Universitat Autònoma de Barcelona

The influence of sea surface temperature on tropical-cyclone intensity

Alfredo Hernández Cavieres

Supervised by: Álvaro Corral & Isabel Serra

11th July 2018

"All models are wrong, but some are useful."
— George Box

Abstract

The influence of global warming on the intensity of tropical-cyclones (TC, also called hurricanes or typhoons in an abuse of language) is a rather controversial topic that has already been addressed in many statistical studies, such as Webster et al. [1]. For the North-Atlantic basin, it has been shown [2] that the probability distribution of the so-called power-dissipation index (PDI, a rough estimation of released energy) is indeed affected by the anual and basin-wide averaged sea surface temperature (SST), displacing towards more extreme values on warm years (high SST). As the PDI integrates (cubic) wind speed over tropical-cyclone lifetime, it is an open question where the PDI increase comes from (higher speed, longer lifetime, or both).

Our empirical results show a remarkable correlation in the joint distribution of lifetime and PDI, and linear regression of the logarithmic variables yields a power-law relation between both. Statistical testing, by means of a permutation test, shows that this relation does not significantly depend on the SST. In other words, the wind speed of a tropical cyclone of a given lifetime will be the same (within statistical fluctuations) in cold and warm years. Nevertheless, the increase of TC lifetime with SST triggers an increase in wind speed as a by-product. Further analysis shows that the longer lifetimes are mainly due to a shift to South-East of the TC genesis point.

Our conclusions are compatible with the view of tropical cyclones as an activation process, in which, once the event has started, its intensity is kept in critical balance between attenuation and intensification (and so, higher SST does not trigger more intensification). Summarizing, storms with the same lifetime should have the same speed and PDI, no matter the SST, as, once the cyclone is activated, the wind speed at each TC stage should not depend on the SST.

Contents

List of Figures

List of Tables

1 Introduction

For the North Atlantic (N. Atl.) and Northeast Pacific (E. Pac.) basin, it has been shown [2] that the probability distribution of the so-called power-dissipation index (PDI, a rough estimation of released energy) is affected by the annual and basin-wide averaged sea surface temperature (SST), displacing towards more extreme values on warm years (high-SST).

As the PDI integrates (cubic) wind speed over tropical-cyclone (TC) lifetime, it is an open question where the PDI increase comes from (higher speed, longer lifetime, or both).

1.1 Tropical-cyclones

To characterise a tropical-cyclone one needs to define a physically relevant measure of released energy. In [3] Emanuel showed that in a tropical-cyclone energy dissipation occurs mostly in the atmospheric surface layer, and that the corresponding dissipation rate per unit area, D, is

$$D \equiv C_D \rho \left\| \vec{v} \right\|^3, \tag{1}$$

where C_D is the surface drag coefficient, ρ is the surface air density, $\left\| \vec{v} \right\|$ is the magnitude of the surface wind velocity.

Thus, integrated over the surface area covered by a circularly symmetric tropical-cyclone of radius r_0 of lifetime τ, the total power dissipated by the storm, PD, is given by

$$PD \equiv 2\pi \int_0^\tau \int_0^{r_0} C_D \rho \left\| \vec{v} \right\|^3 r \, \mathrm{d}r \, \mathrm{d}t \, . \tag{2}$$

However, as stated in [4], since the integral in the expression (3a) will in practice be dominated by high wind speeds, one can approximate the product $C_D \rho$ as a constant and define a simplified power dissipation index (PDI) as

$$PDI \equiv \int_0^\tau v_{max}^3 \, \mathrm{d}t \, . \tag{3a}$$

However, the wind data (as we discuss in § 3.1) is recorded every $\Delta t = 6\,\mathrm{h}$. Therefore, we discretise the expression for the PDI using the rectangle method:

$$PDI = \sum_t v_t^3 \Delta t, \tag{3b}$$

where v_t is the maximum sustained surface wind speed at time t.

Although the PDI is enough to characterise the released energy by a tropical-cyclone, we are interested in the causal relationship between increasing tropical-cyclone PDI and increasing sea surface temperature (SST) proposed by Trenberth [5], as well as the relationship with the lifetime associated to each tropical-cyclone.

Trenberth states that higher SSTs are associated with increased water vapour in the lower troposphere; both higher SSTs and increased water vapour tend to increase the energy available for atmospheric convection and the energy available to tropical-cyclones as a consequence.

By separating the PDI data by low-SST and high-SST years, we get two similar PDI distributions with one major difference: high-SST years should have a longer right tail on account of having more available energy from the sea (this is explored with more detail in § 4.1.2). For this separation (or classification) process we follow the methodology used by Corral et al. in [2], which is a variation of the methodology proposed by Webster et al. in [1].

To perform this classification we need to calculate first the mean SST of each year $\langle \mathrm{SST} \rangle$:

$$\langle \mathrm{SST} \rangle = \sum_y \frac{\mathrm{SST}(y)}{Y}, \tag{4}$$

where $\mathrm{SST}(y)$ is the mean SST of the year y, and Y is the total number of years studied; as usual, the standard error of this mean, is defined as

$$\widehat{\mathrm{se}}(\mathrm{SST}) = \frac{1}{\sqrt{Y}} \sqrt{\frac{1}{Y-1} \sum_y \big(\mathrm{SST}(y) - \langle \mathrm{SST} \rangle\big)^2}. \tag{5}$$

The classification of each year in low-SST and high-SST years is performed depending on whether they are lower or greater than $\langle \mathrm{SST} \rangle$.

If one wishes to consult a thorough technical review article on tropical-cyclone as a thermodynamic system, [6] by Emanuel would probably be the best source available.

Tropical cyclones are classified into three main groups, based on wind intensity: tropical depressions, tropical storms, and a third group of more intense storms, whose name depends on the region. In particular, in the Northeast Pacific or in the North Atlantic, it is called a hurricane. In Table 1 we can see a detailed classification of the tropical-cyclones studied in this text.

Category	1–minute sustained winds	
Tropical depression	≤ 33 kn	(≤ 61 km/h)
Tropical storm	34–63 kn	(63–118 km/h)
Category 1 hurricane	64–82 kn	(119–153 km/h)
Category 2 hurricane	83–95 kn	(154–177 km/h)
Category 3 major hurricane	96–112 kn	(178–208 km/h)
Category 4 major hurricane	113–136 kn	(209–253 km/h)
Category 5 major hurricane	≥ 137 kn	(≥ 254 km/h)

Table 1: Tropical-cyclone classification used by the NHC

1.2 Hypothesis and conclusions

The hypothesis is that the SST does not directly affect the maximum wind speed of a tropical-cyclone: storms of equal lifetime should, in theory, have the same wind speed and PDI, and have the same joint distribution:

$$f(Y \mid X = x)_{\text{low}} = f(Y \mid X = x)_{\text{high}}. \tag{6}$$

The physical reasoning behind this is that once the cyclone is activated, the wind speed should not depend on its underlying SST.

Instead of working with the exact joint distributions f, we study the expected value of the distributions:

$$E(Y \mid X = x)_{\text{low}} = E(Y \mid X = x)_{\text{high}}, \tag{7}$$

where $E(Y \mid X = x)$ is estimated by performing a *ordinary least squares* (OLS) regression analysis on the data sets.

The results show a remarkable correlation in the joint distribution of lifetime and PDI. By considering this joint distribution a bivariate lognormal distribution we perform a linear regression analysis of the logarithmic variables. Statistical testing, by means of a permutation test, shows that there is no significant difference between the regression coefficient estimates for low-SST years and high-SST years.

This gives a strong evidence in favour to the fact that this relation does not significantly depend on the SST. In other words, the wind speed of a tropical cyclone of a given lifetime will be the same (within statistical fluctuations) in cold and warm years.

1.3 Outline

In essence, our methodology consists in comparing the low-SST and high-SST hurricane occurrences distributions to discern any statistical difference between both.

Firstly, in § 4.1, for each basin we explore the joint distributions of PDI and storm lifetime and compare low-SST and high-SST years by studying the expected value of the marginal distributions.

For comparing the two populations obtained after separating the hurricane occurrences by low-SST and high-SST years, in § 4.2 we propose a null hypothesis and test statistics, which are evaluated in § 4.3.

To be able to fit a linear regression model using ordinary least squares, some critical assumptions on the residual errors need to hold. These are tested in § 4.4 using diagnostic plots and specific statistical tests designed to test each assumption.

After finding that some of these assumptions do not actually hold, in § 4.5 we perform a resampling of the observations using a bootstrap method to provide a more accurate and robust regression analysis than the one provided by the ordinary least squares method. Then the same test statistics are evaluated using the bootstrapped coefficient estimates.

The only issue with these test statistics, is that it is hard to tell from theory if they are too big to reject the null hypothesis. It is for this reason that in § 4.6 we propose a permutation test to assess the statistical significance of evidence against (or in favour of) the tested hypothesis that there is no difference in the evolution of a tropical-cyclone once it is activated, regardless of the SST.

To answer why the increase of tropical-cyclone lifetime with SST triggers an increase in wind speed as a by-product, as a first approach in § 5 we perform an exploratory analysis of the geographical properties of the tropical-cyclones, such as genesis and death location of the storms, as well as their path length. Our results show that the longer lifetimes for high-SST are mainly due to a shift to South-East of the tropical-cyclones genesis point.

2 Background

2.1 Simple linear regression

Suppose that there is a quantitative *response*, or *dependent variable*, Y and a *predictor*, or *independent variable*, X. It is assumed that there is some kind of *true* relationship between Y and X, which can be written as

$$Y = f(X) + \epsilon, \tag{8}$$

where f is an fixed but unknown function of X, and ϵ is a random *error term* assumed independent and $\mathcal{N}(0, \sigma^2)$, i.e., distributed normally with mean $\mu = 0$ and variance σ^2.

