ENVIHEI

Title of the PBL challenge: **Data Engineering Meets Climate Science: Exploring CMIP6 Through Problem-Based Learning**

Name of the responsible lecturer

Andrés Navarro Martínez de la Casa

University / Department / Degree program

University of León / Applied Chemistry and Physics / Bachelor's Degree in Data Engineering and Artificial Intelligence

Course / Subject in which implemented

4th Course / Projects in Data Engineering and Artificial Intelligence

Implementation period (dates)

September 10, 2025 – October 31, 2025

Total duration (weeks/hours)

8 weeks / 30 hours

Number of students involved

7 students

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Spanish Service for the Internationalization of Education (SEPIE). Neither the European Union nor SEPIE can be held responsible for them. [2024-1-ES01-KA220-HED-000246733]

1. Project Design

1.1 Brief description of the project and the challenge proposed

The course is a Problem-Based Learning (PBL) experience designed for data engineering students. The core challenge involves applying the full data engineering lifecycle – ingestion, cleaning, analysis, and modelling – to real-world climate problems. Students will process gigabytes of data from climate models (CMIP6) and reanalysis (ERA5) to answer complex sustainability questions, such as detecting climate anomalies, classifying bioclimatic zones using unsupervised machine learning, and evaluating the uncertainty of future climate projections through multi-model ensembles.

1.2 Expected learning outcomes (academic and transversal competences)

Students are expected to demonstrate proficiency in both advanced data engineering and climate science applications. The academic learning outcomes focus on designing and implementing automated ETL (Extraction, Transform, Load) pipelines for high-dimensional geospatial data. Students learn to integrate satellite and modelled data APIs (Copernicus CDS and ESGF) with efficient storage formats. Students also learn to apply unsupervised machine learning techniques – specifically dimensionality reduction (PCA) and clustering (K-Means) – to discover latent spatiotemporal patterns, while rigorously quantifying predictive uncertainty through statistical analysis of multi-model ensembles (CMIP6).

Transversally, the project promotes the ability to manage uncertainty in decision-making processes, which is a critical skill when dealing with probabilistic climate models. Students also master reproducible science standards by using version control systems and documentation best practices.

Finally, the course develops strong data storytelling skills, enabling students to synthesize complex, multidimensional datasets into actionable visual insights that inform diverse stakeholders.

1.3 Targeted sustainability competences

This project directly addresses the targets of three Sustainable Development Goals (SDGs), according to the guidelines of the Sustainable Development Solutions Network (SDSN/REDS):

- **SDG 13 - Climate Action (Main Goal)**: Technical training to analyse essential climate variables and understand the physical basis of climate change using observed and modelled data.
- **SDG 9 - Industry, Innovation, and Infrastructure**: Fostering scientific research and technological capacity through the use of 'big data' and artificial intelligence applied to sustainability.
- **SDG 4 - Quality Education**: Acquisition of advanced digital skills and promoting public access to climate information (Open Data/Open Science).

1.4 Sustainability approach and how it is integrated into the project

Sustainability is not an add-on, but rather the core material of the course. Each technical module is linked to a sustainability issue: calculating climate normals for climate adaptation and resilient infrastructure (ASS01), using ensembles to manage uncertainty in predictions (ASS02), and redefining climate zones under global change scenarios (ASS04). These connections force students to understand the physics of climate change in order to develop effective data solutions.

1.5 Methodologies and tools used (PBL structure, facilitation techniques, etc.)

The Driving Challenge. The project is anchored in a central open-ended question that guides the entire learning process: How can we translate raw, petabyte-scale climate data into actionable insights for sustainability in a changing world? Students act as data engineers tasked with 'decoding' the climate system to address this challenge.

PBL Structure & Scaffolding. The methodology follows a progressive scaffolding approach, moving from guided inquiry to autonomous problem-solving across three phases (foundation, inquiry, and resolution phases):

- **Foundation Phase (The Toolkit)**. Students are introduced to the ecosystem (Linux/Conda) and learn to automate data ingestion using APIs. The focus is on technical competency.
- **Inquiry Phase (The Physics of data)**. Students tackle specific sub-problems (normals, anomalies, and ensembles), which require an understanding of the underlying physical concepts to write the correct code. This phase connects statistics with sustainability.
- **Resolution Phase (The Capstone)**. The final 'Climate Classification System (CCS)' project removes the training wheels. Students must integrate all previous modules (data fetching, cleaning, statistical analysis, and machine learning) to propose a novel solution to the driving challenge.

Facilitation Techniques:

- Just-in-Time Teaching (JiTT): JiTT introduces theoretical concepts (e.g., ensemble, anomaly, PCA) when students need them to solve a coding problem, which maximizes relevance and retention.

- Coding Dojos: Practical sessions in which the lecturer codes live to demonstrate debugging strategies and best practices that students then replicate and extend.
- Peer Code Reviews (In-Situ): Instead of traditional grading, students receive feedback through ‘over-the-shoulder’ code reviews.

Tools & Ecosystem. The technical implementation relied on an industry-standard open-source stack.

- Environment: Linux/WSL with Conda for package management.
- Core Libraries: Python (Xarray, Pandas, Matplotlib, and Cartopy) for data manipulation and visualization.
- Climate-specific tools: CDO (Climate Data Operators) for high-performance NetCDF processing and GDAL for geospatial raster operations.
- Data access: Automated retrieval via Copernicus CDS API and ESGF nodes.
- Collaboration: Git and GitHub are used for version control and project submission.

