A study on factors linked to precipitation*

Zhandong Cao

26 April 2022

Abstract

In agriculture, the degree of regional climate change, especially the regional precipitation, is particularly important. Sufficient rainfall is helpful for the normal growth of crops, but excessive or insufficient rainfall will have adverse effects on crops. At this time, artificial intervention on crop water content is required. In this paper, the yearly precipitation in England and Wales are collected and explored via multiple linear regression model and two sample t-test, then the variability of raindrop is investigated in order to give feasible suggestions to the algriculture.

Keywords: Precipitation, rainfall, regression model, multiple comparison

Contents

*Code and data are available in this GitHub repository: https://github.com/ZhandongCao0601/A-study-on-factors-linked-to-precipitation.git

Introduction

Precipitation is some liquid or frozen dihydrogen monoxide that generated in the atmosphere and drop to the surface of the ground. All plants require water for chemical reactions including but not limited to photosynthesis, therefore precipitation is of vital importance to algriculture. Sufficient rainfall is helpful for the normal growth of crops, but excessive or insufficient rainfall will have adverse effects on crops, even destroying to plants. Understanding the precipitation for agricultural crops is critical for developing cropping systems resilient to stresses induced by climate change (Hatfield (2011)).

Scientists have revealed great impact and interference that human can performed based on the precipitation system. Rainwater harvesting agriculture is performed as a system of maintaining the production of crops. The average annual precipitation (APP) information system, which was created using optimised methods and raster precipitation spatial databases, can quickly and accurately calculate total quantities and spatial shifts of precipitation resources on almost any measurements in the study areas, which is useful for runoff simulation, engineering planning, strategy development, and decision making, along with water management in rainwater harvesting agriculture (HongWei (2005)).

This paper tries to find the influential factors via statistical model on the local precipitation in England and seeks to understand how there variability is influenced. The data comes from Met Office's Hadley Centre and contains the yearly winter raindrops England and Wales for 1951-99. To help us understand the key factors involved in determining precipitation in England and Wales, we raise a few questions and start by answering the following questions thus leading to further insights and possible summations on the precipitations. The question of interest is as follows:

- From the explanatory description analysis, what variables is highly related to the precipitation in England and wales?
- From the statistical methods adopted, what variables significantly influences the percipitation?
- If we divide the yearly raindrop into two time interval (from 1951 to 1975 versus from 1876 to 1999), are the two time period average raindrops differ from each other signifiantly in statistics?

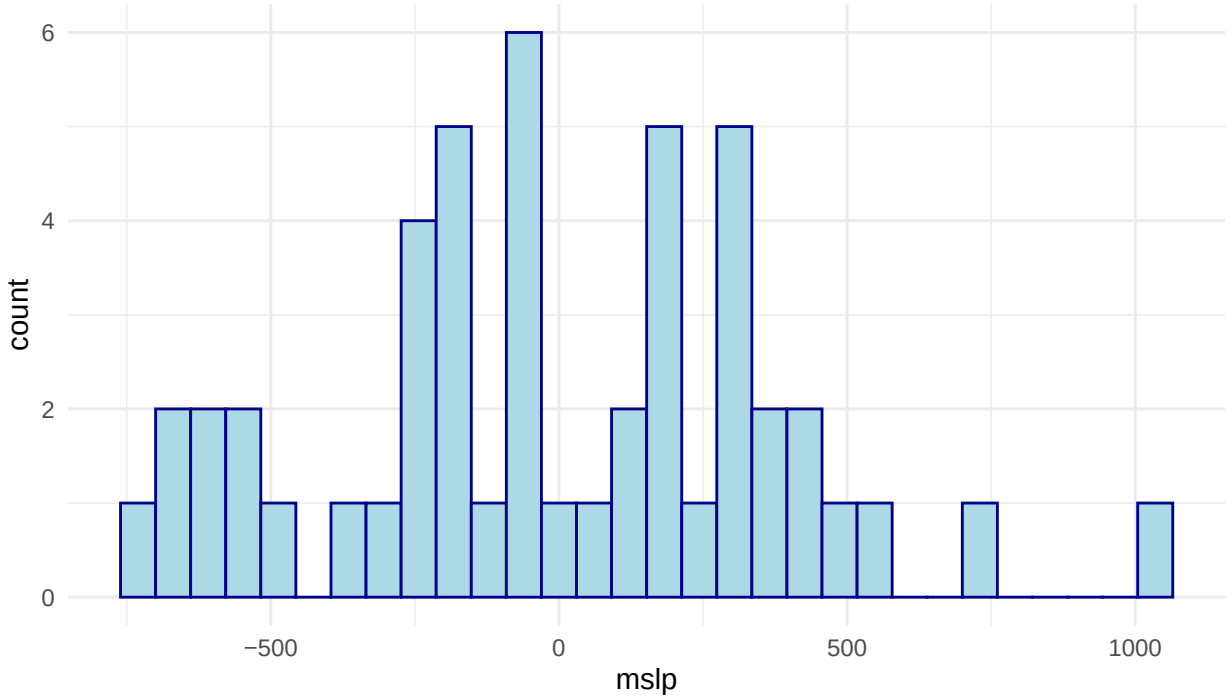
For the first question, we can use basic data analysis methods to determine which variables have a linear relationship with rainfall by drawing a scatter plot of independent versus dependent variable precipitation, and the Pearson correlation coefficient is also an effective method. The second question we want to find statistically significant variables by building a multiple linear regression model. The third problem can be solved by using the t-test of quantitative samples to judge whether the sample means of the two time periods are significantly different through hypothesis testing and P-value.