If f is to be approximated by a linear function, then this relationship can be written as

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{9}$$

Here β_0 is the *intercept* term, that is, the expected value of $Y(X = 0)$, and β_1 is the *slope*, the average increase in Y associated with a one-unit increase in X. The error term ϵ takes into account that (i) the true relationship is probably not exactly linear, (ii) there may be other variables that cause variation in Y, and (iii) there may be measurement error.

The simple linear regression helps building, or fitting, a statistical model for predicting, or estimating, the dependent variable Y based on the dependent variable X.

2.2 Estimation of the coefficients

In practice, the coefficients β_0 and β_1, as well as σ^2 are unknown. The most common method to find the coefficients is the *least squares regression* or *ordinary least squares* (OLS).

Suppose the studied data set (X, Y) consists of n observation pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i-th value of X, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the coefficients estimates. Then, $\epsilon_i = y_i - \hat{y}_i$ is the i-th *residual* of the linear model. One can define the *residual sum of squares* (RSS) as:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2. \tag{10}$$

The OLS approach finds the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise the RSS:

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0, \tag{11a}$$

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0. \tag{11b}$$

From equation (11a), we can show that the minimiser intercept is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{12a}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

On the other hand, replacing $\hat{\beta}_0$ by $\bar{y} - \hat{\beta}_1 \bar{x}$ in equation (11b), we find that the minimiser slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{12b}$$

2.3 Accuracy of the coefficients: standard errors

If we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ on the basis of a particular subset of Y, then the estimates will not be exactly equal to β_0 and β_1. But if we could average the estimates obtained over a huge number of data sets from Y, then the average of these estimates would be exactly equal to β_0 and β_1. This means that the coefficient estimates given by (12) are *unbiased*, or in other words, that the estimators do not systematically over- or under-estimate the true value of the coefficients.

The question, then, arises: how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1. In general, to answer this, we need to compute the variances or *standard errors* associated with β_0 and β_1:

$$\widehat{\mathrm{se}}(\hat{\beta}_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}, \tag{13a}$$

$$\widehat{\mathrm{se}}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{13b}$$

In general, σ^2 is not known, but can be estimated from the data. The estimate of σ is known as the *residual standard error* (RSE), and is given by:

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}. \tag{14}$$

Note that the assumption here, and the OLS as a whole, is that the variance σ^2 is homogeneous, or in other, words $\text{Var}(\epsilon_i) = \sigma^2$, $\forall i$. This property is called *homoscedasticity*.

The term $n - 2$ comes from the fact that the determination of the coefficients is subject to the constraints (11a) and (11b). That leaves $n - 2$ degrees of freedom for the values of y_i in order to determine the linear model exactly.

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed and optimal if the errors ϵ_i are normally distributed. They are often approximately normal for other error distributions, but they are not robust to gross non-normality of errors or to outlying response values.

In § 2.5 we will discuss with more detail some techniques to study the normality and homoscedasticity of the random errors.

2.4 Accuracy of the model: R^2 and correlation

The goodness of a fit can be assessed by measuring the *correlation coefficient* r, or in the more general context of *multiple linear regression*, the R^2 statistic. In this section we will introduce the mathematical expressions of both coefficients and show their equivalence when studying the simple linear regression:

$$R^2 = r^2. \tag{15}$$

2.4.1 R^2 statistic

The statistic R^2 is used in the context of multiple linear regression, and particularly the simple linear regression, and is called the *squared multiple correlation coefficient*, or simply *R-squared*.

The R^2 statistic is given by

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \quad 0 \leq R^2 \leq 1, \tag{16}$$

where TSS is the *total sum of squares*:

$$\text{TSS} = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2. \tag{17}$$

TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS − RSS measures the amount of variability in the response that is *explained* by the regression model:

$$\text{ESS} = \text{TSS} - \text{RSS} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2, \tag{18}$$

and R^2 measures the proportion of variability in Y that can be explained using X as the predictor.

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression model. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, the inherent error σ^2 is high, or both.

2.4.2 Correlation coefficient

A correlation coefficient is a symmetric, scale-invariant measure of association between two random variables, X and Y. It ranges from −1 to +1, where the extremes indicate perfect correlation and 0 means no correlation. The sign is negative when large values of one variable are associated with small values of the other and positive if both variables tend to be large or small simultaneously. The most common correlation coefficient is the Pearson correlation.

The Pearson correlation, or Pearson's r, is rooted in the bivariate normal distribution $f(X, Y)$ (described in § 4.1.1) where the theoretical correlation describes the contour ellipses for the density function. If both variables are scaled to have a variance of 1, then a correlation of zero corresponds to circular contours, whereas the ellipses become narrower and finally collapse into a line segment as the correlation approaches ±1, which is why sometimes r is called the *linear correlation*.

The empirical correlation coefficient is given by

$$r = \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}}, \quad -1 \le r \le 1. \tag{19}$$

It can be shown that in the context of simple linear regression, the square of the correlation coefficient r^2 and the R^2 statistic are identical:

Proof.

$$
\begin{aligned}
R^2 &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \\
&= 1 - \frac{\sum_{i=1}^{n} \left(y_i - (\bar{y} + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \\
&= 1 - \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \\
&= \frac{2\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \\
&= \frac{2\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \hat{y}\right) - \left(\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2} \\
&= \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = r^2. \qquad \text{(15 bis)}
\end{aligned}
$$

□

In simple terms, although they have the same mathematical expression, conceptually r^2 (and r) is an indicator of the strength of the correlation between the variables X and Y; while R^2 is an indicator of the goodness of fit, in the sense of explained variability, of the linear model that predicts the value of Y as a function of X as a causal relationship. Importantly, causality in this context means the direction of causality runs from X to Y and not the other way round.

2.5 Regression diagnostics

If (9) holds with homoscedastic random errors ϵ_j and if those random errors are normally distributed, or if the dataset is large, then standard distributional results will be adequate for making inferences with the OLS estimates. This assumption of homoscedasticity is of big importance in linear regression, as the standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon it being true.

For this reason, if the errors are grossly non-normal or *heteroscedastic*, meaning that their variances are unequal, then those standard results may not be reliable and a *resampling method* (see § 2.6) may offer genuine improvement.

To test the normality, independence, and homoscedasticity of the random errors, one

can check a few diagnostic plots, which could reveal unexplained patterns in the data by the fitted model. In particular, we will focus on:

- Q-Q plot of the residuals.
- Residuals vs fitted values plot.

Additionally, these properties or assumptions on the random errors can be individually tested using some well known statistical hypothesis tests:

- Lilliefors test: normality.
- Correlation test: independence.
- Breusch–Pagan test: homocedasticity.

2.5.1 Q-Q plots

The *Q-Q plot*, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal $\mathcal{N}(\mu, \sigma^2)$. For example, if we run a statistical analysis that assumes our dependent variable is normally distributed, we can use a *normal Q-Q plot* to check that assumption. It is just a visual check, so it is somewhat subjective, but it can help identify obvious distributional anomalies in the data.

A Q-Q plot is built by taking the sample data, sorting it in ascending order, and then plotting them versus quantiles calculated from the theoretical distribution. In the case of studying the residuals, we expect the data to follow a straight line in the normal Q-Q plot.

2.5.2 Residuals vs fitted values plots

The *residuals vs fitted values plot*, or simply residual plot, is a useful graphical tool to help us assess the homoscedasticity of the residuals or, more generally, if the residuals have non-linear patterns.

A residual plot is built by taking the residuals ϵ_i and plotting them versus the fitted values $\hat{y}_i$. The residuals should be centred on zero throughout the range of fitted values, and should be more or less uniformly distributed and have a constant spread throughout the range. The reason behind this is that

$$\mathrm{Cov}(\hat{y}_i, \epsilon_i) = 0. \tag{20}$$

One might wonder why do we not just plot the residuals ϵ_i vs the predictors x_i. Actually, one could plot the residuals vs the predictors, instead of vs the fitted values. However, this plot would be impossible to visualise if we had more than two predictors, so it is of standard practice to look at the residuals vs fitted plot, which is always a 2D plot.

Usually, the residual plot is accompanied by a *locally weighted scatterplot smoothing* (LOWESS) line to help visualise the trend of the residuals. Naturally, under the assumption of normality, the trend should follow a horizontal line centred at $\epsilon = 0$. It is also useful to show the outer quantiles of the residuals to help emphasise and distinguish patterns.

2.5.3 Lilliefors test

The *Lilliefors test*, developed by Lilliefors (1967) [9], is a normality test based on the Kolmogorov–Smirnov (KS) test. It is used to test the null hypothesis that data comes from a normally distributed population when the null hypothesis does not specify which normal distribution; in other words, it does not specify the expected value μ and variance σ^2 of the distribution.

Like most statistical tests, this test of normality defines a criterion and gives its sampling distribution (in this case, a Kolmogorov–Smirnov distribution). When the probability, or p-value, associated with the criterion is smaller than a given α significance level, the alternative hypothesis is accepted (i.e., we conclude that the sample does not come from a normal distribution).

A thing to consider with this test is that with small samples, the KS test is underpowered and fails to detect true violations of normality; and with large samples, the KS test may detect violations of normality which are not important for practical purposes [10]. For this reason it is important when assessing normality to also look at indicators of the degree of normality such as the Q-Q plot stated previously.

In R, the `nortest` package offers the `lillie.test()` function to perform the Lilliefors test.

2.5.4 Correlation test

The *Correlation test* is a test based on either Pearson's product-moment correlation, Kendall's rank correlation tau, or Spearman's rank correlation rho. It is used to test the null hypothesis that the correlation between paired samples (in this case, ϵ and $\hat{Y}$) is zero.

The three methods use different measures of correlation, all in the range $[-1, 1]$ with 0 indicating no correlation. The most used method is Pearson's product-moment correlation, in which the test statistic is based on Pearson's product moment correlation coefficient $r = \mathrm{Cor}(\epsilon, \hat{Y})$ and follows a Student's t-distribution on $n - 2$ degrees of freedom.

