Truth in Data Science

Gobi Govindasamy

Jan 4th 2024

School of Mathematics,
Cardiff University

A dissertation submitted in partial fulfilment of the
requirements for MSC DATA SCIENCE AND ANALYTICS
by taught programme, supervised by Alexander Balinsky

Executive Summary

Data Science has great potential to reveal meaningful insights for scientific and business modernization due to the exponential growth in the availability of data and analytical prospects. But reproducibility crisis acts as a key factor which brings in a flaw to this potential making it difficult to replicate the findings. This Project set out to analyse a real-world case study based on climate change modelling and to assess its reproducibility followed by validating the findings.

An extensive literature review was carried out in the early phase which examines the two different types of statistical approaches, i.e., Bayesian and Frequentist with the help of real-life examples. Following that an extensive study was conducted on processes like P-hacking, and overfitting which can lead to a higher chance of false positive results. These insights were used to investigate the case study in which the R Script provided for climate change is analysed for its legitimacy about the results.

In the given case study, L., McCarthy, G.D., Thornalley, D.J.R. et al(2021) modelled two Bayesian models with R2JAGS for Atlantic Meridional Overturning Circulation(AMOC) strength where one has a fixed changepoint of 50 years whereas the other with single estimated change point. The Bayesian models are first recreated in Python. This is then followed by creating two more non-linear models in the frequentist approach whose results are then cross examined to check the credibility of the project

This segment elaborates on the techniques and resources used to bring life to the approach using Python. Google Colab is a web-based development environment in which Jupyter Notebook service is hosted which combines the power of GPUs and TPUs making it a better cloud-based solution. This programming environment was conducive to the whole project, given the high hardware demand for the development and recreation of the statistical model followed by analysis. The models were primarily developed using Python3 and Python module PYMC followed by ARVIZ for creating models and visualising the results.

Two pronged approach was used considering the findings from both Bayesian and Frequentist models would be the same. The results from the prediction also proved

that findings of the parental research is true, thereby the dignity in the project is established.

This research paves the way for everyone in this field to understand the reasons behind the mistaken results in data science due to practices like P -hacking along with the challenges of replicability and reproducibility of a scientific work.

Acknowledgements :

My sincere gratitude goes out to Mr. Alexander Balinsky, my devoted academic supervisor, whose advice and assistance have been invaluable to me during this dissertation process. The critical analysis, constant support, and commitment to the advancement of my research were priceless. I owe a great deal for the guidance and mentorship.

The ground-breaking work of Caesar, L., McCarthy, G.D., Thornalley, D.J.R. et al (2021) in Atlantic Meridional Overturning Circulation(AMOC) is the foundation for this work. Their groundbreaking publications on Atlantic Meridional Overturning Circulation(AMOC) served as the methodological and theoretical cornerstone for my own research. While drawing from their work, my research uses Frequentist approach to critically evaluate the conclusions made under "Current Atlantic Meridional Overturning Circulation weakest in last millennium". In the field of data science, where decisions and insights are based on the truth, project credibility is critical. "Ensuring the credibility of a project protects against false conclusions and promotes accurate information" is the driving force behind our investigation. A careful assessment of their work, in my opinion, advances our knowledge as a whole and adds to the continuing scientific discussion in this area.

Without the constant encouragement, understanding, and listening ear from my friends and family-especially during tough times, this trip would not have been possible.

Table of Contents :

1. Introduction …….. 07
2. Literature Review…….. 09
3. Methodology and Approach……18
3.1Tools…….. 18
3.2 Data Collection …… 19
3.3 Environmental setup…… 19
3.4 Data Import……. 19
3.5 Initial data Exploration…… 20
3.5.1 Overview of data…….. .20
3.5.2 Data cleaning…… 20
3.5.3 Exploratory data analysis……. .20
4. Data preparation…… 23
5. Statistical Modelling……. 26
5.1 Recreation of Bayesian models……... 26
5.1.1 AMOC change in mean model(Single estimated change model) …… 26
5.1.2 AMOC change in mean model(Fixed change every 50 years)……. 29
5.2 Creation of frequentist models……. 33
5.2.1 AMOC change in mean model(Single estimated change model) …… 33
5.2.2 AMOC change in mean model(Fixed change every 50 years)……. 35
6. Conclusion…… 38
7. Future research……....39
8. References……..…… 40

List of Figures

Figure 1. Kernel Density Plot of Data	**21**
Figure 2. Heatmap showing corelation between Data	**21**
Figure 3. Line plot showing the trends over time	**22**
Figure 4. Plot showing the trends over time in posterior predictive checks	**28**
Figure 5. Plot showing the trends over time in posterior predictive checks over intervals of time	**32**
Figure 6. Plot showing the trends over time in prediction plots	**34**
Figure 7. Plot showing the trends over time in prediction plots	**37**

1.Introduction :

The evolution of big data and advanced techniques for analysis has given many exceptional ideas and discoveries. On the other hand, there are various examples where it is identified that data science can also generate deceiving, tendentious, entirely false conclusion. This project aims to find the origin of the false results in data science and to conduct an factual reproducibility analysis on the recent research based on the climate change.

With the exponential growth of data availability and intricacy of modern analytical methods, assuring research integrity and transparency is very important. However serious challenges are faced in data science in accordance with techniques like p-hacking, overfitting, unquestioned assumptions and biased publications which in turn can lead to the results which fails to reflect the truth. Some notable examples like the Google Flu Trends (Lazer et al.,2014) exposes how precise approaches can also generate a reckless insights about the data which are incorrect. This project will examine the causes which contribute to the findings which aren't reliable in the data science research.

This project includes a case study of a climate change research publication based on the Atlantic Meridional Overturning Circulation(AMOC) from Caesar et al. (2021). The research article claims that the recent trends of AMOC are unprecedented over the millennium. Whereas other researchers have questioned the findings in subject to concerns about the calibration of proxy data, seasonal bias, and circular reasoning. Recreating the analysis gives a factual evaluation of the previous claims. As a significant process for heat redistribution on Earth, the Atlantic Meridional Overturning Circulation (AMOC) plays a crucial role in both climatic fluctuation and change. The AMOC is a delicate nonlinear system that depends on minute variations in the ocean's thermohaline density. Significant AMOC shifts have been connected, for instance, to millennial-scale climate events that occurred during the last glacial epoch.

To summarize, this project will be providing a challenging stance on the reliability on data science. The principles of robust analysis through unbiased reproduction on recent research on climate change will also be demonstrated. In order to ensure that data science lives up to its credibility to reveal the truth and expansion of knowledge

depends upon focusing on clarity, understanding uncertainty and consideration of assumptions. With great availability of data comes immense responsibility. This project aims to highlight both pitfalls and the best practices to progress data science as a more trustworthy one that serves this society.

2.Literature Review :

Our research commences with the elaborate study of errors happening in the field of data analysis followed by the deep dive into the reasons behind the errors. Some of the common errors happening in the analysis of data are P-hacking, Data Dredging, usage of improper statistical tests, ignoring biases and overfitting of models. The preceding errors are clearly explained in the forthcoming context.

The first work which is considered is the parental research based on which this whole project of mine revolves. This paper looks at a variety of proxy records in order to reconstruct the history of the Atlantic Meridional Overturning Circulation (AMOC) over the past 1400 years,. It finds evidence of a relatively stable AMOC up until the 19th century, when it started to weaken. This was followed by a more rapid decline starting in the 1960s, which resulted in an unprecedented modern AMOC weakening over the last millennium that statistical analysis confirms is significantly lower than earlier centuries. Although uncertainties persist due to proxy limitations and short timescales, the results robustly suggest that ongoing anthropogenic climate change may be severely disrupting a major ocean circulation system, with crucial implications for past and future climate change. This highlights the urgent need for additional research on the mechanisms causing AMOC decline to better understand its climatic effects.