1.6 Organization of the project (group formation, timeline, milestones)

8 weeks (30 contact hours + autonomous work):

- **Week 1**: Fundamentals, tools, and data acquisition (APIs).
- **Weeks 2-4**: Climate statistics (normals, anomalies, and ensembles). Individual assignments.
- **Weeks 5-7**: Advanced Machine Learning (CCS Clustering). Students worked in small groups, replicating a professional development environment.
- **Week 8**: Final integration and oral presentations.

1.7 Deliverables requested from students (type, format, deadlines)

For each module (ASS01-ASS04), the following were required:

- **Code**: Clean and documented Python scripts or Notebooks in the repository and/or in Moodle.
- **Technical Report**: PDF document including methodology, visualizations, and discussion of results.
- **Visualizations**: Maps or plots generated with Cartopy/Matplotlib for each assignment.
- **Final Presentation**: Oral defence of the CCS project. Only for the collaborative project (ASS04).

Deadlines: One deliverable per week (seven days after the session), except for ASS04, which is a four-week collaborative project.

1.8 External collaborators involved (mentors, companies, experts...) and their role

No external collaborators were involved in this implementation.

2. Project Evaluation

2.1 Assessment criteria and methods used

The evaluation was continuous and based on specific rubrics for each assignment, with the typical weighting is as follows:

- **Technical correctness and data analysis (40%)**: Accuracy in calculations (e.g., anomalies, seasonal means).
- **Code quality and reproducibility (20%)**: Structure, comments, and efficient use of libraries.
- **Quality of visualizations (20%)**: Communicative clarity of maps and graphs.

- **Report and communication (20%)**: Ability to synthesize and interpret mathematical results.

Details on the specific rubrics can be found in the GitHub repository: https://github.com/navarro-esm/ULE_data-ai-projects/tree/main

2.2 Evidence of student learning (examples, outputs, results)

The final results demonstrate a high level of proficiency in applying data engineering techniques to climate science. Students successfully transitioned from basic scripts to complex, reproducible workflows involving dimensionality reduction (PCA) and unsupervised clustering (K-Means).

Below are direct links to the projects developed by the students, which serve as concrete evidence of the learning outcomes:

- **Group A (Agricultural Vulnerability)**: https://github.com/gortif00/CMIP6_Climate_Classification
- **Group B (Solar Farm):** https://github.com/calejs00/pid
- **Group C (Bear distribution):** https://github.com/SergioAlvesMar5/CMIP6_Climate_Classification_IvanFernandez_SergioAlves/

3. Reflections and Recommendations

3.1 Main achievements and strengths of the project

The main achievement has been the successful integration of 'hard' data engineering with sustainability science. Students learned to recognize the cross-cutting nature of climate: they stopped viewing datasets merely as matrices of numbers and understood their social and environmental impact. Students developed a proactive attitude, solving posed problems and proposing new projects that link engineering and climate, thereby demonstrating a deep understanding of the subject matter.

3.2 Challenges faced and how they were addressed

The challenges encountered were typical of a newly implemented subject, especially since this is the first cohort of students to reach the fourth year of the degree program. As a pilot program with an entirely new group of students, the course required adjustments to technical workflows and pedagogical pacing in real-time to ensure that the learning objectives were met without overwhelming the students. Specifically, two main technical hurdles were addressed:

- **Heterogeneity of Environments**: Managing software environments (Conda) across different operating systems (Windows, Mac, Linux) was complex. Critical tools like CDO had compatibility issues on Windows, but these were resolved by guiding students to use Linux subsystems (WSL) or virtual machines.
- **Data Access Stability**: External repositories (Copernicus, ESGF) occasionally experienced high latency or downtime. To mitigate this issue, deadlines were kept flexible, and local backup datasets (caches) to prevent any blockage in the students' learning progress.

3.3 Lessons learned for future implementations

One of the key lessons learned from this pilot implementation is the importance of immediate engagement to overcome initial scepticism. Although students were hesitant at first, their perception shifted significantly once they started working with real data. Therefore, future editions will introduce high-impact use cases, such as recent extreme weather events, right from the first session to capture the students' interest immediately. Furthermore, student feedback emphasized the need for greater methodological flexibility. Future iterations will give students more freedom in selecting statistical-mathematical techniques to solve problems, encouraging more creative and less guided approaches once the foundational skills are mastered.

4. Supporting Material

4.1 Lecturers' materials and project documentation

Lecturers' materials and project documentation are available in the following link:

https://drive.google.com/drive/folders/17N07p7YOHd7lprLZGN759cnp5EkDBrMi?usp=drive_link

4.2 Deliverables produced by students, final presentations or pitch materials

An example of deliverables produced by students is available in the following link:

https://drive.google.com/drive/folders/1E7oaA-tSpN8NlM2xiN-W1QScFyMKex4I?usp=drive_link

Students' final presentations are available in the following link:

https://drive.google.com/drive/folders/1eSmrix7Wk8tS65pVgI1bX6tDfG6RXmAL?usp=drive_link

4.3 Images/videos (with consent)

Photographic and video records were captured with the explicit consent of the participating students, and are available in the following link:

https://drive.google.com/drive/folders/1S7-AQBQStS3EH22-L_-x_ZpvM12czDyx?usp=drive_link

4.4 Any other relevant material

All the materials, including this report, are available in the following link:

https://drive.google.com/drive/folders/1lOJfNdDqIVrzm_YZyGSwbdy6fbT08z4m?usp=drive_link

Erasmus+
Enriching lives, opening minds.
Co-funded by the European Union
P. No.: 2024-1-ES01-KA220-HED-000246733