In R, the `stats` package offers the `cor.test()` function to perform the Lilliefors test.

2.5.5 Breusch–Pagan test

The *Breusch–Pagan test*, developed by Breusch and Pagan (1979) [11], is a test used in regression analysis to test for homoscedasticity. It is used to test the null hypothesis that residuals are homoscedastic.

The null hypothesis is that the residual variances are all equal, while the alternative hypothesis is that the residual variances are a multiplicative function of one or more variables. To do this, the test fits a linear regression model to the residuals of a linear regression model (by default the same explanatory variables are taken as in the main regression model) and rejects if too much of the variance is explained by the additional constructed explanatory variables.

In R, the `lmtest` package offers the `bptest()` function to perform the Breusch–Pagan test.

Example 2.1. Let us consider the three different simulated models depicted in Figure 1. For each of the models, the random error has been built under a different assumption: (a) homoscedasticity, (b) heteroscedasticity, and (c) non-normality.

For a moment let us forget we know the models behind the data and let us see what we can infer purely from the data, which is generally the case when dealing with real data.

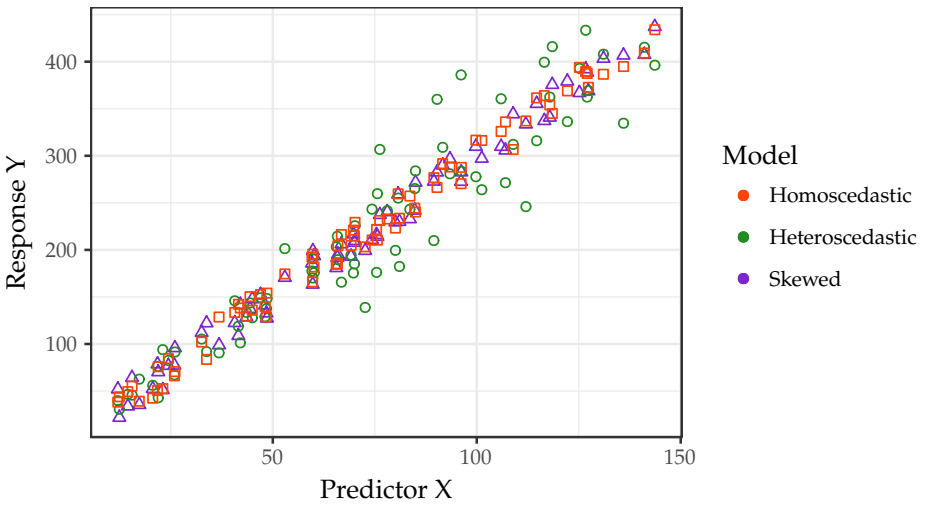
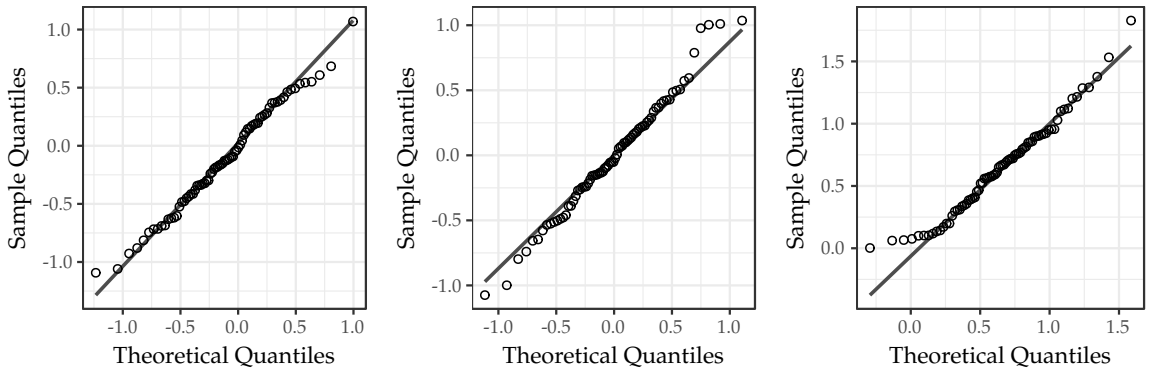


Figure 1: Three simulated models of 75 observations each with different response Y in which it would seem appropriate to fit a linear model using OLS

Looking at the Q-Q plots, at first glance it seems that both Figure 2a and Figure 2b follow a normal distribution for the residuals. A priori, what we can say is that the tails in Figure 2a are slightly lighter than what we would expect under the standard modelling assumptions, while the tails in Figure 2b are heavier.

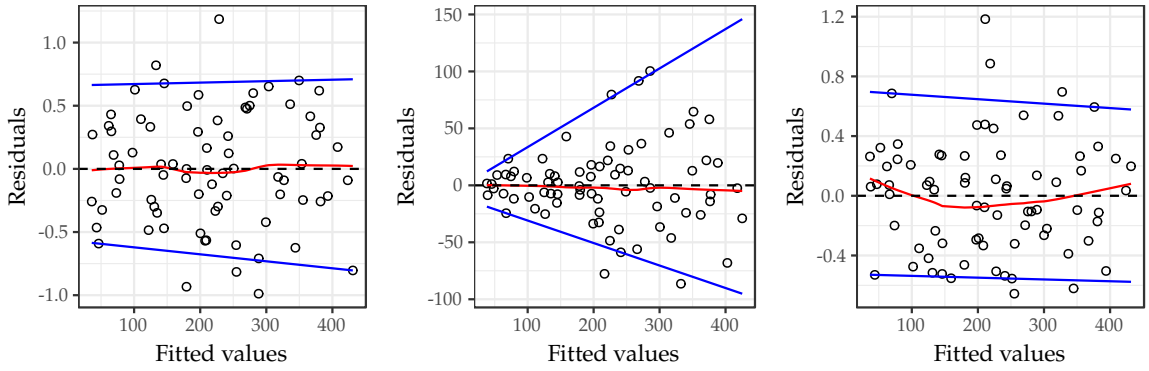
As a contrast, the Q-Q plot gives Figure 2a immediately tells us that we are under a clear case of non-normality and that the fitted model should be probably reconsidered.



(a) Homoscedastic model (b) Heteroscedastic model (c) Skewed model

Figure 2: Normal Q-Q plots for the three models simulated in Figure 1

Let us see if a residual plot can provide additional insight to the underlying models depicted in Figure 1 that could not be observed in the Q-Q plots in Figure 2.



(a) Homoscedastic model (b) Heteroscedastic model (c) Skewed model

Figure 3: Residual plots for the three models simulated in Figure 1, the red line is a smooth fit to the residuals, and the blue lines are the outer quantiles of the residuals

In Figure 3a and Figure 3c we can see that the residuals appear to be roughly randomly distributed. However, for Figure 3c, it is clear that the residuals are skewed, as the point pattern is densest somewhere other than the centre line; this tells us that the residuals do not follow a normal distribution at all. But we already knew this from Figure 2c.

In Figure 3b we can see quite an extreme case of heteroscedasticity, as the magnitude of the residuals tends to increase with the fitted values, giving place to a prominent

funnel shape. This tells us that there is some kind of correlation between the residuals and the predictors: $\epsilon_i \sim \hat{y}_i = f(x_i)$.

As a last step, let us explore the results of performing the Lilliefors, correlation, and Breusch–Pagan tests in Table 2.

Model	Lilliefors	Correlation	Breusch–Pagan
(a) Homoscedastic	0.649	1.000	0.312
(b) Heteroscedastic	0.054	1.000	0.004
(c) Skewed	0.723	1.000	0.830

Table 2: List of p-values associated with the statistical hypothesis tests to respectively analyse normality, independence, and homoscedasticity of the residuals on the three simulated models

From the results in Table 2 we can see that the Lilliefors fails to reject the non-normality of the heteroscedastic and skewed models, when in fact only the homoscedastic model is simulated following a normal distribution, and that there is no correlation between the residuals and the fitted values for any of the models. More importantly, though, the Breusch–Pagan test successfully rejects homoscedasticity in the grossly heteroscedastic model. ▲

To sum up, for linear regression with normal random errors having constant variance, the OLS theory of regression estimation provides clean, exact methods for analysis. But for generalisations to non-normal errors and non-constant variance, exact methods rarely exist, and we are faced with approximate methods based on linear approximations to estimators and central limit theorems. In the next section we will explore how resampling methods have the potential to provide a more accurate analysis.

2.6 Resampling methods

Resampling methods are an indispensable tool in modern statistics. They involve repeatedly drawing samples from a data set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. The term *resampling* is used for any variety of methods for doing one of the following:

- Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping).
- Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomisation tests, or re-randomisation tests).
- Validating models by using random subsets (bootstrapping, cross validation).

Resampling approaches can be computationally expensive, because they involve fitting the same statistical method multiple times using different subsets of the studied data set. However, due to recent advances in computing power, the computational requirements of resampling methods generally are not prohibitive.

2.6.1 The bootstrap in linear regression

The bootstrap is a method to derive properties (standard errors, confidence intervals and critical values) of the sampling distribution of estimators.

In the case of linear regression, when the assumptions on the residual error do not hold, there two quite different resampling methods: (a) resampling of the errors, and (b) resampling of the cases; being the second one more robust to failure of the model assumptions [12].

For resampling the cases (or observations), one assumes the data is the sample from some bivariate distribution $f(X, Y)$. This will sometimes, but not often, mimic reality. Model (9) still applies, but with no assumption on the random errors ϵ_i other than independence.

With f being the bivariate distribution of (X, Y), it is appropriate to take f to be the empirical distribution function (EDF) of the data pairs, and resampling will be from this EDF. The resampling simulation therefore involves sampling pairs with replacement from $(x_i, y_i), \ldots, (x_n, y_n)$. This is equivalent to taking

$$(x_i^*, y_i^*) = (x_I, y_I), \tag{21}$$

where I is uniformly distributed on $\{1, 2, \ldots, n\}$. Then, simulated values $\hat{\beta}_0^*$, $\hat{\beta}_1^*$ the coefficient estimates are computed from $(x_i^*, y_i^*), \ldots, (x_n^*, y_n^*)$ using the OLS method which was applied to obtain the original estimates. This resampling algorithm can be

summarised in Algorithm 1.

