Rufino [1,†]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, 5300-253 Bragança, Portugal; rufino@ipb.pt
[2] Vestas Wind Systems A/S, Vestas Technology Centre Porto, 4465-671 Leça do Balio, Portugal; calvr@vestas.com
[3] Universidade Tecnológica Federal do Paraná, Campus de Ponta Grossa, Ponta Grossa 84017-220, Brazil; leonardoa.2016@alunos.utfpr.edu.br
* Correspondence: balsa@ipb.pt
† These authors contributed equally to this work.

Abstract: The Analogue Ensemble (AnEn) method enables the reconstruction of meteorological observations or deterministic predictions for a certain variable and station by using data from the same station or from other nearby stations. However, depending on the dimension and granularity of the historical datasets used for the reconstruction, this method may be computationally very demanding even if parallelization is used. In this work, the classical AnEn method is modified so that analogues are determined using K-means clustering. The proposed combined approach allows the use of several predictors in a dependent or independent way. As a result of the flexibility and adaptability of this new approach, it is necessary to define several parameters and algorithmic options. The effects of the critical parameters and main options were tested on a large dataset from real-world meteorological stations. The results show that adequate monitoring and tuning of the new method allows for a considerable improvement of the computational performance of the reconstruction task while keeping the accuracy of the results. Compared to the classical AnEn method, the proposed variant is at least 15-times faster when processing is serial. Both approaches benefit from parallel processing, with the K-means variant also being always faster than the classic method under that execution regime (albeit its performance advantage diminishes as more CPU threads are used).

Keywords: hindcasting; meteorological dataset; analogue ensemble; K-means; time-series

Citation: Balsa, C.; Rodrigues, C.V.; Araújo, L.; Rufino, J. Cluster-Based Analogue Ensembles for Hindcasting with Multistations. *Computation* **2022**, *10*, 91. https://doi.org/10.3390/computation10060091

Academic Editor: Gennady Bocharov

Received: 29 April 2022
Accepted: 27 May 2022
Published: 2 June 2022

1. Introduction

Short-term weather predictions by correlation with similar states in the past (analogues) were originally established by Lorenz [1], who suggested that two atmospheric states that are initially very close to each other will remain somewhat similar in the future. This was introduced as an alternative to classical weather forecasting based on systems of equations underlying deterministic Numerical Weather Prediction (NWP) models.

For many years, however, Lorenz's proposal was discarded because of limited historical data on past weather conditions (especially over wide geographical areas) and insufficient computing capacity to implement his approach. Two decades later, van den Dool [2] revisited analogue-based short-range weather forecasting and found it to be feasible and effective when applied to limited geographical areas.

Monache [3] showed the applicability of an analogue scheme (named AN) for post-processing numerical weather forecasts to reduce systematic and random errors. The basic idea is that if previous forecasts (analogues) exist that are similar to the current NWP forecast (predictor), it is possible to produce a AN forecast by averaging the observations corresponding to these previous forecasts. The analogue prediction is then compared with the NWP prediction to infer the prediction error and thus improve the NWP forecast.

Copyrights

Copyright and Licensing

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.

Reproducing Published Material from other Publishers

It is absolutely essential that authors obtain permission to reproduce any published material (figures, schemes, tables or any extract of a text) which does not fall into the public domain, or for which they do not hold the copyright. Permission should be requested by the authors from the copyright holder (usually the Publisher, please refer to the imprint of the individual publications to identify the copyright holder).

Permission is required for:

1. Your own works published by other Publishers and for which you did not retain copyright.
2. Substantial extracts from anyone's works or a series of works.

4. Photographs for which you do not hold copyright.

Permission is not required for:

1. Reconstruction of your own table with data already published elsewhere. Please notice that in this case you must cite the source of the data in the form of either "Data from..." or "Adapted from...".
2. Reasonably short quotes are considered fair use and therefore do not require permission.
3. Graphs, Charts, Schemes and Artworks that are completely redrawn by the authors and significantly changed beyond recognition do not require permission.

Obtaining Permission

In order to avoid unnecessary delays in the publication process, you should start obtaining permissions as early as possible. If in any doubt about the copyright, apply for permission. MDPI cannot publish material from other publications without permission.

The copyright holder may give you instructions on the form of acknowledgement to be followed; otherwise follow the style: "Reproduced with permission from [author], [book/journal title]; published by [publisher], [year].' at the end of the caption of the Table, Figure or Scheme.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040671349758
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	PCAnEn - Hindcasting with Analogue Ensembles of Principal Components
Licensed Content Author	Carlos Balsa, Murilo M. Breve, Baptiste André et al
Licensed Content Date	Jan 1, 2023
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99

Author of this Springer Nature content	yes
Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	7
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the

same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to

the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF

THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040671434833
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Reconstruction of Meteorological Records with PCA-Based Analog Ensemble Methods
Licensed Content Author	Murilo M. Breve, Carlos Balsa, José Rufino
Licensed Content Date	Jan 1, 2024
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99
Author of this Springer Nature content	yes

Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	8
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink

Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author,

Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

computation

MDPI

Article

Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset

Carlos Balsa [1,*], Murilo M. Breve [1], Carlos V. Rodrigues [2] and José Rufino [1]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal; murilo.breve@ipb.pt (M.M.B.); rufino@ipb.pt (J.R.)
[2] Vestas Wind Systems A/S, Design Centre Porto, 4465-671 Leça do Balio, Portugal; calvr@vestas.com
* Correspondence: balsa@ipb.pt