This work by Hand(1983) draws attention to the widespread yet deceptive practice of classifying rules based just on their perceived mistake rate. By separating out the data that was used to construct the model from the data that was not, the apparent error rate provides an optimistic assessment of performance. The author shows how the apparent error rate constantly and occasionally significantly underestimates the true misclassification rate through analysis of real-world situations. A false feeling of a method's accuracy is provided by apparent error rates, notwithstanding their widespread use. The author contends that more accurate methods, such as data partitioning, leave-one-out cross-validation, and bootstrapping, are available to anticipate future performance. These methods are more computationally demanding, but they more accurately capture real-world scenarios where models are applied on fresh data. This study by Brown et al.(2018) analyses invalidating faults in study design, data collecting, statistical analysis, logic, reporting, and communication. It also

looks at common errors and underlying causes that compromise scientific rigour. The authors illustrate poor data, analysis errors, reasoning errors, poor communication, and contributing factors such as ignorance, flawed study inception, publication pressures, excitement overriding objectivity, insufficient resources, and conflicting priorities through examples and surveys demonstrating nontrivial prevalence across disciplines. The issues of overburdening researchers are noted, and suggested solutions focus on education, gatekeeping requirements, incentives, resources, and cultural shifts to promote intrinsic desire for rigour. The authors contend that although there are obstacles to identifying and fixing mistakes, coordinated efforts by all parties involved can gradually improve scientific self-correction and knowledge advancements. Nevertheless, in the human endeavour of science, perfection will remain unattainable; But perfection will always be difficult in human endeavours like science; maintaining rigour while advancing will call for constant process improvement, personal honesty, and teamwork. The study emphasises the significance of using appropriate statistical approaches to objectively evaluate classification systems and steer clear of overly optimistic claims of performance by demonstrating the considerable bias in apparent error rates. By using these vitally crucial tools for diagnostic and decision-making, this helps to assure robust methodology and findings.

In method comparison studies, this study by Pankaj K. Choudhary & Kunshan Yin (2010) investigates frequentist and Bayesian methods for numerous comparisons and method pair ranking using agreement metrics. For establishing tight simultaneous coverage bounds, the authors find that noninformative priors and bootstrap approaches outperform typical frequentist methods. Although it necessitates more implementation concerns, the Bayesian approach offers additional benefits such as posterior probability inferences and the incorporation of useful priors. For sample sizes of thirty or more, the methods perform well with normality assumptions for the dependent variables. While the report offers a helpful comparison of methods for measuring and rating agreement between them, it ignores study design issues pertaining to necessary sample sizes. It is still necessary to conduct additional research on nonparametric modelling techniques and study planning.

The article by Min-ge Xie and Kesar Singh(2013) give a thorough analysis of current developments in confidence distribution theory and technique is given here. Many forms of distributional inference, such as bootstrap distributions, p-value functions,

fiducial distributions, and normalised likelihoods, are unified by the current notion of a confidence distribution. The debate centres on frequentist concepts and applications, where confidence distributions provide access to statistical techniques that were not before available. The philosophical argument between frequentist and Bayesian schools is not reopened, although pointing out certain distinctions and parallels with Bayesian processes. Rather, confidence distributions serve to bridge and close gaps between disparate statistical processes and philosophies by offering a single, comprehensive framework that encompasses a wide range of statistical inference techniques.

Finally, the study by Borra and Di Ciaccio(2010) provides a comprehensive evaluation of approaches, such as cross-validation, bootstrap, and covariance penalty techniques, for measuring the prediction error of non-linear regression models. The authors indicate that the repeated-corrected 10-fold cross-validation and parametric bootstrap produce the most stable and accurate estimates of extra-sample prediction error across conditions across simulations with varied sample size, signal-to-noise ratio, and model stability. The findings demonstrate how overfitting has a significant impact on estimator performance, with parametric bootstrap and corrected cross-validation showing the most resilience. For applied researchers who must consistently validate and compare non-linear predictive models, the publication offers useful information.

Followed by the study of common errors happening in the analysis of data, an extensive study was caried about the two different statistical approaches i.e., Bayesian approach and Frequentist approach with the help of simple real-life examples which makes the understanding of the concepts more easier and faster to everyone followed by a comparative study between the approaches in order to explore the difference between them.

Bayesian hypothesis and modelling :

In the field of statistics and probability, Bayesian offers a remarkable perspective when it comes to analysing data and hypotheses testing. It's different from the traditional approach, "Frequentist", where population parameters and long-run averages are focused upon whereas the Bayesian approach embraces uncertainty and the beliefs are updated that are based on the information before-hand. These qualities makes

Bayesian valuable when it comes to decision making and scientific inquiries which includes prior knowledge and data's continuous flow acts as a key

The fundamental elements of Bayesian hypothesis and modelling comes together as follows :

1.Priors and Likelihoods :

- Prior : The existing belief or knowledge about the hypothesis or the parameter which is being investigated is encapsulated into this, which in turn is expressed as a probability distribution reflecting the degree of uncertainty of the true values.
- Likelihood : the probability of observing data which includes a given value of parameter or the hypotheses validity. In this the fit of data with our beliefs is very important.

2. Bayes' Rule: when the prior and likelihood are combined using the bayes rule the magic happens. Bayes rule is a mathematical formula which updates your initial beliefs in the presence of observed data. Posterior distribution is the outcome of this which represents the refined and quantified beliefs on the parameter or hypothesis later

3.From prospects to conclusions:

The foundation for the further analysis and decision-making is the posterior distribution. The Credible intervals are calculated which are used to quantify the likely values' range, the competing hypothesis can be compared with the help of bayes factors and predictions are made based on the updated system of beliefs.

Advantages of Bayesian approach:

- Prior knowledge incorporation : Integration of opinions and existing facts can be done in a smoother way, which leads to results that more sensible and with refinement.
- Constant updating: when there is emerge of new data, the beliefs can be quickly adjusted, which offers a framework that is dynamic and flexible to changes for exploring and prediction making.

- Measuring uncertainty : The confidence associated with conclusions are understood clearly with the help of probabilistic output, thereby elevating clear and well informed making of decision.

Drawbacks of Bayesian approach:

- Selecting informational Priors : The selection of priors which will be appropriate will be important for obtaining results that are reliable and worthy. Skewness in analysis can be resulted from biases in the prior.
- Computation difficulty : Specific Bayesian models may be requiring specialized software and skilfulness which makes it computationally demanding.
- Results interpretation : The communication and translation of results is necessary though the language of probability may not be familiar with all audience.

Applications of Bayesian approach:

Bayesian methods have a wide range of applications in various fields which includes:

- Engineering and science : Estimation of physical constants, predictive models development and analysing experimental data.
- Economics and Social science : Testing the effectiveness of interventions, examining economic trends and analysing voter behaviour
- Artificial intelligence and Machine Learning : Creating and training of probabilistic models, improving the accuracy of prediction and incorporation of user feedback.

Bayesian hypothesis and modelling offers a powerful foundation for scientific inquiry purposes, decision-making across wide range of domains and data analysis with the help of continuous updating of beliefs through data and uncertainty.

Frequentist Hypothesis and Modelling

Like Bayesian method relies on the facts and beliefs updated with each new data, the Frequentist method takes a different viewpoint. The foundation of this approach lies on the concept of repetitive sampling and frequencies of long run. It inquires: “if this

particular experiment can be replicated numerous times, how often the outcomes could be seen like this? ”

Here comes the unfold of frequentist methods and modelling :

1.Hypothesis testing:

- Null Hypothesis : This hypothesis states that there is no presence of significant relationship between the variables from data which are considered. This is represented as H0.
- Alternative Hypothesis : this hypothesis states the possible alternative of H0, that is there is significant relationship between the variables which are under consideration. This is represented as Ha.