Algorithm 1: Resampling the cases using bootstrap

for $r \leftarrow 1$ ***to*** R **do**
  (i) Sample $i_1^*, \ldots, i_n^*$ randomly with replacement from $\{1, 2, \ldots, n\}$.
  for $j \leftarrow 1$ ***to*** n **do**
    (ii) Set $x_j^* = x_{i_j^*}$, $y_j^* = y_{i_j^*}$
  (iii) Fit OLS regression to $(x_i^*, y_i^*), \ldots, (x_n^*, y_n^*)$.
  (iv) Calculate estimates of $\hat{\beta}_{0,r}^*$ and $\hat{\beta}_{1,r}^*$.

A great advantage of bootstrap is its simplicity. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as percentile points, proportions, and correlation coefficients. Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality [13].

2.6.2 Permutation tests for comparing two populations

Permutation tests or randomisation tests are widely used in nonparametric statistics where a parametric form of the underlying distribution is not specified (or known).

Permutation tests for comparing two populations can be widely used in practice because of flexibility of the test statistic and minimal assumptions [15]. Consider sample of m observations from population A and n observations from population B. Assume that under the null hypothesis H_0 there is no difference between a certain property of both populations; this could be the mean, the median, etc. Then any permutation of the observations between the two populations has the same chance to occur as any other permutation.

One could also perform parametric tests to test H_0, but they require the assumptions about the distribution of the characteristic in the population to be fulfilled. Permutation tests do not need fulfil the assumption about conformity with normal distribution and are as robust as parametric tests [16].

The essence of permutation tests is to determine a test statistic and then to evaluate the sample distribution of this statistic for all permutations of the populations. When calculations affect large number of permutations, a Monte Carlo method is applied (i.e., the permutations are random).

The steps for performing this kind of permutation test can be summarised in Algorithm 2.

Algorithm 2: Permutation test for comparing two populations

(i) Define the null hypothesis, H_0, and the alternative.
(ii) Consider a test statistic that compares the populations which is large (small) if the null hypothesis is not true, and small (large) if it is true.
(iii) Calculate the true statistic of the data, T.
for $r \leftarrow 1$ ***to*** R **do**
    (iv) Create a new data set consisting of the data, randomly rearranged. Exactly how it is rearranged depends on the null hypothesis.
    (v) Calculate the statistic for this new data set, T^*.
    (vi) Compare the statistic T^* to the true value, T.
(vii) If the true statistic is greater (lower) than 95% of the random values, then one can reject the null hypothesis at $p < 0.05$.

A great advantage of the permutation tests exist for any test statistic, regardless of whether or not its distribution is known. Thus one is always free to choose the statistic which best discriminates between hypothesis and alternative and which minimises losses.

3 Data

3.1 Hurricane tracks

3.1.1 Description of the database

Although Corral et al. analyse several ocean basins, we focus only on the North Atlantic (N. Atl.) and the Northeast Pacific (E. Pac.) Oceans. The reason to do this is the abundance of research on these two basins and the precision of the database provided by the National Hurricane Center (NHC) [17]: the HURDAT [18].

Since both basins directly concern USA territories (especially the N. Atl.), the government's efforts on improving the tracking and prediction technologies, routine satellite imagery has been used since as early as the late 1960s.

A major change between our data sets and the ones used by Corral et al. [2] is that the second-generation hurricane database (HURDAT2), has been developed this decade [19]. The improvements of the revised version are mainly:

(i) Inclusion of non-developing tropical depressions.

(ii) Inclusion of systems that were added to the database after the end of each season.

Also, the ongoing post-storm analysis reviews of the tropical-cyclones have revised several storms [20], particularly important in the 1851–1960 era.

Recently, in June 2018, Delgado et al. [21] have revised and updated the HURDAT2 data set. This revision includes storms from 2017, as well as a revision of the 1954–63 era for the North Atlantic data. Nonetheless, the analysis is done using the 2016 data sets, as it would require a lot of effort to clean the new data and make sure no major change has been done introduced into historical data.

These raw data sets used can be downloaded from `http://www.aoml.noaa.gov/hrd/hurdat/hurdat2-1851-2016-apr2017.txt` (N. Atl.) and `http://www.aoml.noaa.gov/hrd/hurdat/hurdat2-nepac-1949-2016-apr2017.txt` (E. Pac.).

3.1.2 Data structure

The format of the HURDAT2 data sets is documented at [22, 23]. A record of data is recorded once every 6 hours for each storm (although there are additional records for certain storms, specially those marking the landfall of a storm). The record is comprised of the date and time, storm identifier, system status (cf. tropical-cyclone category), latitude and longitude of the centre of the storm, the sustained surface wind speed (in knots) observed in the storm, and several other properties that are not relevant in this study.

In Table 3 one can see the structure of the cleaned data illustrate the variables we use

in the study directly available in the raw data sets, as well as the format (data type[1]) of the observational record data.

storm.id <chr>	storm.name <chr>	n.obs <int>	date.time <dttm>	status <fctr>	lat <dbl>	long <dbl>	wind <dbl>	storm.year <dbl>
AL011851	UNNAMED	13	1851-06-25 00:00:00	Hurricane	28.0	-94.8	80	1851
AL011851	UNNAMED	13	1851-06-25 06:00:00	Hurricane	28.0	-95.4	80	1851
AL011851	UNNAMED	13	1851-06-25 12:00:00	Hurricane	28.0	-96.0	80	1851
AL011851	UNNAMED	13	1851-06-25 18:00:00	Hurricane	28.1	-96.5	80	1851
AL011851	UNNAMED	13	1851-06-26 00:00:00	Hurricane	28.2	-97.0	70	1851
AL011851	UNNAMED	13	1851-06-26 06:00:00	Tropical storm	28.3	-97.6	60	1851

Table 3: Excerpt of the North Atlantic data set

A map displaying the studied hurricane tracks for the North Atlantic and Northeast Pacific basins can be seen in Figure 6.

3.2 Sea surface temperature

3.2.1 Description of the database

There are several sea surface temperature (SST) databases, with different time-steps (e.g., daily, weekly, monthly, and so on), domains (i.e., global or specific regions), and data resolutions; each used for different analyses of climatological nature [24, 25].

The Met Office [26] Hadley Centre's sea ice and sea surface temperature database, HadISST1 [27], is a unique combination of monthly globally complete fields of SST and sea ice concentration on a latitude-longitude grid from 1871. In Figure 4 we can see a sample from the data set to illustrate the grid structure.

Although there is a revised HadISST.2 database [28], we use the HadISST1 database, as it is the one used in Corral et al.'s, Webster et al.'s and several other authors's climate analyses.

The main reason to do so, however, is that HadISST.2 contains more ocean grid boxes and introduces a different method to calculate the monthly temperatures, making it quite incompatible with HadISST1 (as opposed to the revised HURDAT2 database, which is just an improved version of the old database, without introducing changes in the methodology of analysis or in the structure of the data).

[1] In computer science and computer programming, a data type or simply type is a classification of data which tells the compiler or interpreter how the programmer intends to use the data.

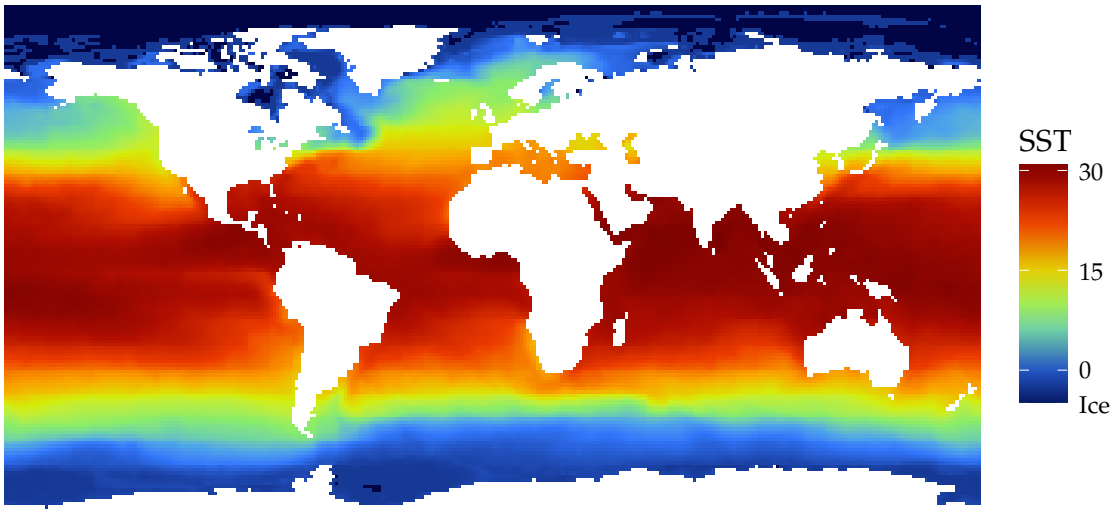


Figure 4: Global SST (in °C) map from December 2015

3.2.2 Data structure

The format of the HadISST1 database is documented at [29]. The data are available in netCDF format, which is constructed using raster data.

A raster brick consists of a matrix of cells (or pixels) organised into a grid where each cell contains a value representing information, such as temperature in our case. Each matrix can also be comprised of layers (as illustrated in Figure 5); in the HadISST1 database, each matrix layer represents a different month.

