

|  |
| --- |
| Coventry University  5011CEM Big Data Programming Project Specification Document  visualization and simple Ensemble comparison |

**Author: Armandas Barkauskas**

**Document Version: 1.4**

**19/03/2020**

**Table of Contents**

[1. Introduction 3](#_Toc30950202)

[2. Project Requirements 3](#_Toc30950203)

[2.1 Related documents 3](#_Toc30950204)

[2.2 Terms/Acronyms and Definitions 3](#_Toc30950205)

[3. Risks and Assumptions 3](#_Toc30950206)

[4. Out of Scope 3](#_Toc30950207)

[5. System/ Solution Overview 3](#_Toc30950208)

[5.1 Context Diagram/ Interface Diagram/ Data Flow Diagram, Application Screen Flow, Sitemap, Process Flow 3](#_Toc30950209)

[6. Project Management 4](#_Toc30950210)

[7. References 4](#_Toc30950211)

[8. Open Issues 4](#_Toc30950212)

[Appendix 4](#_Toc30950213)

# Introduction

The aim of this project is to help the scientists to visualize data models and make

This aim of this project is to use multiple observation models for Europe area and provide more accurate results. The first approach of doing this is simple mean ensemble, which is basically taking an average of all models combined for each location point. According to, Hyde, it was found that the accuracy of model ensembles can be improved up to 18% when using the DDC clustering algorithm (Hyde and Angelov 2014). Project will involve using the sub-spacing and parallel processing concepts as running big data processes sequentially takes a long time. Ultimately, both ensembles and comparison results will be plotted in a simple map for a visualisation part. More information about DDC algorithm can be found at <https://ieeexplore.ieee.org/document/6930157>.

# Project Requirements

Originally, there are 14 climate models worldwide, but for this project only 7 of them are going to be used due to the long execution time. The data is stored in standard format that is used by climate scientists. Additionally, we are going to analyse Europe area only (in 0.1-degree precision). It is required to use big data techniques to reduce the processing time to less than 2 hours. These techniques will include sub-spacing, parallel processing, display of the big data. The flowchart depicts the core idea of program flow and expected working principles. All the major steps done, will be recorded in the logbook in the excel format.

## Related documents

|  |  |  |
| --- | --- | --- |
| **Component** | **Name (with link to the document)** | **Description** |
| Code base | <https://github.coventry.ac.uk/barkausa/big-data> | Program code for Matlab excluding the big data files |
| DDC report | <https://ieeexplore.ieee.org/document/6930157> | Find out more about the DDC algorithm from its authors |
| Flowchart | /flowchart.drawio and /images/flowchart.png | Find the program flowchart as a project file or image |
| Installation guide | /README.md | Contains short description from coding perspective |

## Terms/Acronyms and Definitions

|  |  |  |
| --- | --- | --- |
| **Term/Acronym** | **Definition** | **Description** |
| observation models |  | Model is a measurement made by a different organization, which is not completely accurate |
| SME | Simple mean ensemble | An average of all provided models (in our case 7) for every location point within Europe |
| DDC algorithm | Data density clustering algorithm | An algorithm that clusters data points. Its output is being used to create a CBE |
| CBE | Cluster based ensemble | More accurate ensemble then SME, where generation of it is much more complicated than SME |
| Sub-spacing |  | Dividing a current data set to smaller data chunks |
| Parallel processing |  | Usually, splitting independent tasks to run on a different thread to execute faster. Closely related to sub-spacing. |
| flowchart |  | Depicts the main logical program steps visualized using a special symbol syntax |
| logbook |  | An excel sheet, were the major project steps are being recorded |

# Risks and Assumptions

Learning a new language like Python or Matlab to do the project takes which does not produce any working result at that time. The hardware processor will have up to 8 cores due to the money limitation, meaning that the performance still can be scaled horizontally later. There is a low chance of a need for strong graphic card. For this project we are going to install Matlab with addons, which takes huge amount of disk space and processing power while installing. Data is encoded in an unusual data format: netCDM. Additional, data files are huge in size, so we need to handle them well, to make the program run fast and reliable. Any multiple processor computer should work fine for this project. This is needed because we are going to use a parallel processing technique to reduce the execution time. Also, we need to handle corrupted files in case of an accident. So of the program parts might include 3-rd party code, which may have bugs or can be hard to understand. ~~Data must be normalized to make use of the DDC algorithm~~