2. Statistical Tests and p-value:

- T-tests, Chi-square tests and Regression analyses are some of the tests employed in frequentist approach in order to quantify the proof against the null hypothesis.
- The p-value act as the main output here which denotes the probability of seeing data as intense as what was observed in which the null hypothesis is assumed true.

3.Denying to turn down H0 :

- P-value below 0.5 is treated as a low one, which gives enough proof to deny the null hypothesis and decide that the detected effect is not likely to be due to chance.
- In contrast, a high value of p-value does not conclude that the null hypothesis is accurate. Whereas it shows that there is insufficient evidence to reject it, which is often referred to as “failing to reject null hypothesis”.

4. Estimation of Confidence Intervals :

- Confidence intervals are provided by the frequentist approach around estimated parameters like mean, coefficients etc, in a model. This confidence interval demonstrates the value range within which the real parameters are likely to lie.

Advantages of the frequentist Approach :

- Universally acknowledged and inferred : these frequentist approaches and methods are more reliable and acts as a backbone of statistics effecting their results interpretable in a easy way and more comparable in various studies.
- Explicit guidelines for making decisions : The frame work of "reject or failing to reject" offers an approach which is straightforward when it comes to drawing conclusions depending on the available p-values and confidence interval.
- Sturdy with outliers : these frequentist approaches are less susceptible to data points that are extreme compared to Bayesian approach.

Disadvantages of the frequentist Approach :

- Concentrating on testing for null hypothesis: The attention on denying or failing to deny the null hypothesis may overshadow examining the effect size and practical significance.
- Two-fold analysis of p-values : Interpreting the p-values wrongly acts as conclusive evidence or concealment of a hypothesis which can lead to incorrect results.
- Restricted handling of uncertainty : Frequentist approaches doesn't incorporate the prior knowledge explicitly or update the belief in a constant manner thereby it potentially overlooks the valuable piece of data

An Essential component of scientific analysis :

Regardless of the constraints and complications, the frequentist methods acts as the primary support for scientific exploration and analysis of data. the reasons which makes frequentist approach, a valuable and reliable tool in order to understand real life aspects include testing of hypothesis, transparent criteria for decision making and its wide range of application.

Therefore having a clear cut idea of both the statistical approaches of Bayesian and Frequentist enables us to determine the most accurate approach based on some factor like the research question, data available and required degree of uncertainty.

Here an event is described using both approaches and it's outcome is discussed for the better understanding of both approaches. The event which is considered here is tossing coin and it's outcomes are studied here.

Coin toss : Frequentist vs. Bayesian :

While we are investigating the fairness of coin, our curiosity will be in knowing whether the coin lands on heads and tails with equal probability(50/50) or if the coin is biased towards one side(head/tail).

Here comes the Frequentist approach:

1. Setting the Null Hypothesis(H0) : The coin which is taken is fair(50% heads and 50% tails)
2. Setting the Alternative Hypothesis(HA) : The coin is biased thereby it favours either head or tails.
3. Experiment : Flip the coin for 'n' time and record the results. Here let n be 100 times which gives 70 heads and 30 tails.
4. Statistical test : Perform Chi-square test in order to access the likelihood of observing 70 heads or more extreme deviations from 50/50 where H0 is assumed to be true.
5. p-value: check the p-value. If p-value is less than 0.05, the null hypothesis is rejected. This suggests the observed from 50/50 is unexpected and confirms the presence of biasness in the coin.
6. Confidence Interval : calculate the confidence interval around the proportion of heads(i.e., 70 heads 70%). This range explains where the true probability of heads lies in the population of coin flips.

Bayesian Approach :

1. Prior : we commence with the incorporation of prior belief about the fairness of coin. This can be set to 50/50 if there is no presence of prior information on the fairness of coin or it can be affected by the factors like appearance of the coin or earlier observations.
2. Likelihood : let's consider the possibility of observing 70 heads(30 tails) for diverse range of true probabilities of heads(range is between 0% and 100%).

3. Bayes' Rule: By the application of Bayes' rule, we combine both prior and likelihood in order to update the belief about the coin's fairness based on the observed data.
4. Posterior Distribution : Our new updated belief about the original probability of getting heads is reflected by this new distribution. After the incorporation of evidence , there might be a slight change in probability(for example it may come down to 63%).

Comparison between approaches :

- Frequentist approach focuses on rejecting or failing to reject the null hypothesis(H0) whereas the Bayesian approach updates beliefs continuously based on the data.
- Frequentist depends on p-values and confidence intervals, whereas Bayesian provides on posterior distribution for the parameter of interest.
- When prior knowledge is minimal frequentist works best and Bayesian updates the prior information by updating it with existing information.

So the final thought is both frequentist and Bayesian approaches provide insights that valuable . But the assumptions and goals makes one to choose between the appropriate approach. In this, if the prior belief on fairness of coin is well-known Bayesian approach may be better choice than the frequentist approach. However, if there is minimal prior knowledge about fairness of coin and if the focus is on the testing of hypothesis, then Frequentist approach will be better one for consideration. So it's only about knowing there strength and limitations thereby choosing the approach which act aligned with the objective of the research .

In conclusion, this comprehensive literature survey sheds light on this diverse area of expertise, providing valuable insights which may acts as a preparation for the further exploration along with advancement in this field.

3.Methodology and Approach :

Based on the objective, the work is segmented into sub-tasks incorporating transcoding, data preparation, statistical modelling, and cross-validation. The mentioned tasks are elaborated in the upcoming sections. Now, here we commence with transcoding the available R script from the research done the authors. This was done with the help the of python .we'll dive deeper into the process of transcoding and newer statistical models in the following sections. The terminal part will be discussing about the cross validation of results thereby establishing the legitimacy of the project previously carried out by the authors on the findings about "Current Atlantic Meridional Overturning Circulation weakest in the last millennium".

3.1 Tools :

Python is one of the languages which balances versatility ,flexibility, readability and scalability followed by its universal usage makes it a more powerful programming language for wide extent. Due to these factors, Python is considered to be informative regarding any problems one may come across. Like this, Python has extensive libraries and modules for the creation, recreation and accessing the reliability of the statistical models. One such feature-packed library in Python is PyMC3, which is used for probabilistic programming. PyMC3 enables the users to express and fit Bayesian statistical model for the data analysis in a efficient manner. Though the authors has created the Bayesian models using JAGS in R while recreating the same model in python PyMC3 is used instead of Pyjags (Python module which provides an interface JAGS environment for creating Bayesian models) due to various reasons in which some are listed below,

- PyMC3 permits the user to diagnose and criticize the Bayesian models within python without connecting to external JAGS code.
- Probabilistic programs are customizable when it comes to PyMC3 which can be further used in other libraries of Python ecosystem.
- The readability and expressiveness of the code written in PyMC is more when it is compared to that of code written with the JAGS.
- PyMC uses efficient samplers which provides better performance and scaling for big data.

3.2 Data Collection :

The dataset provided, which was earlier collected by the authors that acted as the foundation of this research. The dataset used by the authors in order to explore about Atlantic Meridional Overturning Circulation (AMOC) is the one which reconstructed with the help of several other datasets used in other research projects. There are nearly a total of 11 datasets used for the creation of proxy dataset which will be used here. The earliest of the records date back to around 400 AD and the most recent ones are from the early 20th century which covers almost span of around 1600 years. These datasets are used to create a dataset for the span of 146 years starting from 1871 till 2016. Statistical change point analysis was individually applied to each series in order to test for significant differences which in turn may indicate the changes in AMOC. The mean values for non-coinciding interval of 50 years which goes back in time were tested for every proxy for deciding whether there were lower values in the latest period when compared with the past. The regions from which were the proxy datasets are created is around the North Atlantic region.