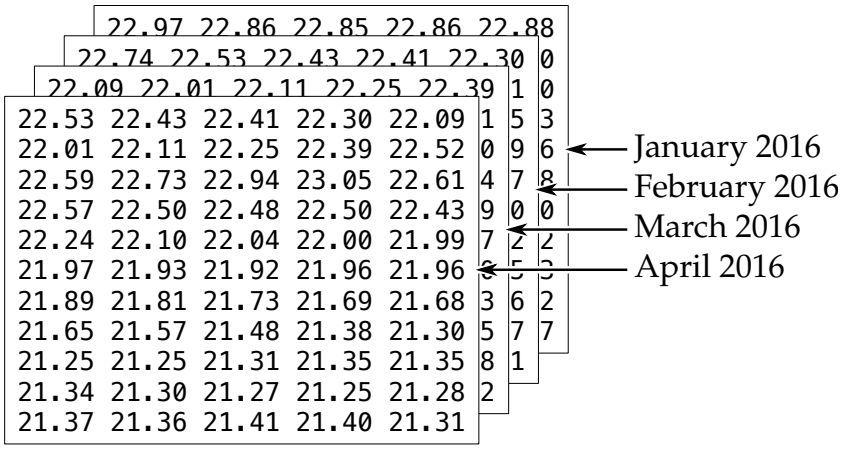


Figure 5: Simple diagram of the internal structure of a raster brick

The data set used can be downloaded from `http://www.metoffice.gov.uk/hadobs/hadisst/data/HadISST_sst.nc.gz`.

In Table 4 one can see the structure of the result from the SST classification of the years into low-SST and high-SST to illustrate the variables we use, as well as the data type of the SST data. Note that the classification is unique to each of the basins.

year	sst	sst.norm	sst.class
<date>	<dbl>	<dbl>	<chr>
1966-01-01	27.47	0.9979934	low
1967-01-01	27.19	0.9879054	low
1968-01-01	27.34	0.9933687	low
1969-01-01	27.75	1.0083072	high
1970-01-01	27.36	0.9940200	low
1971-01-01	27.04	0.9825272	low

Table 4: Excerpt of the results from the SST analysis for the North Atlantic basin

3.3 Activity windows

Even though recent improvements have been made to the HURDAT2 database, [20, 22, 23], following the methodology of Corral et al., we intentionally limit this study to the satellite era.

In [1], Webster et al. go into more details about the activity windows for the hurricane tracks data as well as the sea surface temperature used by researchers in the past. In Table 5 we can see a summary of the spatial and temporal activity windows we use for each basin based on the information available in the previously mentioned papers; we also include the amount of analysed tropical-cyclones N, the amount of occurrences in low-SST years N_{low}, the amount of occurrences in high-SST years N_{high}, as well as the size of the entire data set N_{tot}.

Basin	Years	Season	Longitude	Latitude	N	N_{low}	N_{high}	N_{tot}
N. Atl.	1966–2016	June–October	90°W–20°W	5°N–25°N	771	365	406	1756
E. Pac.	1986–2016	June–October	120°W–90°W	5°N–20°N	594	238	356	1071

Table 5: Spatial and temporal activity windows for each basin

In Figure 6 we can see a map showing all the storms analysed for both basins (N. Atl. and E. Pac.), already divided by SST class, and the spatial window for the $\langle \text{SST} \rangle$ calculation highlighted in green.

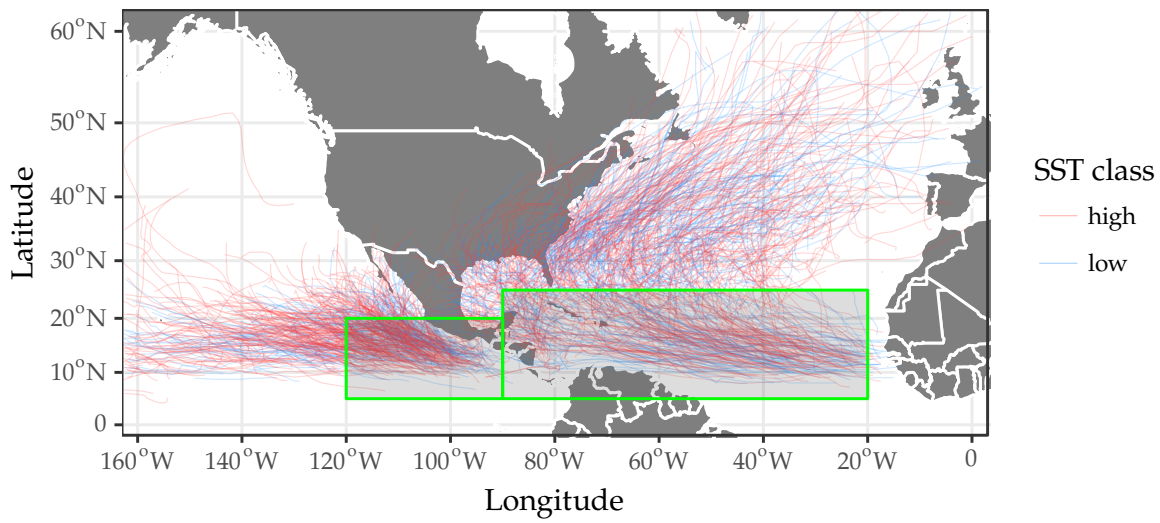


Figure 6: Tropical-cyclones best tracks for the North Atlantic and Northeast Pacific Oceans

3.4 Unified data set

For this study, we have developed a unified data set that summarises the relevant variables of each analysed tropical-cyclone using data from the HURDAT2 and the HadISST1.

In Table 6 one can see the structure of the unified data set to illustrate the variables we use, as well as their data type.

storm.id <chr>	storm.name <chr>	n.obs <int>	storm.duration <dbl>	storm.pdi <dbl>	max.wind <int>	mean.wind <dbl>	mean.sq.wind <dbl>	storm.year <int>	basin <chr>	sst <dbl>	sst.norm <dbl>	sst.class <chr>
AL011966	ALMA	42	252	34632626747	110	56.4	3750	1966	NATL	27.6	0.998	low
AL021966	BECKY	9	54	3413930334	65	46.1	2353.	1966	NATL	27.6	0.998	low
AL031966	CELIA	36	216	7839872104	70	35.4	1488.	1966	NATL	27.6	0.998	low
AL041966	DOROTHY	37	222	21340832518	75	54.9	3211.	1966	NATL	27.6	0.998	low
AL051966	ELLA	26	156	4646503652	45	37.7	1487.	1966	NATL	27.6	0.998	low
AL061966	FAITH	69	414	120569417711	110	79.0	6722.	1966	NATL	27.6	0.998	low

Table 6: Excerpt of the North Atlantic data set

This unified data sets for the North Atlantic and Northeast Pacific basins can be downloaded from the GitLab repository of this project [30] in `CSV` format.

Alternatively, these data sets have been packaged into an R package called `HurdatHadISSTData` [31], and can be installed using the `devtools` package:

```
1 library(devtools)
2 install_git("https://gitlab.com/aldomann/hurdat-hadisst-data.git")
```

The names of these data sets in the `HurdatHadISSTData` package are:

- `tc.pdi.natl` – Data set for the North Atlantic basin.

- `tc.pdi.epac` – Data set for the Northeast Pacific basin.
- `tc.pdi.all` – Data set for both basins.

3.5 On non-developing systems

Tropical-cyclones that surpass the 33 kn wind speed threshold are called *developing systems*, while the ones that do not do so are called *non-developing systems*. In terms of hydrodynamics, the major difference between developing and non-developing systems is that the developing have a distinct area of low height or low pressure centred on the system, i.e., they form cyclones [32].

This means that even though all tropical-cyclones behave thermodynamically in the same way throughout their lifetime (the PDI calculation is valid for all their life span), the wind speeds evolve rather differently depending on the development status of the storm.

Apart from this physical description, from a statistical point of view, one can see that non-developing systems constitute a *nonstationary* time series for the North Atlantic data (Figure 7 and Table 7); this means the distribution of the storms alters with alterations in time.

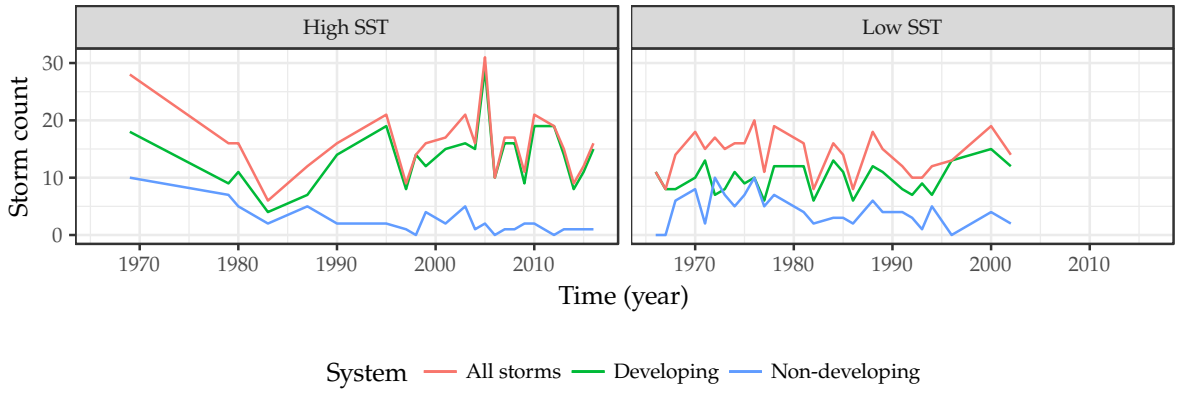


Figure 7: Time series of storm occurrences for the North Atlantic basin, emphasising into non-developing systems and developing systems

Subset	Non-developing	Developing	All storms
All storms	0.0205	≥ 0.1	≥ 0.1
Low SST	0.0233	≥ 0.1	≥ 0.1
High SST	0.0429	≥ 0.1	≥ 0.1

Table 7: List of p-values associated with the Kwiatkowski–Phillips–Schmidt–Shin test to analyse stationarity of the storm occurrences for the North Atlantic basin

For the Northeast Pacific data (Figure 8 and Table 8), stationarity is rejected only for the time series without separation of the storms by SST class.