This paper first provides the basic statistics as well as figures that display the overall distribution and other information about the original dataset in **Data** section. To solve the questions of interest above, some methods such as linear regression model, t test for two observational series are adopted, **Methodology** section gives a brief concept and introduction of the methods used in our paper. **Result** shows the R outcome and the result yield from our statistical methods and **Discussion** section will be elebaorated around the understanding of percipitation in England and Wales based on the methods result, the drawbacks of the methods as well as the possible future works. The data set will be processed using R (R Core Team (2020)), and a few packages inside R called tidyverse (Wickham et al. (2019)). Most figures and tables are also being done with R using ggplot2 (Wickham (2016)), dplyr(Wickham et al. (2021)), kableextra (Zhu (2021)), moment(Komsta and Novomestky (2015)), reshape2(Wickham (2007)), and gridExtra(Auguie (2017)).To generate this R markdown report, knitr (Xie (2014)) package was also used.

Data

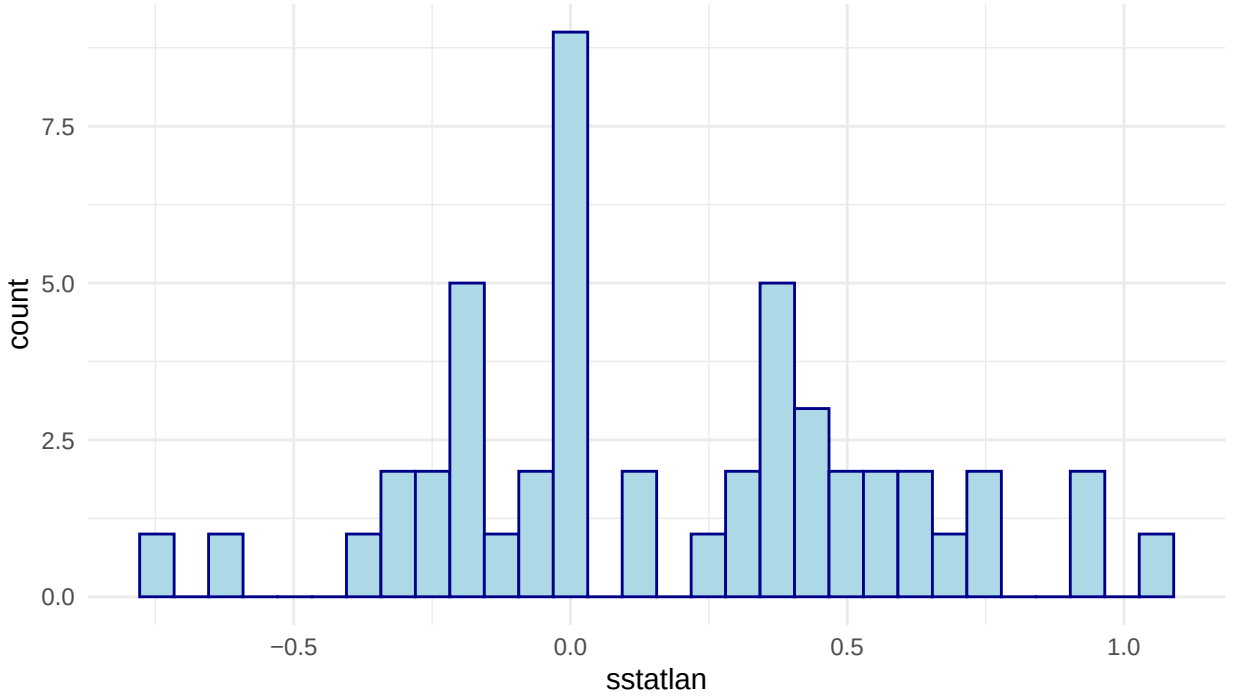
First of all, below is a quick EDA of the data and some quick view of the Distributions of the variables.