3.3 Environment Setup :

It's necessary to note that the configuration needed may change depending on the software stack, architecture of hardware and the use case specified by the user. Here the code works in a environment which configures the environment variables for libraries that are working on numerical computation, which specifically affects the behaviour of Theano Library Intel Math Kernel Library(MKL).This setting is necessary for compatibility purposes and avoiding the conflicts with other libraries. This may in turn help in optimizing the performance of numerical computations by configuring MKL and BLAS based on the architecture of system. Issues and conflicts which arise during code execution will be addressed with these configurations

3.4 Data Import :

Pandas is a Python package which is used data manipulation and analysis. Pandas reads the given data into Data Frames from different sources which will be available

in a variety of formats. Data Frames are structured from of data in rows and columns with the labels. This is one the vital facility provided by Pandas.

Read_csv() is the primary method for are different methods to load csv files and it is the same method implied in here to import the dataset.

3.5 Initial Data Exploration:

3.5.1 Overview of Data:

Here first few rows of the data were explored using the head() function. It is necessary for knowing general features of data like its feature names, data types and sample values. Following this, shape() function is implied to evaluate the dimensions of data since gaining knowledge on the size of dataset helps in determining the computational and processing methods in future aspects. It also reveals the number of features available for analysis.

3.5.2 Data Cleaning :

Handling missing values : Using the isnull() function, the presence of missing values is examined and to have amore detailed view of which columns have missing values sum() method is called upon this Boolean dataframe which in return the number of missing values in a column wise manner. Since the data present is very simple and has no missing values no cleaning of data is needed

3.5.3 Exploratory Data Analysis :

KDE plot and heatmap : Probability density function is studied with the help of KDE plot. Here, a two dimensional Kernel Density Plot is created for the “y”, “lower” and “upper” columns of the dataset. The plots show that the variables follow similar type of distribution and have strong corelation which is also confirmed with the heatmap showing their strong corelation.

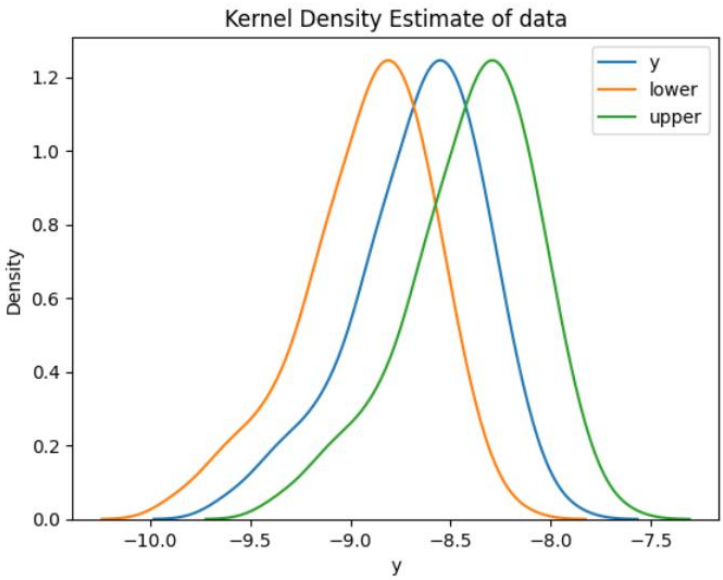

Figure 1. Kernel Density Plot of Data

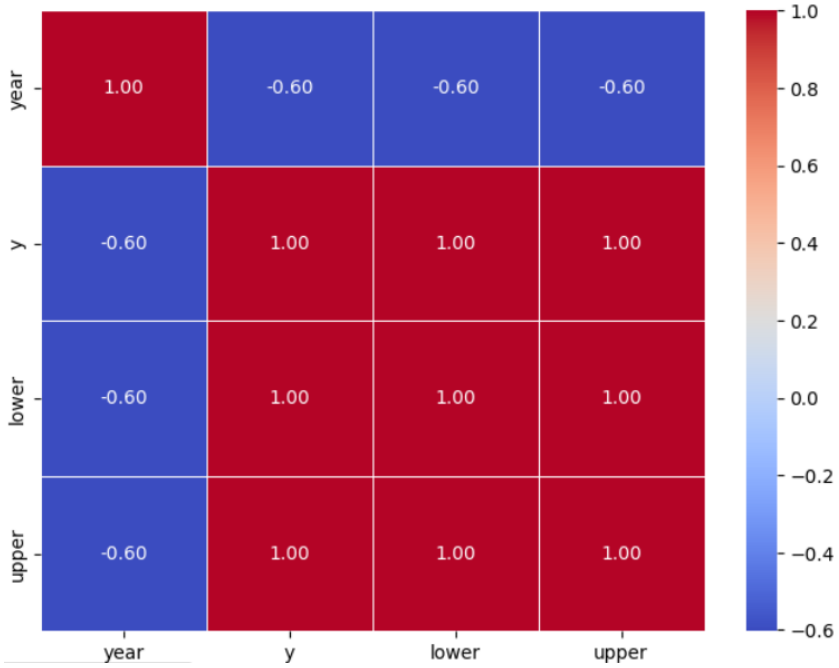

Figure 2. Heatmap showing corelation between Data

Line plot : Trend and uncertainty of variables in the data is established using line plot between the columns “y”, “lower” and “ upper” over “year” , where “lower” and “ upper ” fills the line formed by “y” over year. This give a range in which the true values tend to fall , thereby giving a sense of precision of the estimates.

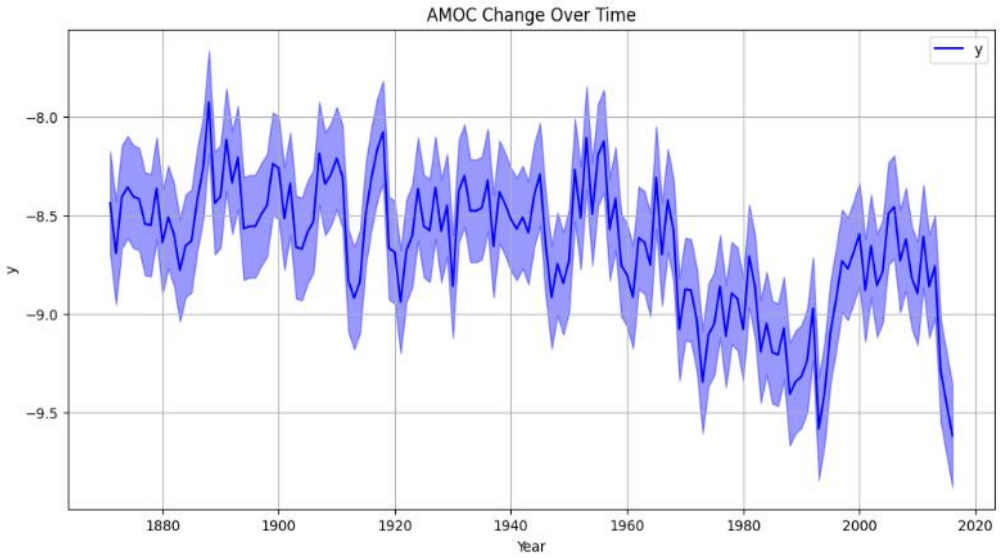

Figure 3. Line plot showing the trends over time

The Atlantic Meridional Overturning Circulation (AMOC) is plotted to show how it changes over time. The graphic indicates that since the 1880s, the AMOC has been waning. There have been times of very rapid decline, such as the early and late 20th centuries, but the pattern of weakening is not linear.