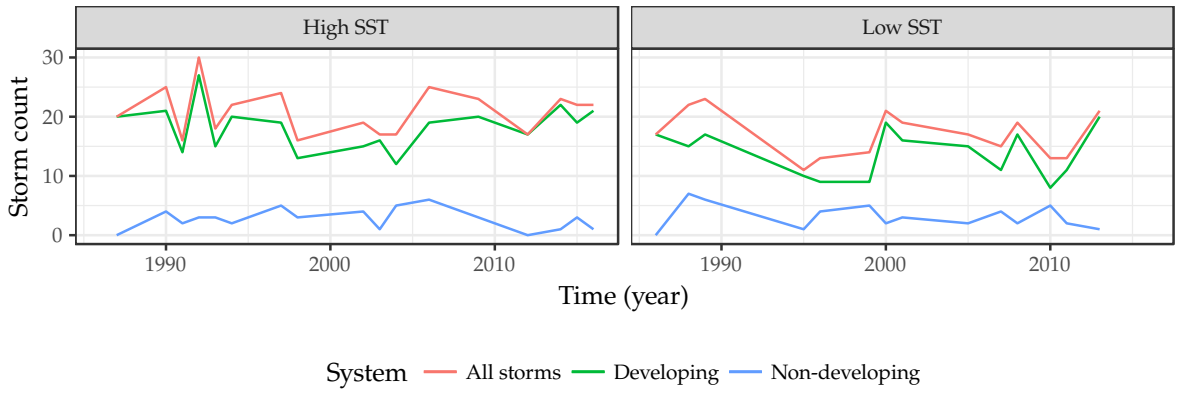


Figure 8: Time series of storm occurrences for the Northeast Pacific basin, emphasising into non-developing systems and developing systems

Subset	Non-developing	Developing	All storms
All storms	0.0294	≥ 0.1	≥ 0.1
Low SST	0.0732	≥ 0.1	≥ 0.1
High SST	0.0733	≥ 0.1	≥ 0.1

Table 8: List of p-values associated with the Kwiatkowski–Phillips–Schmidt–Shin test to analyse stationarity of the storm occurrences for the Northeast Pacific basin

The issue with this is that when nonstationary time series are used in a regression model one may obtain apparently significant relationships from unrelated variables. This phenomenon is called *spurious regression* [33]. For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. And if the mean and variance of a series are not well-defined, then neither are its correlations with other variables.

One should be cautious, therefore, about trying to extrapolate regression models fitted to nonstationary data.

Although we do not perform a direct regression analysis on the time series of storm occurrences, our data is temporally distributed following this time series. Thus, to be on the cautious side, as they represent different physical systems and may introduce problems in the regression analysis, for our study, we opt to exclude non-developing systems altogether.

One could argue that this is not necessary for the Northeast Pacific basin, but to be consistent in the methodology applied to both basins, we exclude non-developing systems as well.

4 Regression analysis using resampling methods

4.1 Distributional properties of the data

It is important to notice that the relationships between a storm's PDI and its lifetime are of non-linear nature. This naturally means that our regressions need to follow a so-called log–log model:

$$\log \Psi = \beta_0 + \beta_1 \log \Phi + \epsilon, \tag{9 bis}$$

where $\log \Psi \equiv Y$ and $\log \Phi \equiv X$.

We do not know exactly how the PDI and the lifetime of the storms are exactly correlated; we just suspect there is a correlation. For this reason, we not only compute the $Y(X)$ fit for each data set, but also $X(Y) = \hat{\beta}_0 + \hat{\beta}_1 Y$, which is calculated as stated in § 2.2 just interchanging the role of X and Y.

Before performing the regression analysis we should study, however, the bivariate distribution of the data and the marginal distributions of the PDI and the lifetime of storms to see any difference between the two populations that are result of the separation of the storms by SST class.

4.1.1 Bivariate lognormal distribution

The bivariate lognormal has two variables, X_1 and X_2, that are jointly related, and has five parameters, μ_1, μ_2, σ_1^2, σ_2^2, and r:

$$f(X_1, X_2) \sim \mathcal{BVLN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, r). \tag{22}$$

The marginal distributions are lognormally distributed, i.e., the logarithm of them is normally distributed:

$$\log X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad \log X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2), \tag{23}$$

and when the value of one of the variables is known, the distribution on the other is also normally distributed. This, naturally implies that the joint distribution of the logarithm of the variables follows a bivariate normal distribution:

$$f(\log X_1, \log X_2) \sim \mathcal{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, r). \tag{24}$$

Figure 9 shows a bivariate normal distribution to illustrate both the joint distribution between the two variables X_1 and X_2 and their respective marginal distributions.

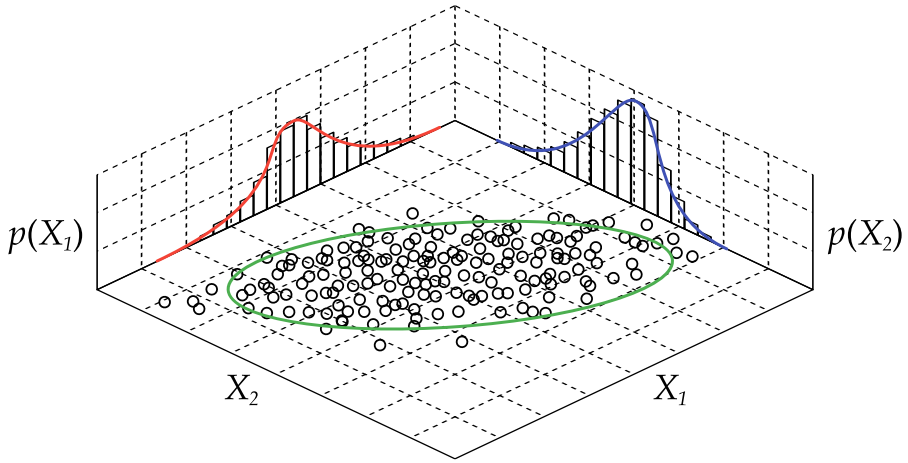


Figure 9: Bivariate normal distribution $f(X_1, X_2)$ and the marginal distributions of X_1 and X_2

In Figure 10 and Figure 11 we can see the bivariate lognormal distributions of the PDI and lifetime of the storms for the North Atlantic and Northeast Pacific basins separating storms by SST class.

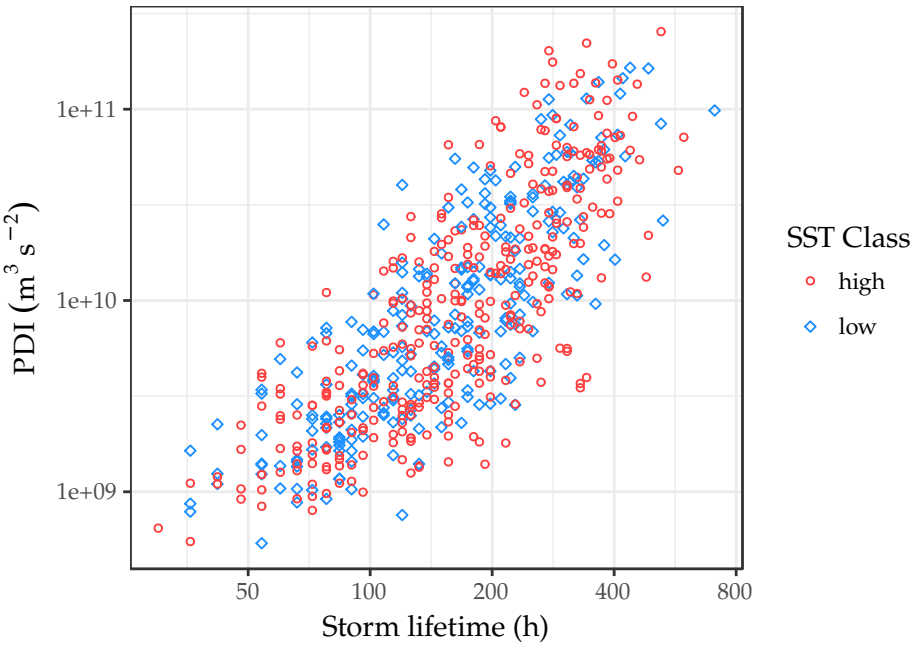


Figure 10: Bivariate lognormal distribution $f(PDI, \text{lifetime})$ of the hurricane observations for the North Atlantic basin

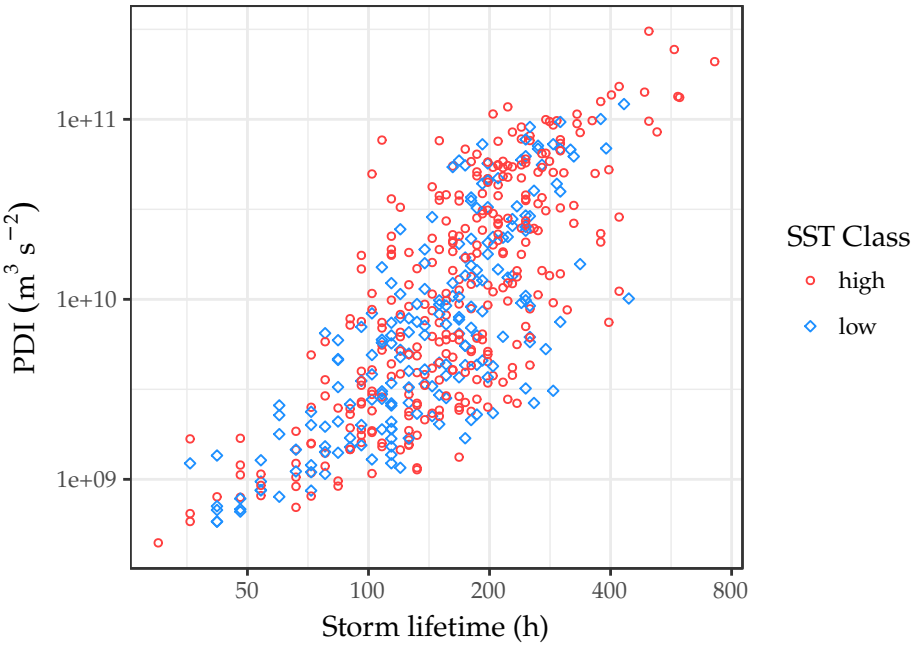


Figure 11: Bivariate lognormal distribution $f(PDI, \text{lifetime})$ of the hurricane observations for the Northeast Pacific basin

The first thing to notice is that, as our hypothesis predicts, although the joint distributions seem to overlap, the expected value $(\mu_{PDI}, \mu_{\text{lifetime}})$ seems to be different for the distributions associated to each SST class. To see this more clearly, we should perform a descriptive univariate analysis of the marginal distributions.

