A study on factors linked to precipitation[*]

Zhandong Cao 1005689391

2022/04/24

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/STA304final.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

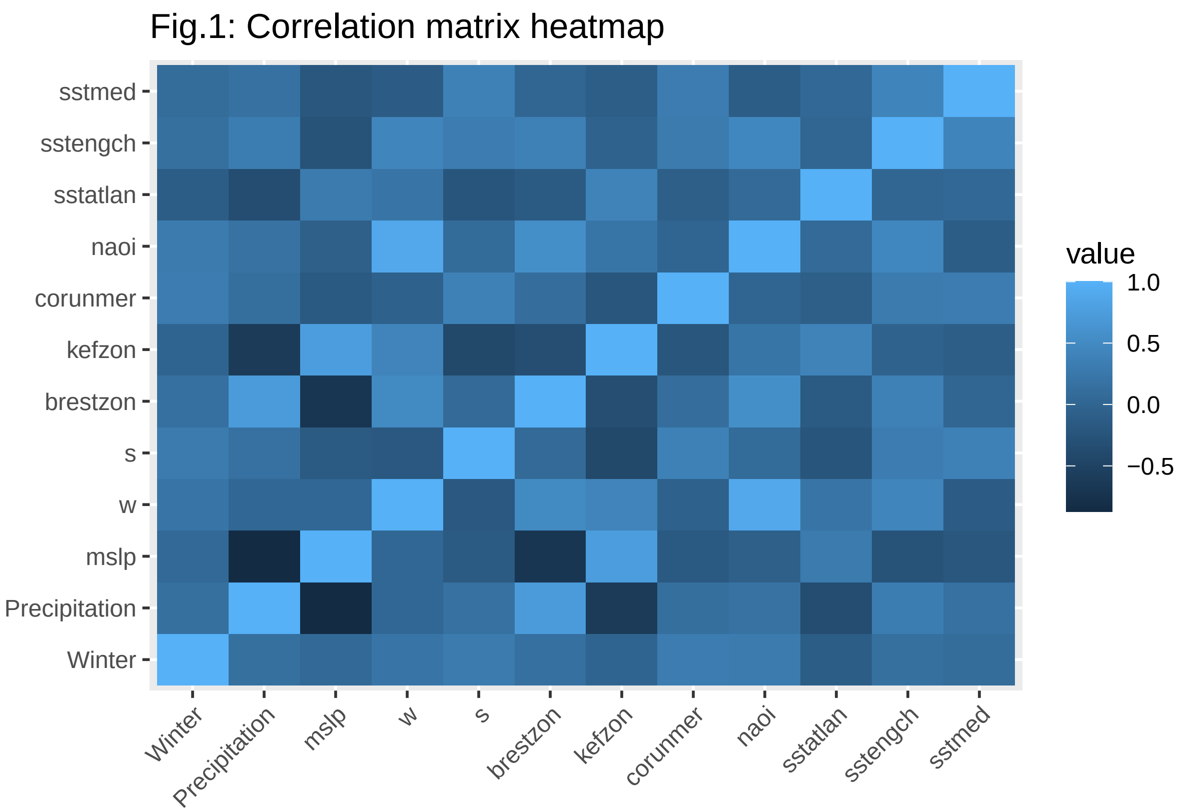
The data comes from Met Office's Hadley Centre and contains the yearly winter raindrops England and Wales for 1951 to 1999(Alexander and Jones (2001) Jones and Conway (1997) Gregory J. M. and Wigley

(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.