4. Data Preparation :

It is one of the important step when it comes to rewriting an R script in python followed by applying statistical modelling like frequentist modelling and Bayesian modelling. So here the data is changed in accordance with the R script from the research work carried out by the authors earlier in parental research.

In this the data is prepared in accordance with the type of statistical model earlier applied in the parental research for the process of reproducing the r script and also according to the requirements for the construction of new statistical models which are based on the frequentist approach. The different ways in which the data is prepared is listed below :

1.Vectorization :

This is performed on entire arrays of data simultaneously without the need of iterating through individual elements using loops which enhances the speed of operations followed by increasing its conciseness and compatibility. Vectorization is more than just code elegance and performance; it's a complex jewel in data preparation. It optimises memory utilisation for huge datasets, coordinates computing performance through matrix operations, and streamlines feature engineering activities like dimensionality reduction. Its capacity to process data in parallel lets current processors perform computational symphonies and guarantees numerical stability for algorithms that have certain numerical preferences. Text data is also included in this dance, being converted into vectors for tasks related to topic modelling and sentiment analysis in natural language processing. But with great power comes great responsibility; data type knowledge, memory optimisation, and algorithm alignment are all important factors to take into account. In the end, vectorization makes data scientists into masters of quick, elegant, and algorithmic analysis by enabling them to move through the information landscape with agility and precision.

2.Scaling:

Scaling becomes an important tool in data analysis, tying distant feature sets together. Similar to how a conductor adjusts the pitch of an orchestra, scaling makes ensuring that every feature has a distinct melody and keeps features with higher values—that is, louder instruments—from overpowering their more subdued counterparts. Scaling

promotes algorithm fairness and interpretability by converting information into a shared numerical range, producing accurate and nuanced insights. Although there are many different scaling strategies, each with advantages and disadvantages of their own, careful application and selection guarantees that every feature adds its own song to the data symphony, enhancing the analysis in the end. Numerical features are scaled in certain algorithms in order to achieve the same result as parental research. Larger magnitude characteristics may unintentionally skew the model's learning in favour of their influence. By reducing this bias, scaling encourages more accurate predictions and a more equitable representation of all features. With the help of this easy technique, models may train efficiently, make precise predictions, and uncover hidden patterns in your data.

3.One-hot encoding :

One key method that turns inert category characteristics into binary melodies that resonate is one-hot encoding. One-hot encoding generates a single binary feature for every different category, allowing for their smooth integration with their numerical counterparts, much like a conductor creating separate sections for each instrument. In doing so, it enables algorithms that are trained on numerical dances to recognise the subtleties of each category's contribution, promoting interpretability and feature independence. The capacity of one-hot encoding to reveal the latent harmonies in categorical features enhances the depth and subtlety of the data symphony, guaranteeing that all voices are heard in the end, even though dimensionality expansion and sparsity are important factors to take into account.With the help of this magic trick, algorithms can comprehend distinctions between categories, interpret them with ease, and even give them varying weights. It can add more features, but it's a really useful tool for finding the hidden mysteries in your categorical data.

4.Feature engineering :

In order to produce new characteristics that could be more instructive for the model, it entails modifying the current data. The model can capture possible non-linear interactions between time and the dependant variable by squaring the year variable. When linear patterns are insufficient, this is important. It broadens the feature space, which might improve the model's capacity to recognise intricate patterns. Feature creation is the process of creating more features out of preexisting ones, like carving

more instruments out of raw material. Imagine if "average speed" is extracted from "distance" and "time," adding a new melodic line to the analysis.

Feature transformation is the process of reshaping already-existing features to improve their algorithmic suitability—think of it as tuning and polishing the instruments. Among these polishing, methods are scaling, normalisation, and discretization, which guarantee that every instrument plays in the same key and range.

Choosing the most pertinent and instructive elements is known as feature selection; it's similar to picking the appropriate instruments for the composition you want. Eliminating superfluous or cacophonous elements simplifies the ensemble and concentrates the examination on the most memorable tunes.

5. Intercept inclusion :

The intercept appears as a subtle component, but its presence or absence can have a subtle effect on the composition as a whole. Its inclusion guarantees that all aspects resonate in their intended range, similar to how pitch is adjusted in an orchestra, resulting in precise and comprehensible insights. When all independent variables are zero, the intercept shows the dependent variable's baseline value. Think of it as the analysis's primary key signature, from which the feature contributions' harmonies rise and fall. By ensuring that models do not force the "zero-feature" forecast to lie on an artificial origin, the inclusion of the intercept promotes more realistic interpretations.

A better understanding is provided by two separate scenarios:

Included Intercept: The model forecasts the value of the dependent variable even in the absence of any characteristics. This is especially important when a feature's zero value does not always indicate that its influence is absent. Think about forecasting the price of real estate; a zero "number of bedrooms" does not mean that there is no influence from the home style; an intercept takes this initial cost of ownership into account.

Intercept Excluded: The model might have misinterpreted feature contributions if it had assumed that the zero-feature scenario would line up with the origin. Although it needs to be interpreted carefully, excluding the intercept might be appropriate in some situations where a true "zero-feature" state exists.

5. Statistical Modelling :

In this research, four statistical models are created based on approaches from Inferential statistics(Bayesian and Frequentist). The Bayesian model created here are recreation of the Bayesian model which was earlier created by the authors in R script. So the concept of Bayesian model is directly applied as in accordance with the author's approach.

5.1 Recreation of Bayesian Models :

The Bayesian model created by the authors was created in R using JAGS() . this consists of two Bayesian models which are created for finding AMOC change in mean model using a fixed change over every 50 years and single estimated change point.

5.1.1 AMOC change in mean model(Single estimated change model) :

This model aims to detect and analyse changes in the dataset, which is accomplished by considering one changepoint for the whole of data. Based on this shift, data are allocated to "before" and "after" shift ,assuming each area has its own average value. With the help of analysing uncertainties, the happening of shift and the differences in averages are explored. Before modelling, data is prepared in such a way, it is suitable for modelling. The Steps include,

Scaling:

In this data, the column "year" is scaled by dividing the values by 1000 . This particular rescaling of values by the factor of 0.001 is carried out because the same convention is followed in the parental research. This ensures the consistency during the analyses.

Calculation of variability :

The uncertainty or variability associated with "y" is calculated with the help of "upper" and "lower" values from the data, likely revealing a range or confidence interval for every data point.

Followed by this, the total timespan is calculated with the help of minimum and maximum year calculations.

Model Building :

The model estimates the changepoint present within the datapoint thereby allocating the data “before” and “after” changepoint.

Model definition :

The building of model starts with setting Priors, Error and likelihood.

Priors:

The mean of each interval is assumed to follow normal distribution with standard deviation of 10 and mean of 0 . The variable “alpha” is assigned to for storing this value. The value of error in standard deviation is constrained and it made to remain positive. This is done using Half-Cauchy distribution, which is assigned to variable “sigma_err”. The change point location is assumed to be distributed in a uniform way between the range of minimum and maximum years.

Error :

The incorporation of both error and uncertainty in observation is calculated into tau, which signifies the precision also the inverse of variance.

Likelihood :

Each observation is assigned to an interval depending on whether the year is before or after the changepoint in data. The observation of model are distributed uniformly which rely on means based on the intervals assigned (mu) and precision(tau).