4.1.2 Descriptive univariate analysis of the marginals

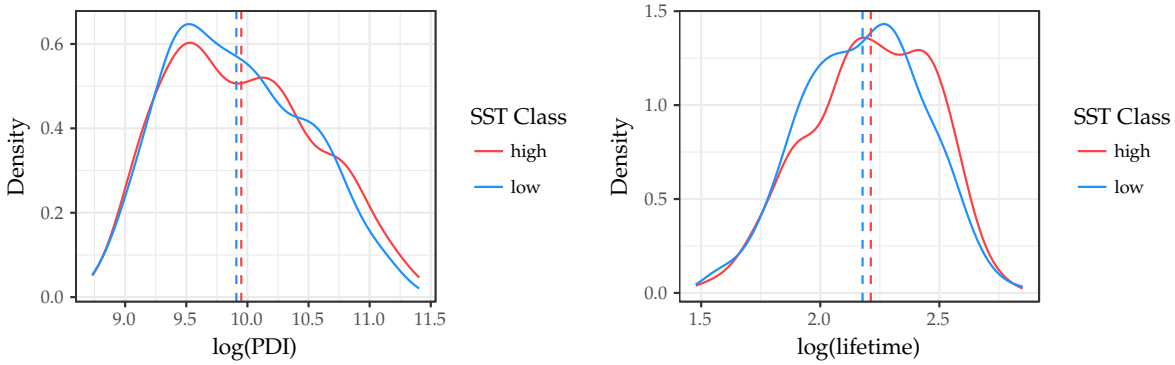
In Figure 12a and Figure 12b we can see the marginal distributions of the PDI and lifetime, in logarithmic scale, for the North Atlantic basin data, separating the data by SST class. For the marginal analysis we also show the expected value μ as a dashed line. Similarly, in Figure 13a and Figure 13b we show the same marginal distributions for the Northeast Pacific basin data.

A statistical summary of the marginals for both basins can be seen in Table 9 and Table 10.

We can see that the marginals do not follow exactly a normal distribution for the PDI. This is expected, as Corral et al. [2] show that the PDI distributions can be characterised by a power-law decay in their central regions. The lifetime marginal distributions do, as it is expected from a bivariate lognormal distribution, roughly follow a normal distribution.

The results show that the expected value $(\mu_{PDI}, \mu_{\text{lifetime}})$ of the joint distributions are displaced to the right-upper corner for high-SST years. This can be clearly seen in the

fact that high-SST years have a longer right tail on account of having more available energy from the sea, and as a result displace the mean of the population μ to higher values as well. This is both reflected in the PDI and the storm lifetime.

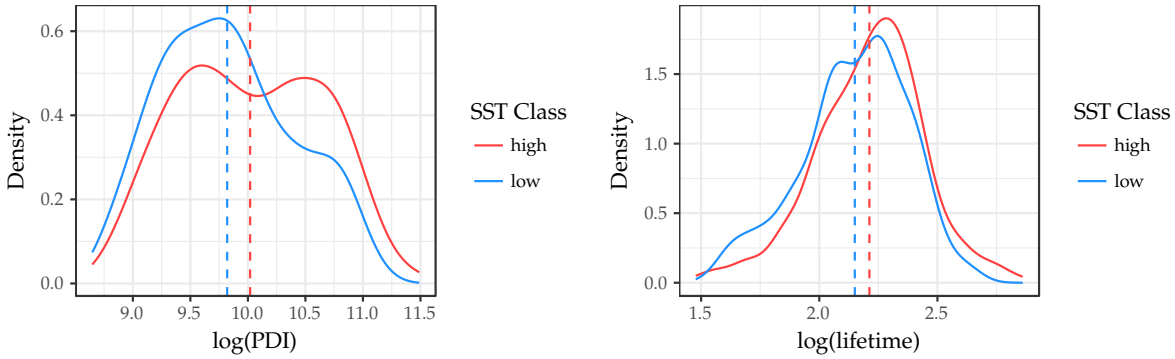


(a) Marginal distribution of the PDI (b) Marginal distribution of the lifetime

Figure 12: Marginal analysis for the variables of the bivariate lognormal distribution for the North Atlantic basin data

Marginal variable	Data	Mean μ	Median
log(PDI)	Low-SST	9.91 ± 0.04	9.86 ± 0.04
	High-SST	9.95 ± 0.03	9.91 ± 0.04
log(lifetime)	Low-SST	2.18 ± 0.02	2.19 ± 0.02
	High-SST	2.21 ± 0.01	2.23 ± 0.02

Table 9: Statistical summary for the low-SST and high-SST subsets of the marginals of the bivariate lognormal distribution for the North Atlantic basin data



(a) Marginal distribution of the PDI (b) Marginal distribution of the lifetime

Figure 13: Marginal analysis for the variables of the bivariate lognormal distribution for the Northeast Pacific basin data

Marginal variable	Data	Mean μ	Median
$\log(PDI)$	Low-SST	9.82 ± 0.04	9.78 ± 0.05
	High-SST	10.00 ± 0.04	9.99 ± 0.04
log(lifetime)	Low-SST	2.15 ± 0.02	2.18 ± 0.02
	High-SST	2.21 ± 0.01	2.23 ± 0.02

Table 10: Statistical summary for the low-SST and high-SST subsets of the marginals of the bivariate lognormal distribution for the Northeast Pacific basin data

4.2 Test statistics to compare the two populations

The marginal analysis of the data performed in § 4.1.2 suggests that there are slight differences in the PDI and lifetime marginal distributions for the populations associated to the different SST classes, and thus their joint distribution.

However, theory suggests that there is no difference in the evolution of a tropical-cyclone once it is activated. Therefore, we should expect that

$$f(Y \mid X = x)_{\text{low}} = f(Y \mid X = x)_{\text{high}}. \tag{6 bis}$$

To analyse this, we propose following null hypothesis:

$$H_0 : \hat{\beta}_{0,h} = \hat{\beta}_{0,l} \wedge \hat{\beta}_{1,h} = \hat{\beta}_{1,l}. \tag{25}$$

The simplest way to test this is to build and calculate statistics that compare both populations that are near to zero if H_0 is true. There are many possible test statistics that could help us test the null hypothesis, but assuming both populations follow the same trend under a linear regression approach, it seems straightforward to compare the coefficient estimates directly:

$$T^{(1)} = \left|\hat{\beta}_{0,h} - \hat{\beta}_{0,l}\right|, \quad T^{(2)} = \left|\hat{\beta}_{1,h} - \hat{\beta}_{1,l}\right|, \quad T^{(3)} = \left|R_h^2 - R_l^2\right|. \tag{26}$$

Polko-Zając [16] propose alternative statistics that not only consider the nominal value of the coefficient estimates, but take into account their standard errors as well:

$$T^{(4)} = \frac{\left|\hat{\beta}_{0,h} - \hat{\beta}_{0,l}\right|}{\widehat{\text{se}}(\hat{\beta}_{0,h} - \hat{\beta}_{0,l})}, \quad T^{(5)} = \frac{\left|\hat{\beta}_{1,h} - \hat{\beta}_{1,l}\right|}{\widehat{\text{se}}(\hat{\beta}_{1,h} - \hat{\beta}_{1,l})}, \quad T^{(6)} = T^{(4)} + T^{(5)}. \tag{27}$$

4.3 Analysis using ordinary least squares

Before we can calculate the test statistics $T^{(i)}$, we need to fit a linear regression model on the data using the OLS method.

In Figure 14 we can see the regression models that fit the joint bivariate lognormal distribution of PDI and lifetime of storms of the populations associated to the different SST classes for the North Atlantic basin. The coefficient estimates obtained for each of the four resulting regression models are shown in Table 11.

With the obtained coefficient estimates we calculate the test statistics $T^{(i)}$associated to the model PDI(lifetime) and the inverse lifetime(PDI) to compare occurrences in low-SST years and in high-SST years for the North Atlantic basin. These are shown in Table 12.

We can see that the regressions for low-SST years and high-SST years are statistically compatible (Table 11), although for the PDI(lifetime) regression the coefficients seem to be much closer, as can be clearly seen in Figure 14.

The results shown in Table 12 might be surprising at first, especially the value of $T^{(1)}$ for the lifetime(PDI) regression, as it is quite big. This is actually an expected result that derives from the relative position of the joint distributions for the two SST classes.