Fig 0.1: Distributions of the variables (barplot of mslp)



We can see that the majority of the data was Distributed from -250 to 500 mslp.

Fig 0.2: Distributions of the variables (barplot of Sea−surface temperature)



We can see from the above plot that the the sea-level temp is mostly 0 and evenly distribute on the x-axis.

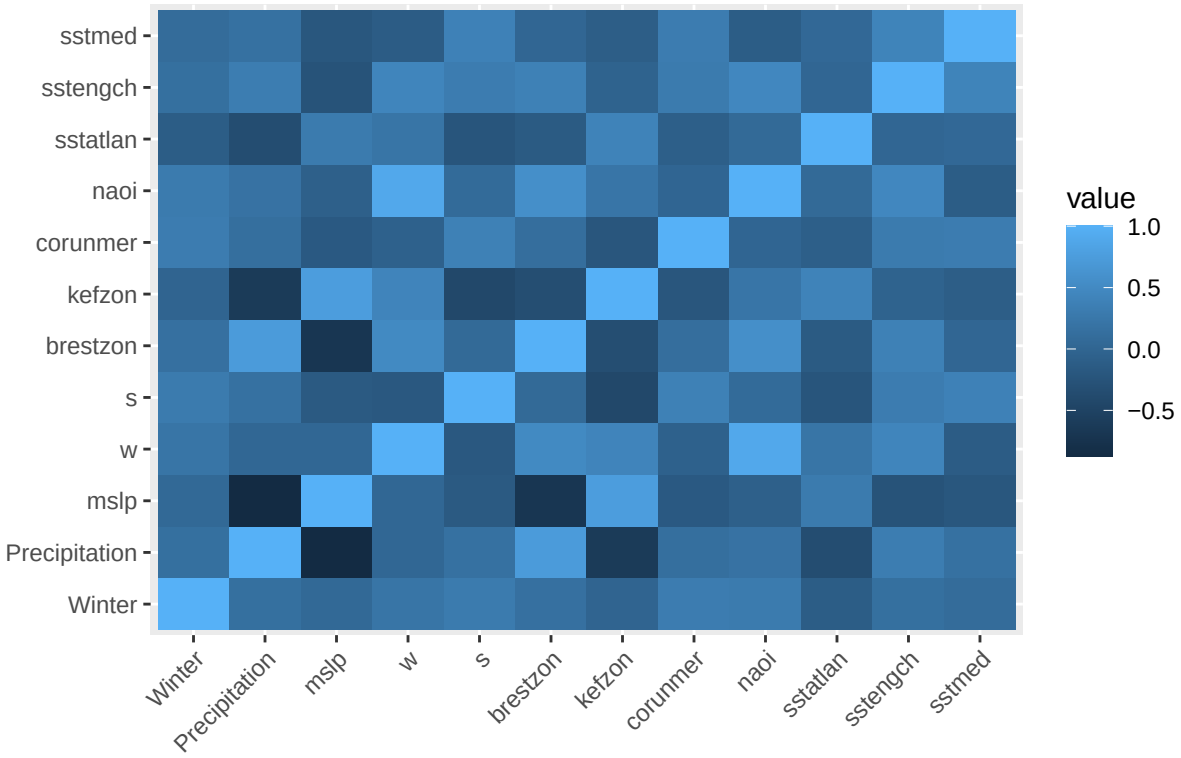
The data comes from Met Office's Hadley Centre and contains the yearly winter raindrops England and Wales for 1951 to 1999(Alexander and Jones (2001) Jones and Conway (1997) Gregory J. M. and Wigley (1991) T. M. L. Wigley and Jones (1987) J. M. L. Wigley T. M. L. and Jones (1984)). The description for each variable can be viewed from the table below. We first obtain information about the data as a whole through descriptive statistics. Descriptive statistics include calculating the mean, median, quantile, kurtosis, and skewness of a sample. To help the readers better understand the variables involved with climate terms, we divide the ten independent variables into three sub-sets:

1. Pressure and wind factors (mean sea-level pressure, westerly and southerly wind-force, North Atlantic oscillation index) which can be significant controlling factors of Earth's weather and climate.
2. Water vapour flux components (zonal at $Brest^1$, $Keflavik^2$ and $La\ Coruna^1$) which influence the precipitation by different gas content.
3. Sea-surface temperatures (eastern North Atlantic, $English\ Channel^1$ and $Mediterranean\ Sea^2$) which is associated to the amount of atmospheric water vapor over the oceans and the risk of raindrops.