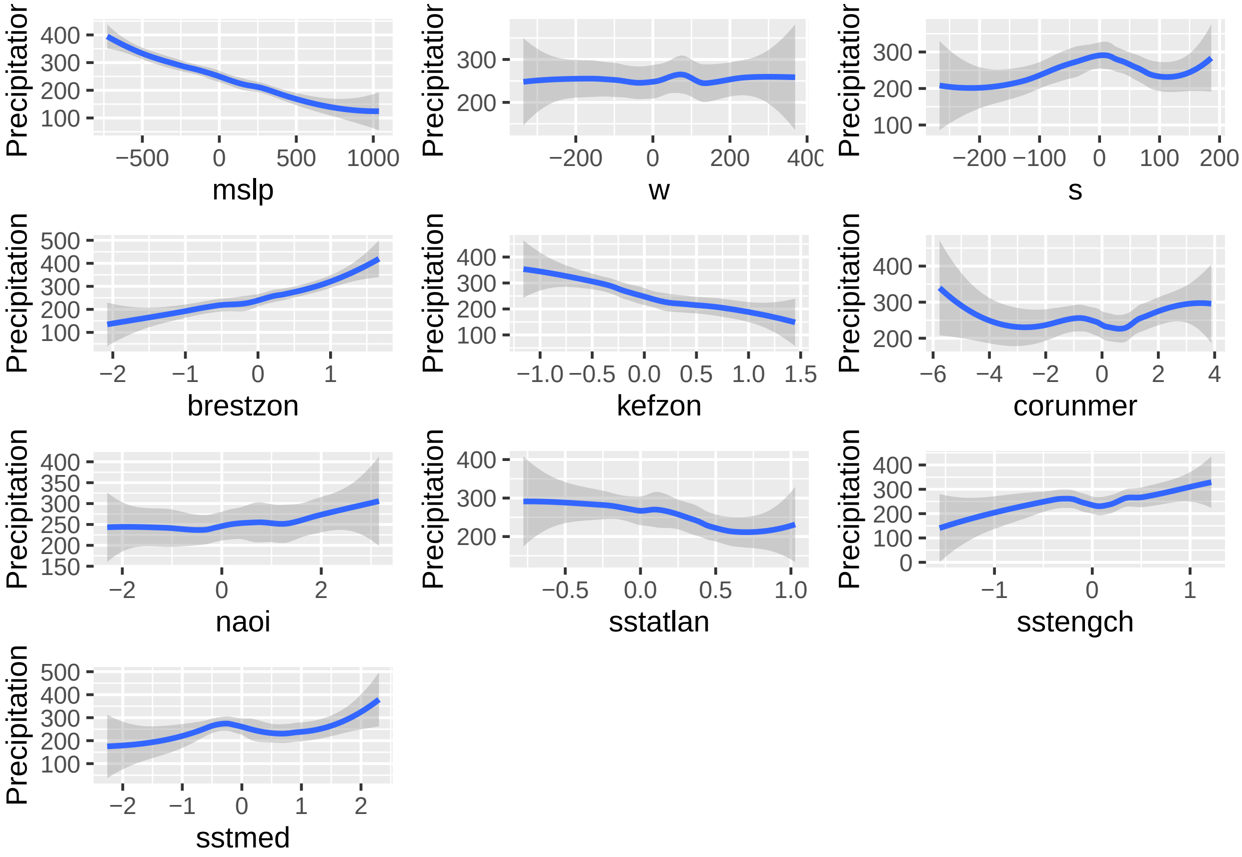


The square of Pearson correlation coefficient is equivalent to R^2 for the simple linear regression model. Pearson correlation coefficient formula is as follows:

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

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.

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:

```
##
## Call:
## lm(formula = Precipitation ~ mslp + brestzon + kefzon + sstatlan +
##     sstengch, data = precipitation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.498 -16.157   3.078  28.150  63.656
```

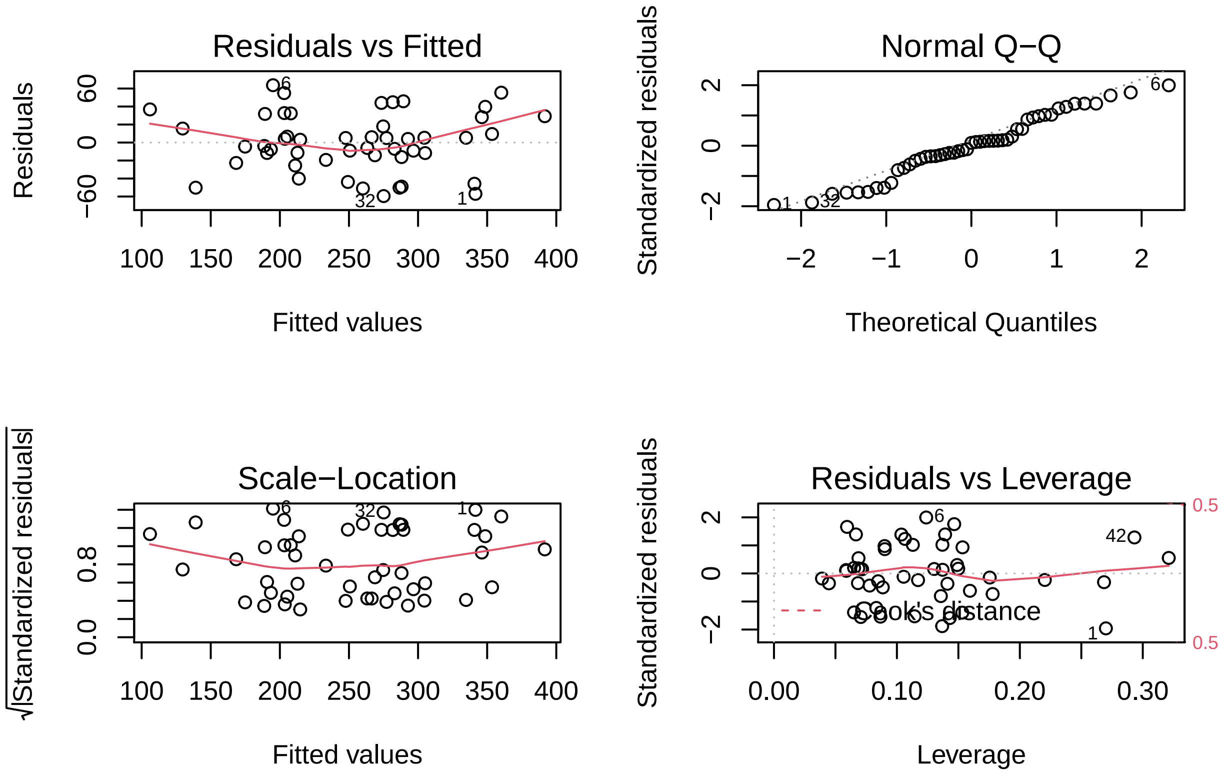
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 254.46036    5.38287  47.272  < 2e-16 ***
## mslp         -0.12783    0.02785  -4.590 3.83e-05 ***
## brestzon     18.42217    9.62720   1.914   0.0623 .
## kefzon        1.07303   15.02184   0.071   0.9434
## sstatlan    -23.05318   13.14552  -1.754   0.0866 .
## sstengch      9.63816   10.86659   0.887   0.3800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.09 on 43 degrees of freedom
## Multiple R-squared:  0.7997, Adjusted R-squared:  0.7764
## F-statistic: 34.33 on 5 and 43 DF,  p-value: 5.758e-14
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -0.99646, df = 46.13, p-value = 0.3242
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -61.77327  20.86227
## sample estimates:
## mean of x mean of y
##  243.1125  263.5680
```

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.

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.