Sampling :

The samples are stimulated using Markov Chain Monte Carlo(MCMC) posterior distributions under the consideration of both priors and data. The posterior distribution is estimated for each parameter with the help of sample analysis which is stored in “trace”

Learning from model :

- The changepoint where the sudden change occurs is being identified therefore the data is classified into before and after change considering all the uncertainties thereby revealing its reliability. The changepoint of the model is identified to be around the year 1968(95% credible interval between 1967-1970), which suggests a shift in mean of y within that specific range of time.
- The estimated mean helps in finding the presence of change point. the mean before changepoint is -8.499(-8.545 to -8.454) and the mean after the change point is -8.986(-9.052 to -8.923).
- The non overlapping credible intervals for before and after the changepoint signifies the presence of statistically significant differences in the mean of y after the occurrence of changepoint.
- The posterior predictive checks displays the valid agreement between observed vs predicted data, which thereby conveys that the model adequately captures the detected pattern in the time series. The posterior prediction plot is given below,

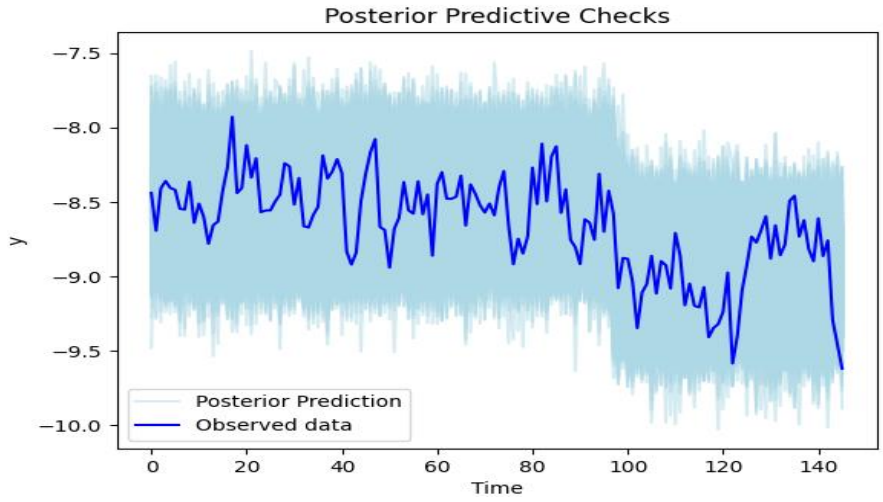

Figure 4. Plot showing the trends over time in posterior predictive checks

5.1.2 AMOC change in mean model(Fixed change every 50 years) :

This model aims to detect and analyse changes in the dataset, which is accomplished by dividing the data into intervals(50 years) thereby assuming each interval has it own mean value and error magnitude in the data. Then it infers the locations of changepoints with the help of intervals identified with notable different means.

In order to achieve this the data is prepared in a specific way like,

Identification of Year Range :

- The maximum and minimum value of year is found from “year” column of the dataset.
- The “year” values are extracted into a separate vector.

Establishing changepoints :

- Vectors are created in order to store the possible changepoints which are spaced in the interval of 50 years apart between the maximum and minimum values.
- Ensure that the first changepoint is at the minimum year or within 10 years of it.
- Now the number of change points is calculated.

Assigning observation to intervals:

- Vectors are created to store the interval assignments and names.
- Iterate through the changepoints and observations: which assigns each observations to an interval based on year value and the boundaries of changepoint. This updates the created vector with interval assignments.

Final adjustments :

- The last observation is assigned to the last interval and the first interval is named based on its boundaries.

So the key data transformations include,

- Assignment of each individual observation to an interval based on the changepoints that are predefined.
- Mapping is created between observations and their corresponding intervals.

- Descriptive interval names are generated.

The purpose of doing this is to prepare data for the particular changepoint analysis by grouping observations into intervals with characteristics that are potentially distinct each other. this enhances modelling and inference by allowing different parameters for model of each interval .Following this the changepoint model is constructed which helps in changes over intervals of time (50 years).

Model Building :

This model estimates the mean value and uncertainty present within each interval of 50 years as specified before. Here comes the breakdown of components of model for a better understanding of model;

Adjustment of inputs:

The interval to which each data belongs are held in an array.

For ensuring that the data is valid for the model, np.clip() function is used for adjusting it. This function makes sure that the adjusted value falls within the permissible range of 0 to 1 less than total number of changes in intervals as there is only n interval and the index of 0 represent the first interval and n-1 means the last .

Model definition :

The probabilistic relationships of the model is defined with pm.Model().

Priors :

The variable “interval_mean” is created assuming each interval has its own mean value, and these value follow normal distribution with the mean of 0 and standard deviation of 1. This acts as a starting point. Following this, the model estimates the actual mean values for every interval based on the data points.

The standard deviation of error in the observations are represented by “sigma_err”. The value of “sigma_err ” is always positive which is ensured by “pm.HalfNormal”.

The error and observation of uncertainty are combined into single measure precision, inverse of variance. This is represented by “tau”.

Likelihood :

The model gets connected with the real data in this point. “y_obs” represents the observed data points. We assume each data point follows normal distribution with mean equal to the mean of interval and a precision of “tau”. This signifies that the estimated mean of interval is closer to the data points present within the same interval.

The observed data points that are used in the model comes from the “y” column of the data.

Sampling :

This is followed by the sampling by drawing thousands of samples from the model using Markov Chain Monte Carlo (MCMC). Different possible combinations of parameter values which fit the data well are allowed to explore.

The posterior distribution is estimated for each parameter with the help of sample analysis which is stored in “trace”. The most likely value of each parameter and the range of plausible values are estimated from this distribution, thereby providing the measure of uncertainty present in our estimates.

Learning from Model :

- The intervals with significantly different means can be identified with the help of posterior distributions, which suggests the changepoints in the data. The uncertainty in the estimates of mean for each interval can be studied from the width of posterior distributions.
- Estimates of mean : For each interval of time, the model has estimated distinct mean values which has brought a downward trend into light. The mean estimates are,

  Interval 1(earliest interval) : -8.456

  Interval 2 : -8.512

  Interval 3(latest interval) : -8.953

- the clear separation of third interval’s credible interval and the earlier intervals indicts the presence of apparent shift in the mean after the changepoint.

- The mean estimates and uncertainty are plotted using violin plot for better visualisation and understanding purposes.

  The plot is given below,

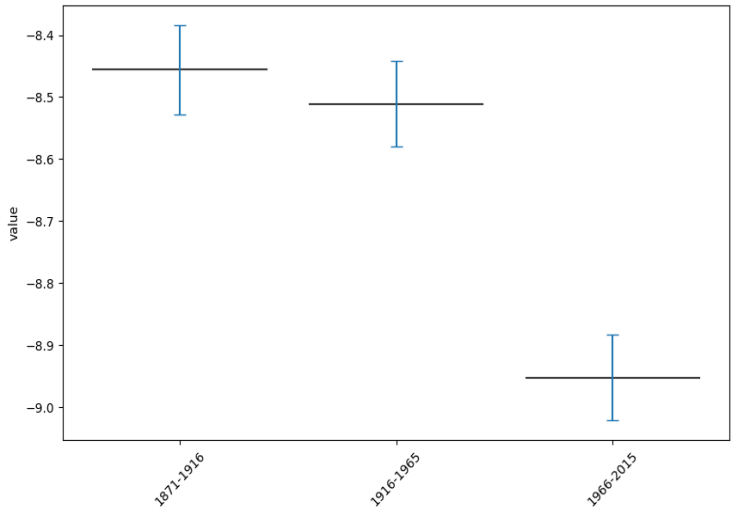

Figure 5. Plot showing the trends over time in posterior predictive checks over intervals of time.