Variable	Description	Mean	Median	SD	Range	IQR	Skew	Kurt
rainfall	Winter precipitation	253.6	252.7	72.1	332	85.7	0.30	2.92
mslp	Mean sea-level pressure	-12.35	-53.2	382.8	1765	529.3	0.18	2.87
w	The westerly wind component	18.61	3.57	166.3	704.9	237.1	-0.06	2.42
s	The southerly wind component	-7.12	14.63	100.3	452.9	125.1	-0.31	2.75
brestzon	The zonal water vapour $flux^1$	0.04	0.22	0.80	3.74	1.19	-0.33	2.71
kefzon	The zonal water vapour $flux^2$	0.06	-0.04	0.56	2.61	0.73	0.41	2.76
corunmer	The meridional water vapour $flux^1$	0.13	0.14	1.88	9.65	2.15	-0.61	3.86
naoi	The North Atlantic oscillation index	0.35	0.14	1.35	5.46	1.66	-0.03	2.50
sstatlan	Sea-surface temperature	0.17	0.10	0.41	1.81	0.61	0.05	2.49
sstengch	Sea Surface Temperature $anomalies^1$	0.06	0.08	0.49	2.77	0.59	-0.30	4.41
sstmed	Sea Surface Temperature $anomalies^2$	0.11	-0.10	0.84	4.56	1.20	0.30	3.68

The correlation matrix heatmap helps us to quickly observe the correlation between dependent and independent variables in the dataset. The heatmap below displays the correlation coefficients, the darker blue color the bar is, the closer the coefficient is to 1, the lighter blue color the bar is, the closer the coefficient is to -1.

Fig.1: Correlation matrix heatmap

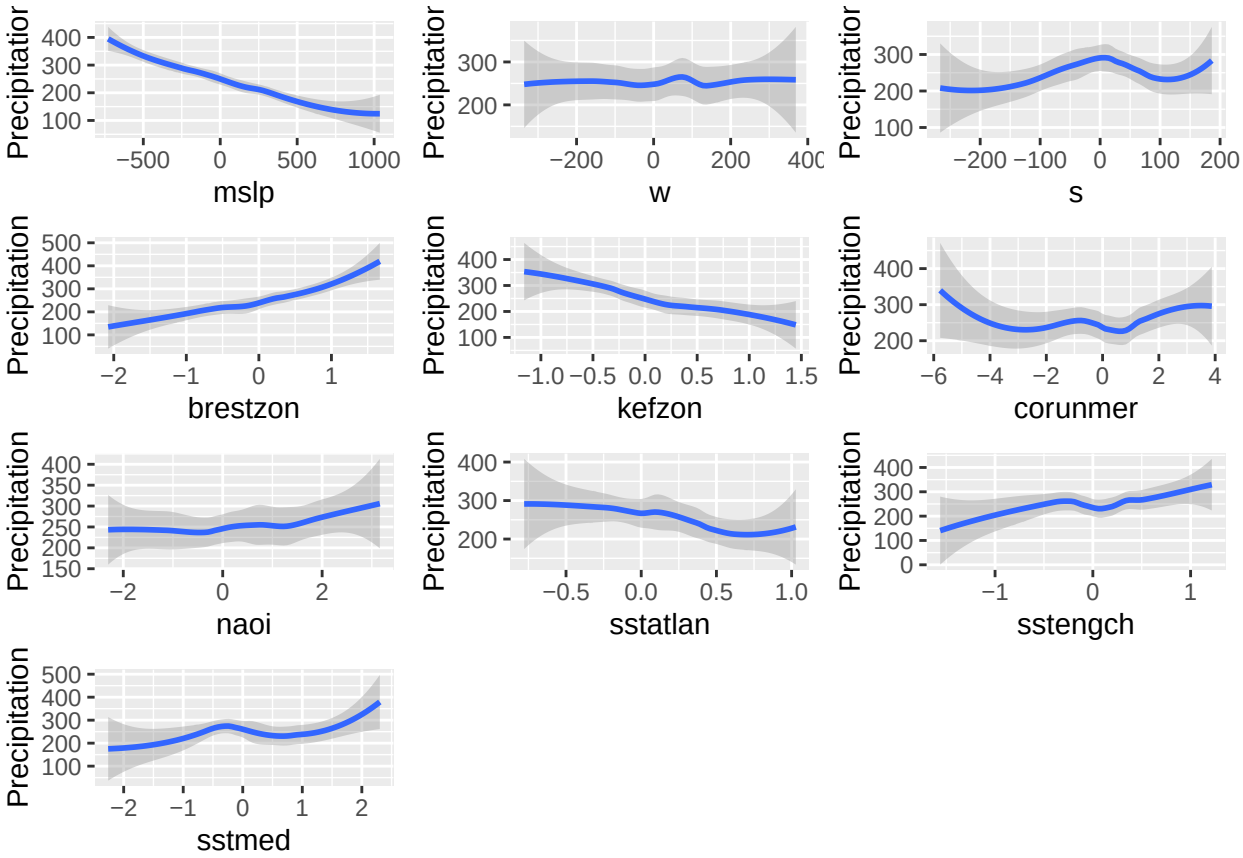


The square of Pearson correlation coefficient is equivalent to R^2 for the simple linear regression model. Pearson correlation coefficient formula is as follows:

$$r_{XY} = \frac{Cov\,(X, Y)}{\sqrt{Var\,(X) * Var\,(Y)}}$$

The absolute value that is closer to 1 for the independent variable and the dependent variable precipitaion indicates the nessecity to take the independent variable into consideration for building the regression model. For most of the case, the absolute value of ρ means there might exist linear relationship between two variables. From the heatmap and R calculation result, we see five variables satisfies this: mslp, brestzon, kefzon, sstatlan and sstengch.

The Pearson correlation coefficient gives us a initial relation between the variables, the next goal is to look at the specific scatter plot to see if the linear relation exist.



Now we provide the scatter plot of percipitation versus other independent variable. The plot shows mslp, brestzon, kefzon, sstengch and sstmed seem to have linear relation with the percipitation variable. This is consistent with the Pearson correlation coefficient result. So in our regression model, these five variables are taken into consideration.

Methodology

This paper tries to find the influential factors via statistical model on the local precipitation in England and seeks to understand how there variability is influenced. By answering the three questions we raise above, we are more familiar with and gain insight of the whole dataset. To answer the first question: **From the explanatory description analysis, what variables is highly related to the precipitation in England and wales?**, we use the Pearson correlation coefficient and provide the analysis result in **Data** section. For the second question: **From the statistical methods adopted, what variables significantly influences the percipitation?**, we are going to utilize the multiple linear regression model. The initial model tries to take the precipitation as the response variable and five variables (mslp, brestzon, kefzon, sstengch and sstmed) as the predictors to build the regression relation. The regression equation is given below:

$$Precipitation = \beta_0 + \beta_1 mslp + \beta_2 brestzon + \beta_3 kefzon + \beta_4 sstengch + \beta_5 sstmed + \epsilon$$

By looking at the sign and magnificance of regression coefficient, we are able to exlpore the relations between the five key factors and the precipitation variable. The last question: **If we divide then yearly raindrop into two time interval (from 1951 to 1975 versus from 1876 to 1999), are the two time period average raindrops differ from aech other signifiantly in statistics?** is related to the overall variation of the precipitation in England and Wales. We are interested in answering this as we might learn whether some instant strategies are necessary to adopt if the precipitation varies dramatically. To answer this

question, two sample t test are used. We manly focus on the average precipitation level as well as consider the variation within each time interval. The hypothesis test method are used and the statistic (follows student t distribution if null hypothesis are true) are given as follows:

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},\ df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n1)^2}{n_1 - 1} + \frac{(s_2^2/n2)^2}{n_2 - 1}}$$

By looking at the P-value of the hypothesis test, we are able to answer question 3.

The other two key methods that I will use to anaylise the data is by examing P-value and performing a t-test on the statistical model that was build.

P-value:

The p-value is a number produced conducted during a statistical hypothesis test that represents the likelihood of finding of a set of observations.It is used in hypothesis testing to help determine if the null hypothesis should be rejected. The lower the p-value, the more likely the null hypothesis will be rejected. Usually, we consider the P-value as low when it is less than 0.05, which we say there are some evidence to reject the null hypothesis.

T-test:

A t-test is an inferential statistic method that is used to see if there is a significant difference in the means of two groups that are connected in some way. A t-test is a hypothesis testing technique that may be used to assess an assumption that is relevant to a population.

Results

What is linked to the precipitation

In Data section, we see the variables that have Pearson correlation coefficient between precipitation that is higher than 0.3 or smaller than -0.3 are: mslp, brestzon, kefzon, sstatlan and sstengch. In addition, the smooth line fitted by built-in R function tells us the same result that these five variables is more linearly related to precipitation. By looking at the five variables, we can see pressure and wind-force (ocean sea-level pressure over the British Isles), water vap-flux components (Brest and Keflavik) and sea-surface temperatures (eastern and English) are all related to the precipitaiton in England and Wales.