Abstract: The reconstruction or prediction of meteorological records through the Analog Ensemble (AnEn) method is very efficient when the number of predictor time series is small. Thus, in order to take advantage of the richness and diversity of information contained in a large number of predictors, it is necessary to reduce their dimensions. This study presents methods to accomplish such reduction, allowing the use of a high number of predictor variables. In particular, the techniques of Principal Component Analysis (PCA) and Partial Least Squares (PLS) are used to reduce the dimension of the predictor dataset without loss of essential information. The combination of the AnEn and PLS techniques results in a very efficient hybrid method (PLSAnEn) for reconstructing or forecasting unstable meteorological variables, such as wind speed. This hybrid method is computationally demanding but its performance can be improved via parallelization or the introduction of variants in which all possible analogs are previously clustered. The multivariate linear regression methods used on the new variables resulting from the PCA or PLS techniques also proved to be efficient, especially for the prediction of meteorological variables without local oscillations, such as the pressure.

Keywords: hindcasting; forecasting; analog ensemble; principal component analysis; partial least square; multivariate regression

Citation: Balsa, C.; Breve, M.M.; Rodrigues, C.V.; Rufino, J. Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset. *Computation* **2023**, *11*, 98. https://doi.org/10.3390/computation11050098

Academic Editor: Shengkun Xie

Received: 29 March 2023
Revised: 7 May 2023
Accepted: 9 May 2023
Published: 12 May 2023

1. Introduction

Filling gaps in observed time series is an important problem in many areas of applied sciences that depend on data analysis. Without this filling, data reconstruction is difficult or even impossible. This assumption is particularly true in weather forecasting, where the amount of stored information is growing four times faster than the world economy [1]. In view of this, big data analytics can help to improve predictions by uncovering patterns and correlations in the data [2] and reconstructing missing data in areas where there is limited information. Conversely, this growth in data also means that the amount of missing data is increasing, which makes accurate reconstruction a crucial task. To handle this challenge, forecasting methods must be able to handle large amounts of data, multiple data sources and a wide variety of meteorological variables. This requires advanced methodologies that can adapt to the particular characteristics of big data in weather forecasting.

Despite the general abundance of weather data available, there are still many regions without historical data records. These locations, which may be remote or under-developed, have the potential to be significant generators of renewable energy. However, without historical weather data, it is difficult to accurately predict the potential for energy generation in such places. Therefore, there is a growing need for methods that can generate weather data from limited inputs and locations, with the purpose of running simulations of environmentally driven systems that target these locations. This may greatly enhance our understanding of the potential for renewable energy generation, and may facilitate the development of sustainable energy systems in such regions [3].

Copyrights

Copyright and Licensing

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.

Reproducing Published Material from other Publishers

It is absolutely essential that authors obtain permission to reproduce any published material (figures, schemes, tables or any extract of a text) which does not fall into the public domain, or for which they do not hold the copyright. Permission should be requested by the authors from the copyright holder (usually the Publisher, please refer to the imprint of the individual publications to identify the copyright holder).

Permission is required for:

1. Your own works published by other Publishers and for which you did not retain copyright.
2. Substantial extracts from anyone's works or a series of works.

4. Photographs for which you do not hold copyright.

Permission is not required for:

1. Reconstruction of your own table with data already published elsewhere. Please notice that in this case you must cite the source of the data in the form of either "Data from..." or "Adapted from...".
2. Reasonably short quotes are considered fair use and therefore do not require permission.
3. Graphs, Charts, Schemes and Artworks that are completely redrawn by the authors and significantly changed beyond recognition do not require permission.

Obtaining Permission

In order to avoid unnecessary delays in the publication process, you should start obtaining permissions as early as possible. If in any doubt about the copyright, apply for permission. MDPI cannot publish material from other publications without permission.

The copyright holder may give you instructions on the form of acknowledgement to be followed; otherwise follow the style: "Reproduced with permission from [author], [book/journal title]; published by [publisher], [year].' at the end of the caption of the Table, Figure or Scheme.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 02, 2025

This Agreement between Carlos Balsa / Universidade de Santiago de Compostela ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	6040680154524
License date	Jun 02, 2025
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Optimal Latent Variables Number for the Reconstruction of Time Series with PLSR
Licensed Content Author	Carlos Balsa, Hugo Dupuis, Murilo-M. Breve et al
Licensed Content Date	Jan 1, 2024
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99

Author of this Springer Nature content	yes
Title of new work	PhD in Mathematics and Applications
Institution name	Universidade de Santiago de Compostela
Expected presentation date	Sep 2025
The Requesting Person / Organization to Appear on the License	Carlos Balsa / Universidade de Santiago de Compostela
Requestor Location	Mr. Carlos Balsa EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, 5370-202 Portugal
Order reference number	10
Billing Type	Invoice
Billing Address	Instituto Politécnico de Bragança EsACT - Instituto Politécnico de Bragança Campus do Cruzeiro Avenida 25 de Abril, Cruzeiro, Lote 2. Mirandela, Portugal 5370-202
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the

same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to

the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF

THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

This research is developed within two main areas of Applied Computational Mathematics: Simulation and Data Analysis.

In the simulation component, the study investigates the motion of a passive particle on a spherical surface under the influence of multiple point vortices. The aim is to find an energy-minimising trajectory that respects constraints such as specified departure and arrival points, as well as a fixed travel time.

In the data analysis component, the research focuses on reconstructing incomplete meteorological time series by addressing gaps caused by missing records. The proposed methodology seeks to maximise the use of available information from predictor stations while ensuring both high accuracy and computational efficiency.