5.2 Creation of Frequentist Models :

In order to establish the credibility of results obtained and the approach used in the parental research, the same models are newly created using a different approach, frequentist approach. In this approach, the data is prepared first, followed by its application in the statistical model constructed and the results are cross validated.

5.2.1 AMOC change in mean model(Single estimated change model) :

Here the quadratic model aims to explore the potential non-linearities thereby establishing the relationship between time and AMOC temperature and explore the changepoint present in.it .The data preparation is as follows,

Squared term creation : in order to capture potential relationships, a new column “year_squared” is created with the square values of existing “year” column. This enables the model to detect the curvature in the relationship between variables.

Model Building :

Followed by the preparation of data, the data is employed in a quadratic ordinary least squares(OLS) regression model which estimates the best fitting curve establishing the relationship between the temperature and time.

Model definition :

The model include the following variables,

Year_squared : it defines the curvature in the relationship and represents the quadratic effect of time. This act as the independent variable.

y : the outcome variable which is being modelled is represented by “y”. its behaviour is investigated in relation to time. This represents the AMOC temperature values.

Constant term : “const” is added to the model using the command “sm.add_constant(X)”, which accounts for the baseline effects that are not captured by the time-based variables. This ensures more accurate representation of the relationship.

Regression technique: Ordinary Least Squares(OLS) is the technique implied here to find the relationship between the independent and dependent variables by fitting a parabola in it.

Learning from model:

- The key finding is the presence of negative co efficient for “year_squared” which is statistically significant. This gives an idea that the relationship between is not simply linear but also shows a downward curvature. This exhibits the rate of decline in temperature of AMOC in the upcoming years.
- The negative sign of “year_squared” coefficient shows the presence of concave parabola which is downward.
- The R-squared value indicates the improvement in simple linear model. This indicates that there is variance in AMOC temperatures which are unexplained. Here there is a presence of 61% of variance since R-squared value is 0.388.It also brings to light that there may be other factors influence on the decline of AMOC apart from time.
- The plot of observed vs predicted capture the overall trend of data but some deviations highlights the limitation of model in capturing the full complexity of AMOC fluctuations.

Model Forecast :

The plot present below is the prediction plot of future values of the variable which is trained already in the statistical model above. The prediction chart is,

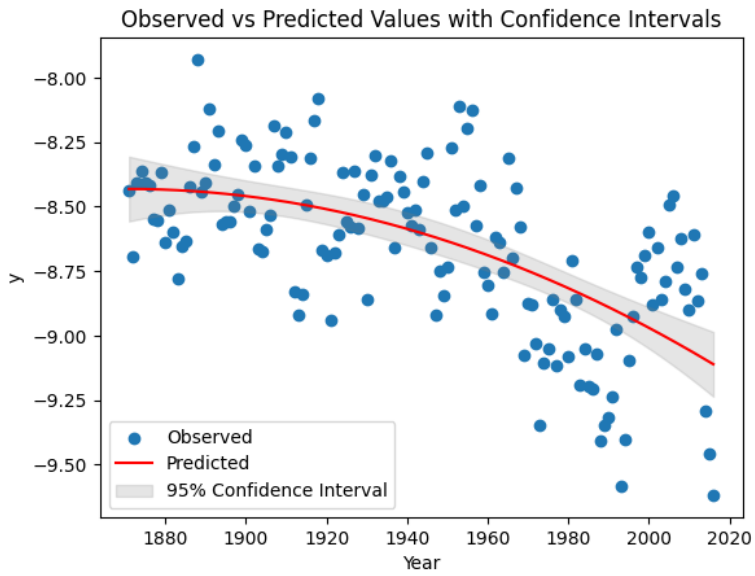

Figure 6. Plot showing the trends over time in prediction plots

5.2.2 AMOC change in mean model(Fixed change every 50 years) :

Using a method known as linear regression, this code examines the potential effects of numerous factors (independent variables) on a significant response (dependent variable). Here “y” acts as the dependent variable. The data is divided into intervals of 50 years. Then, using intervals found with noticeably differing means, it infers the locations of changepoints.

Here the data is prepared in such a way that enables to infer the results in a correct way. The data preparation includes,

Defining intervals :

Intervals of 50 years are established. this is done by calculating the maximum and minimum value of years followed by creating a list of changepoints at regular intervals of 50-year to define distinct time periods. This is done as same as similar to preparation of data for the Bayesian model.

Data categorization :

Each observation is assigned to a specific period based on the corresponding year, thereby creating a new column “Period” in the dataset. For performing this, “np.digitize” is used .

One-Hot Encoding :

The newly created column “period” is categorical in nature. This is converted into multiple binary columns using pd.get_dummies, thereby representing each period as a separate feature.

In order to reduce redundancy and prevent multicollinearity issues in subsequent analysis, one of the newly created columns is dropped.

These data preparation techniques prepares the data in such a way by encoding temporal information in such way that machine learning models can handle it effectively. The representation of periods as separate feature enables the model to identify pattern and relationships specific to different timespans.

Model Building :

Here a linear regression model is constructed which looks into possible variations in AMOC temperatures over various time periods.

Model definition :

The model includes the following parameters,

Dependent variable : Here “y” column from data act as the dependent variable

Independent variable :

- Period_2 : It is a binary indicator variable in which 1 denotes the years present within the second period and 0 for all other years. Second period defines the period of 50 years after the initial period
- Period_3: it is similar to period_2, but indicates the years present in the third period. Third period indicates years that are 100 years after the initial period.
- Const : The constant term accounts for any baseline effect which are not captured by the periodic variables.
- Regression technique :The model developed has used Ordinary Least Squares (OLS) which minimizes the sum of squared residuals between the actual and predicted AMOC temperatures. This finds the best fitting linear relationships between dependent and independent variables.

Learnings from Model :

- The coefficients of both periods (Period_2 and Period_3) are negative which indicates that AMOC temperatures are lower in period_2 and period_3 along with a stronger drop in period_3.
- The coefficient of period_2(-0.0984) is nearly 6 times greater than the coefficient of period_3(-0.53) thereby bring the steeper decline of AMOC temperature during the period_3 compared to that of period_2.
- The R-value of the model(0.480) conveys that the model explains about the variation in AMOC temperatures, which is 48%. This also implies that the model captures an ample portion of the trend observed.

- The plot of observed vs. predicted temperature shows that ability of the model to follow the trend of data in general, in spite of some deviations in the later years, recommend the model doesn't capture the potential influences.

Model Forecast :

The plot present below is the prediction plot of future values of the variable which is trained already in the statistical model above. The prediction chart is,