What significantly impact the precipitation

In data section, mslp, brestzon, kefzon, sstengch and sstmed seem to have linear relation with the percipitation variable from the scatter plot and the Pearson correlation coefficient, so the five variables out of the ten variables are going to be used in the multiple linear regression model. The regression model equation is as follows:

$$Y = \beta_0 - 0.128X_1 + 18.422X_2 + 1.073X_3 - 23.053X_4 + 9.638X_5 + \epsilon$$

The vairable and its representation is:

- Y: precipitation
- X_1: mean sea-level pressure above British Islesl

- X_2: zonal water vap-flux anomaly at Brest
- X_3: zonal water vap-flux anomaly at Keflavik
- X_4: above sea temperature at eastern Atlantic
- X_5: above sea temperature at English

The regression output is given below:

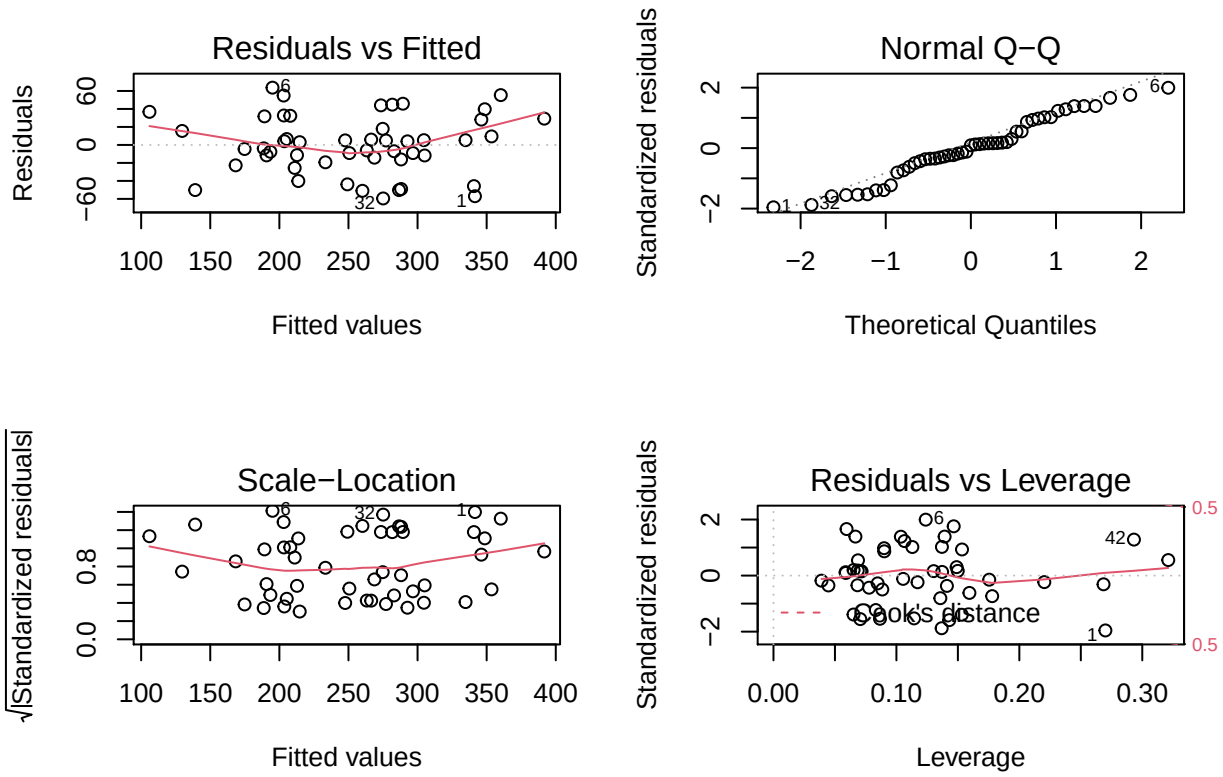
Variable	Description	Estimate	Std. Error	t value	P-value
mslp	Mean sea-level pressure	-0.12783	0.02785	-4.590	3.83e-05
brestzon	The zonal water vapour $flux^1$	18.42217	9.6272	1.914	0.0623
kefzon	The zonal water vapour $flux^2$	1.07303	15.02184	0.071	0.9434
sstatlan	Sea-surface temperature	-23.05318	13.14552	-1.754	0.866
sstengch	Sea Surface Temperature $anomalies^1$	9.63816	10.86659	0.887	0.38

As we can see, the overall impact for the five variables are:

- If mean sea-level pressure raises 1 millibars, the average percipitation is expected to decrease 0.128 millimetres.
- If zonal water vap-flux anomaly at Brest raises 1 kilograms per metre per millibar per second, the average percipitation is expected to increase 18.422 millimetres.
- If zonal water vap-flux anomaly at Keflavik raises 1 kilograms per metre per millibar per second, the average percipitation is expected to increase 1.073 millimetres.
- If above sea temperature in eastern raises 1 Degrees Celsius, the average percipitation is expected to decrease 23.053 millimetres.
- If above sea temperature in English Channel raises 1 Degrees Celsius, the average percipitation is expected to increase 9.638 millimetres.