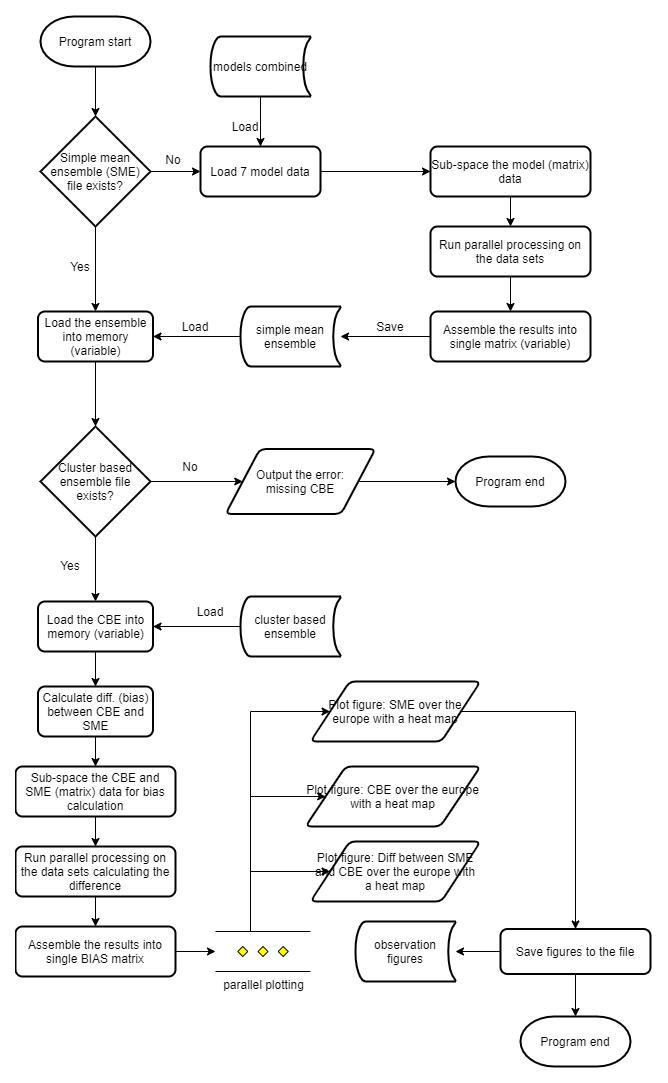
# Out of Scope

We are not going to create a CBE as our goal is to compare ensembles rather than generate a CBE itself. Instead, CBE data is being given in “.csv” format, which will be converted into “.mat” matrix file. We will not handle the data bigger than 1GB size in this project as it causes long wait times. Thus, this might produce slightly imprecise results.

# System/ Solution Overview

The program is going to start by loading the Simple Mean Ensemble (SME) and Cluster Based Ensemble (CBE) models. These models are going to be used to calculate the difference between the ensembles. All results are going to be plotted on the map and figures will be saved.

# Context Diagram/ Interface Diagram/ Data Flow Diagram, Application Screen Flow, Sitemap, Process Flow

****

# Project Management

* Logbook
* Logical folder/file structure
* Git version control system
* Flowchart

The main project management will be handled by logbook. This is the place where project development will be explained in the timestamped steps. Besides, to run project smoothly there must be a logical folder structure. All data will be held under the control of Git/Github. Ultimately, the core logic of the program is explained in the Data Flow Diagram.

# References

* Matlab documentation: <https://www.mathworks.com/help/matlab/>
* Data density-based clustering: <https://ieeexplore.ieee.org/document/6930157>
* University material/ data source

# Open Issues

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Issue** | **Raised On** | **Solution/ Decision** | **Resolved On** | **Status** |
| Specification must be improved | 19/02/2020 | Go over the text again and improve the quality. Add more references |  | To do |

# Appendices

N/A