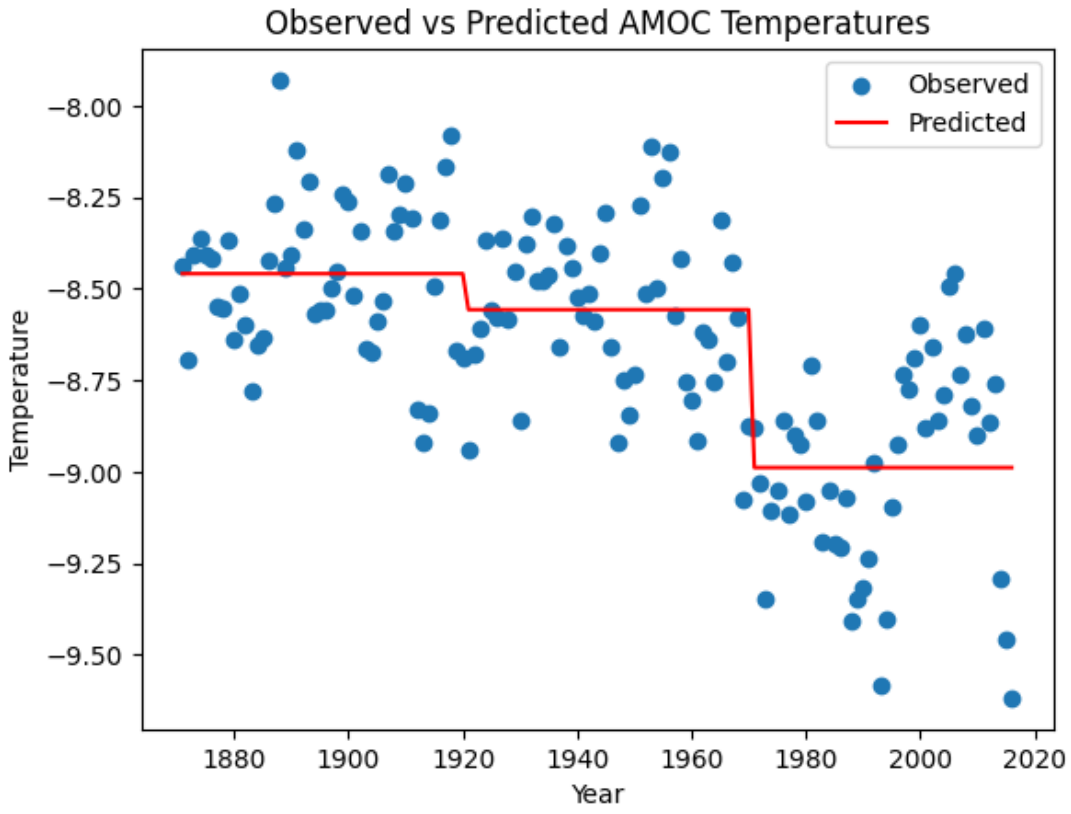

Figure 7. Plot showing the trends over time in prediction plots

6. Results and Conclusion:

Based on the observations from the results of statistical models constructed using the frequentist approach and the recreated Bayesian model on Atlantic Meridional Overturning Circulation (AMOC),that oceanic belt of heat in the Atlantic region, it is clearly evident that there is a presence of downward trend of overtime as stated in the parental research paper “after a long and relatively stable period, there was an initial weakening starting in the nineteenth century, followed by a second, more rapid, decline in the mid-twentieth century, leading to the weakest state of the AMOC occurring in recent decades”. These observations acts as an powerful statement to the power of realm of data science about its truth seeking power, this also acts as a reminder of the requirement of cautious inquiry along with optimum level of uncertainty.

The newly created model using the frequentist approach reveals that there is slight decline in AMOC, while the decline in the nineteenth century unfolds subtly, thereby the precipitous drop happening in the mid twentieth century becomes a dimming star. This paints a more detailed picture on the historical behaviour of AMOC. With the help of long-term trend and offering details of difference in the character of AMOC weaking, the earth’s climatic system makes us better understand of the behaviour of AMOC and it course in the future.

In conclusion, the investigation of AMOC ‘s weaking acts an appealing example of data science’s truth-seeking.it urges us to remain careful in our enquiry and embrace the uncertainty that is inherent and which accompanies scientific endeavours.

7. Future Research :

Although the current study has clarified how the “year” may affect the temperatures of the Atlantic Meridional Overturning Circulation (AMOC), there are still a lot of options for future research to further understand this intricate dynamical system. It would be advantageous to include a wider variety of explanatory variables in the current analysis. Including variables like salinity variations, atmospheric circulation patterns, and even human influences like greenhouse gas emissions might enhance the analysis and reveal more strands influencing AMOC evolution. Future studies can provide a more comprehensive picture of the forces influencing AMOC behaviour by taking into account the interactions between these other variables.

Deeper insights might also be unlocked by expanding the current modelling framework. To gain a more comprehensive knowledge of AMOC dynamics, more advanced model designs that can capture temporal dependencies and non-linear connections between variables should be used. Consider employing powerful statistical or machine learning methods, such as dynamic models or recurrent neural networks. These methods may reveal complex relationships and hidden patterns that are missed by more straightforward models, opening the door to more precise forecasts of the AMOC's future trajectory.

8. References :

1. Adams, R.P. and MacKay, D.J.C. 2007. Bayesian Online Changepoint Detection. arXiv [stat.ML]. Available at: http://arxiv.org/abs/0710.3742.
2. Atenas, J., Havemann, L. and Timmermann, C. 2023. Reframing data ethics in research methods education: a pathway to critical data literacy. International journal of educational technology in higher education 20(1), p. 11. Available at: http://dx.doi.org/10.1186/s41239-023-00380-y.
3. Brown, A.W., Kaiser, K.A. and Allison, D.B. 2018. Issues with data and analyses: Errors, underlying themes, and potential solutions. Proceedings of the National Academy of Sciences of the United States of America 115(11), pp. 2563–2570. Available at: http://dx.doi.org/10.1073/pnas.1708279115.
4. Caesar, L., McCarthy, G.D., Thornalley, D.J.R., Cahill, N. and Rahmstorf, S. 2021. Current Atlantic Meridional Overturning Circulation weakest in last millennium. Nature geoscience 14(3), pp. 118–120. Available at: http://dx.doi.org/10.1038/s41561-021-00699-z.
5. Cheng, L., Trenberth, K.E., Fasullo, J., Boyer, T., Abraham, J. and Zhu, J. 2017. Improved estimates of ocean heat content from 1960 to 2015. Science advances 3(3), p. e1601545. Available at: http://dx.doi.org/10.1126/sciadv.1601545.
6. GLM: Linear regression — PyMC 5.10.3 documentation. 2021. Available at: https://www.pymc.io/projects/docs/en/stable/learn/core_notebooks/GLM_linear.html
7. Hand, D.J. 1983. Common errors in data analysis: the apparent error rate of classification rules. Psychological medicine 13(1), pp. 201–203. Available at: http://dx.doi.org/10.1017/s0033291700050212.
8. Model checking and diagnostics — PyMC 2.3.6 documentation. Available at: https://pymcmc.readthedocs.io/en/latest/modelchecking.html.
9. Lazer, D., Kennedy, R., King, G. and Vespignani, A. 2014. Big data. The parable of Google Flu: traps in big data analysis. *Science (New York, N.Y.)* 343(6176), pp. 1203–1205. Available at: http://dx.doi.org/10.1126/science.1248506.
10. Paskhaver, B. (2017). *Data Analysis with Pandas and Python*
11. Rahmstorf, S. 2002. Ocean circulation and climate during the past 120,000 years. Nature 419(6903), pp. 207–214. Available at: http://dx.doi.org/10.1038/nature01090.
12. Rahmstorf, S., Box, J.E., Feulner, G., Mann, M.E., Robinson, A., Rutherford, S. and Schaffernicht, E.J. 2015. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. Nature climate change 5(5), pp. 475–480. Available at: http://dx.doi.org/10.1038/nclimate2554.
13. Simone Borra, A.D.C. 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics and Data Analysis 54, pp. 2976–2989. Available at: http://dx.doi.org/10.1016/j.csda.2010.03.004.

14. statsmodels.regression.linear_model.OLS - statsmodels 0.14.1. 2023. Available at: https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html.
15. Time Series analysis tsa - statsmodels 0.14.1. 2023. Available at: https://www.statsmodels.org/stable/tsa.html.
16. Xie, M.-G. and Singh, K. 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review: Confidence distribution, the frequentist distribution estimator. Revue internationale de statistique [International statistical review] 81(1), pp. 3–39. Available at: http://dx.doi.org/10.1111/insr.12000.