Under 90% confidence level (P-value is smaller than 0.1), variables mslp, brestzon and sstatlan are significant. The R^2 of the regression model is 0.80, which means 80% variation in raindrop is explained by the multiple regression model. In conclusion, when we consider what factor plays key role in determining the precipitation in England and Wales, we should see the mean sea-level pressure over the British Islesl jave significant impact on the precipitation, the zonal water vap-flux anomaly at Brest has significant impact on the precipitation and the sea-surface temperatures in eastern North Atlantic statistically significant influence the precipitation

In addition, the mean sea-level pressure and the sea-surface temperatures over the British Islesl have positive impact and the zonal water vap-flux at Brest have negative impact, the zonal water vap-flux at Brest and the sea-surface temperatures over the British Islesl have much more impact than the mean sea-lvel pressure (18.42 versus -0.13 and -23.05 versus -0.13).



The above four plots gives us the goodness of fit for the regression model. We can tell from the residuals versus fitted plot that the residual is overall zero mean but are not spread equally around, which means the variane is not constant. The normal Q-Q plot tells us the normality are satisfied for the regreesion model and there seems to be no high leverage points.

Are precipitation amount varies over time

If we divide then yearly raindrop into two time interval (from 1951 to 1975 versus from 1876 to 1999), are the two time period average raindrops differ from aech other signifiantly in statistics? The two sample t test (independent and unequal variance assumption needed) can be utilized to answer the question. The test statistic is -0.996, numberically means the average precipitation decreases over the time period as the sign is negative. But if we take the variation in each samples into consideration, the P- value is 0.32, which means the negative sign (decerase in average precipitation in England and Wales) is within expection, the decrease results from the slight variation in the two samples. So statistically we say there is no difference in mean of raindrops for the two time period (from 1951 to 1975 versus from 1976 to 1999).

Discussion

The discussion section is divided into two subsections. The first subsection is about the interpretation of the result for the three questions of interest and the understanding of precipitation in England and Wales, the second subsection is about underlying drawbacks in the methodology and possible future adjustment works in these methods.

Interpretation of result and understanding in precipitation

From the statistical analysis result, we see pressure and wind-force (ocean sea-level pressure over the British Isles), water vap-flux components (Brest and Keflavik) and above sea temperatures (eastern and English) are all related to the precipitaiton in England and Wales.

Further more, the regression model result tells us the mean sea-level pressure over the British Islesl significantly influence the raindrops, if mean sea-level pressure raises 1 millibars, the average percipitation is expected to decrease 0.128 millimetres; the zonal water vap-flux anomaly at Brest significantly influence the raindrops, if zonal water vap-flux anomaly at Brest raises 1 kilograms per metre per millibar per second, the average percipitation is expected to increase 18.422 millimetres; the above sea temperatures in eastern North Atlantic significantly influence the precipitation, if above sea temperature in eastern raises 1 Degrees Celsius, the average percipitation is expected to decrease 23.053 millimetres.

By looking at the three significant factors, we see the zonal water vap-flux anomaly at Brest and above sea temperatures over the British Islesl have much more impact than the mean sea-lvel pressure (18.42 versus -0.13 and -23.05 versus -0.13).

To better gives feedbacks for the algriculature in England and Wales, the basic mean sea-level pressure is needed to monitor. In addition, the zonal water vap-flux anomaly at Brest is associated with the precipitation in England but not the zonal water vap-flux anomaly at iceland (which is closer to England), the possible reason is that they are at the same latitude. So the suggestion is to monitor the zonal water vap-flux anomaly of the geographical location that have the same latitude to the England and Wales. The sea-surface temperatures over the British Islesl is neccesary to monitor as it is closer to England mainland and is more likely to influence the nearby precipitation. When we talks about the overall variation of the precipitation in England and Wales, we observe there is no significant difference for the two seperated time periods, so no instant strategies are necessary to adopt as the precipitation does not vary dramatically.

Underlying drawbacks and future adjustment works

In the methodology and result part, we see some underlying drawbacks for the methods. The regression diagnostics shows the residuals are not having the constant variance, which means some transformations such as box-cox transformations are needed to solve the problems. In addition, we only takes five variables into consideration in the regression model. In the regression model, $R^2 = 0.7997$, which means 79.97% variation in damage percent can be explained by the distance. If we can collect more essential information(i.e. the local in-land temerature) and take more variable into consideration, then we can raise the R^2. However the higher R^2 might result from overfitting, so the future work should consist of both collecting more variates as well as variable selection and dimension reduction procedure.

Appendix

Data sheet:

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was innitialy created to serve as agricultural use. It is to record tha precipitation of area across the UK.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset was created by a group of scientist from the 90s, they are:Alexander and Jones (2001) Jones and Conway (1997) Gregory J. M. and Wigley(1991) T. M. L. Wigley and Jones (1987) J. M. L. Wigley T. M. L. and Jones (1984)
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - This part of information is missing.
4. *Any other comments?*
    - N/A

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - This data set was recorded in the UK only, and it served as national data. and is sepreated by different measurement that the scientiset need. They are sea-level data, water vapour data, and wind data.
2. *How many instances are there in total (of each type, if appropriate)?*
    - sea-level data:4
    - water vapour: 4 _ wind data: 2
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - It is almost impossible to consider all factors when it comes to a study of nature, also, this is limited by the tech they have in the 90s.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - all data is numeric.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - There are no associated label or target to each instance.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- There are a few missing data across the data set, but they are rather little considering the size of the data.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - There are no particular relationships between the instances.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no errors, noise, or redundancies in thea dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - This dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There are no data contained in the data set that will be considered condidential.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - This data did not include any information that will be considered offensive.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The dataset studies the information of nature, and should not include any sub-population.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - Due to the nature of this data-set, it is not possible to identify any individuals in this data-set.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - This data set did not include any data that will be considered sensitive in any aspect.
16. *Any other comments?*
    - N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - This data associated with each instance is through the scientifical method.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - The procedure was by examing the nature pieces of information measured by the scientists.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - The dataset is not a sample from any existing larger set.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - The data was collected by the scientists in the UK.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data was collected from 1984 to 2001.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - There are no ethical review processes included.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - The data was not collected by myself.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - There are no individual needed to be questioned.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - There are no individual needed to be questioned.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - There are no individual needed to be questioned.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - There are no analysis of potential impact of the dataset.

12. *Any other comments?*
    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Yes, there are cleaning done by removing the Nas (missing data) from the data.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - The 'raw' data is saved in the input folder.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R the programing language used, it is available to the public.
4. *Any other comments?*
    - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The data-set may be used in other tasks, but the information was not publiced.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - Code and data are available in this GitHub repository: https://github.com/ZhandongCao0601/STA304final.git
3. *What (other) tasks could the dataset be used for?*
    - Any nature related task.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - There are no information about the composition of the dataset.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - N/A
6. *Any other comments?*
    - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The data-set will not be distributed to third parties outside of the entity.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The data will be distributed on GitHub.
3. *When will the dataset be distributed?*
    - The data-set will be distributed in April 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - The dataset will not be distributed under a copyright or other intellectual property.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - There are no third parties imposed IP-based or other restrictions on the data associated with the instances.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - There are no export controls or other regulatory restrictions apply to the dataset or to individual instances.
7. *Any other comments?*
    - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Zhandong Cao (Leo)
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - By E-mail at "Zhandong.cao@mail.utoronto.ca"
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There are no erratum
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - There are no plane to update the data-set. It is hard to get data from UK.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - There are no pplicable limits on the retention of the data associated with the instances.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older version will not be hosted.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - The repository can be extend/augment/build on/contribute to. I (the writer of this report) will not be responsible for any responsivlity.

8. *Any other comments?*

    - N/A

References

Alexander, L. V., and P. D. Jones. 2001. “Updated Precipitation Series for the u.k. And Discussion of Recent Extremes.” https://doi.org/10.1006/asle.2001.0025.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Gregory J. M., P. D. Jones, and T. M. L. Wigley. 1991. “Precipitation in Britain: An Analysis of Area-Average Data Updated to 1989.”

Hatfield, Agron. J. 2011. “Climate Impacts on Agriculture: Implications for Crop Production.” https://doi.org/10.2134/agronj2010.0303.

HongWei, Tian-Gang-Liang, Jian-LongLi. 2005. “Study on the Estimation of Precipitation Resources for Rainwater Harvesting Agriculture in Semi-Arid Land of China.” *Agricultural Water Management* 71 (1): 33. https://doi.org/10.1016/j.agwat.2004.07.002.

Jones, P. D., and D. Conway. 1997. “Precipitation in the British Isles: An Analysis of Area-Average Data Updated to 1995.”

Komsta, Lukasz, and Frederick Novomestky. 2015. *Moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. https://CRAN.R-project.org/package=moments.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2007. “Reshaping Data with the reshape Package.” *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wigley, J. M. Lough, T. M. L., and P. D. Jones. 1984. “Spatial Patterns of Precipitation in England and Wales and a Revised Homogeneous England and Wales Precipitation Series.”

Wigley, T. M. L., and P. D. Jones. 1987. “England and Wales Precipitation: A Discussion of Recent Changes in Variability and an Update to 1985.”

Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.