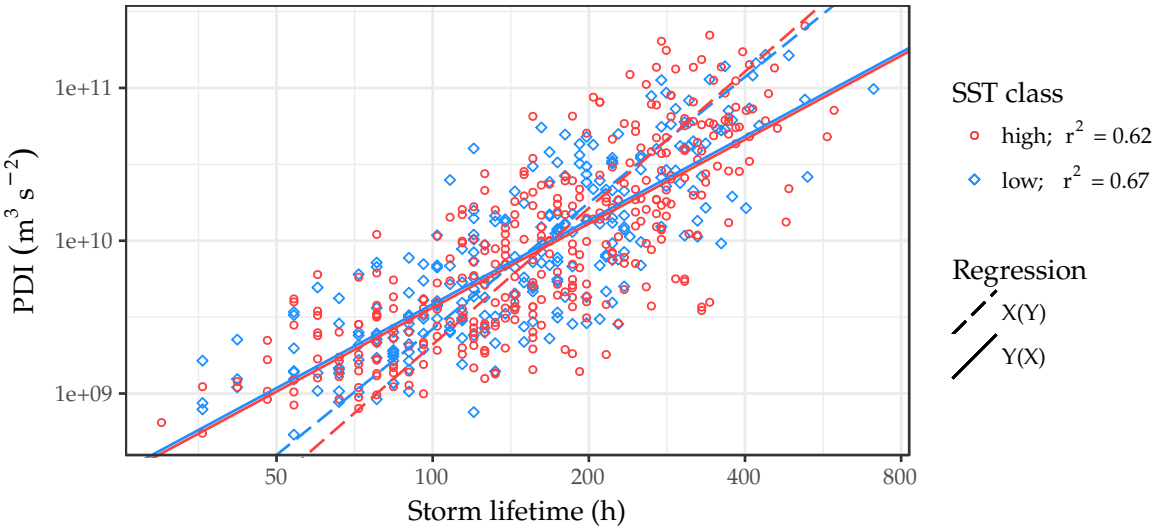


Figure 14: Scatterplot of the joint distribution and regression analysis for the PDI and lifetime of storms for the North Atlantic basin

X	Y	SST class	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
lifetime	PDI	Low	5.94 ± 0.18	1.82 ± 0.08	0.67
		High	5.91 ± 0.17	1.83 ± 0.08	0.62
PDI	lifetime	Low	-1.44 ± 0.16	0.37 ± 0.02	0.67
		High	-1.14 ± 0.14	0.34 ± 0.01	0.62

Table 11: Linear regression coefficients obtained performing OLS on the North Atlantic basin data

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.025	0.001	0.051	0.101	0.010	0.111
PDI	lifetime	0.299	0.028	0.051	1.388	1.295	2.683

Table 12: Value of the test statistics for North Atlantic basin data set using OLS

For the Northeast Pacific we have a similar situation. In Figure 15 we can see the regression models that fit the joint bivariate lognormal distributions; the coefficient

estimates obtained for each of these four resulting regression models are shown in Table 13. The values of the test statistics for this basin are shown in Table 14.

We can see that the regressions for low-SST years and high-SST years are statistically compatible as well (Table 13). As it also happens for the North Atlantic, for the PDI(lifetime) regression the coefficients seem to be much closer, as can be clearly seen in Figure 15.

The results shown in Table 14 are definitely surprising, in particular the value of $T^{(1)}$ for the PDI(lifetime regression, as we expect this value to be smaller from theory. How much smaller, however, is not something we can calculate or predict from theory.

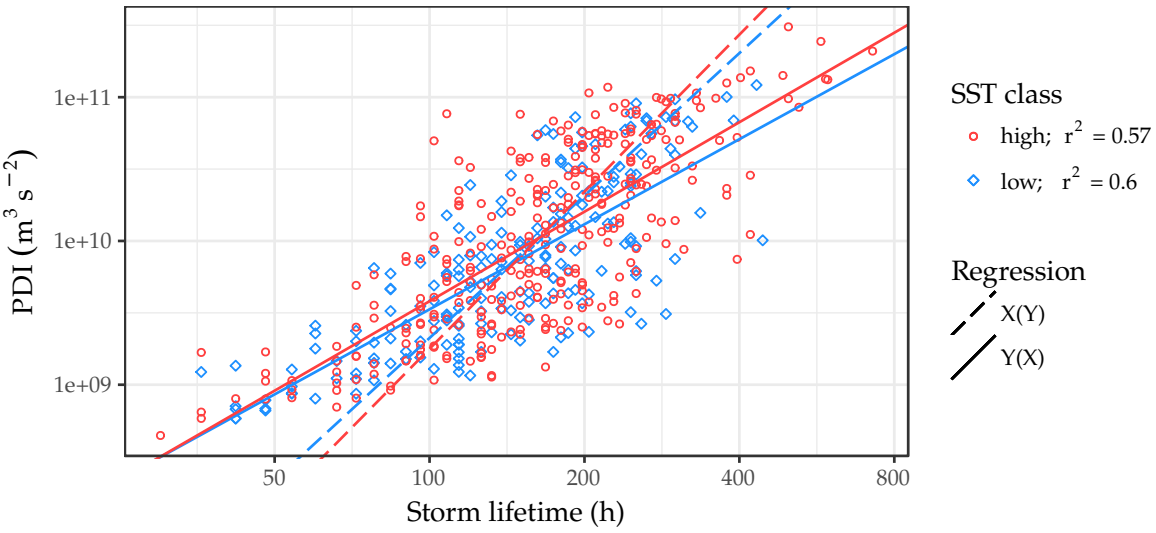


Figure 15: Scatterplot of the joint distribution and regression analysis for the PDI and lifetime of storms for the Northeast Pacific basin

X	Y	SST class	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
lifetime	PDI	Low	5.59 ± 0.25	1.97 ± 0.12	0.60
		High	5.45 ± 0.23	2.07 ± 0.10	0.57
PDI	lifetime	Low	-0.83 ± 0.18	0.30 ± 0.02	0.60
		High	-0.55 ± 0.14	0.28 ± 0.01	0.57

Table 13: Linear regression coefficients obtained performing OLS on the Northeast Pacific basin data

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.149	0.103	0.027	0.437	0.661	1.099
PDI	lifetime	0.285	0.028	0.027	1.273	1.254	2.527

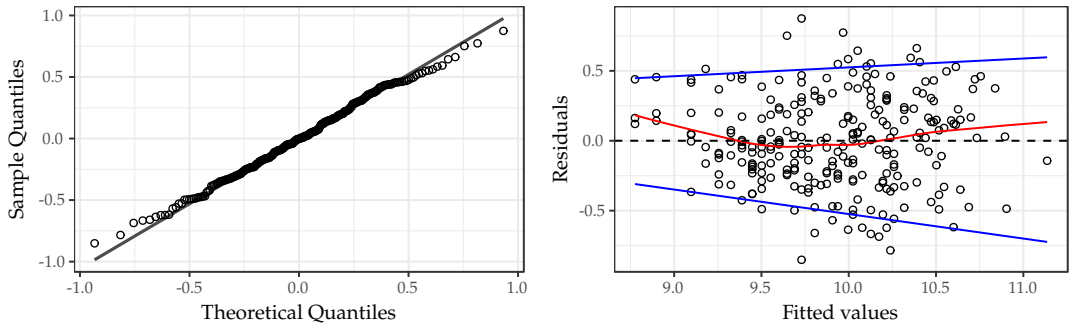
Table 14: Value of the test statistics for Northeast Pacific basin data set using OLS

4.4 Testing the linear regression assumptions

As stated in § 2.1, in linear regression the standard errors and hypothesis tests associated with the linear model rely on the random error ϵ being normal, independent, and homoscedastic.

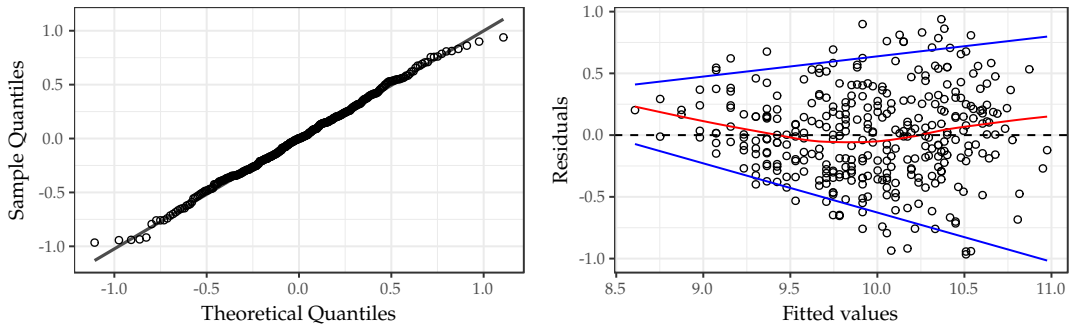
Therefore, to be sure the results found on § 4.3 are reliable, we should test these assumptions using the regression diagnostic tools introduced in § 2.5. First we will analyse the North Atlantic basin data, and then we will analyse the Northeast Pacific basin data.

The first step is to have a look at the diagnostic plots. The diagnostic plots for the North Atlantic basin, separating the data by SST class can be observed in Figure 16.



(a) Q-Q plot (low SST)

(b) Residuals vs fitted (low SST)



(c) Q-Q plot (high SST)

(d) Residuals vs fitted (high SST)

Figure 16: Diagnostic plots to analyse the residuals for the North Atlantic basin

In the Q-Q plots shown in Figure 16a and Figure 16c we can see that the residuals are mostly normal; if anything, the distribution has light tails. The residual plots in Figure 16b and Figure 16d, however, tell show us gross heteroscedasticity, specially for high-SST years.

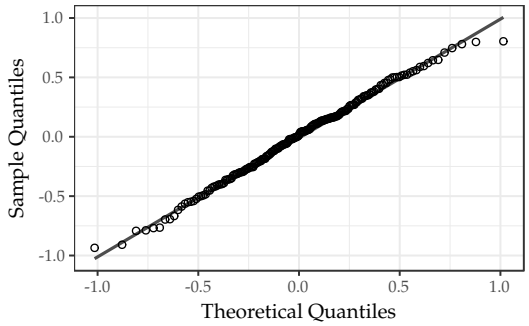
To confirm this numerically, let us explore the results of performing the Lilliefors, correlation, and Breusch–Pagan tests for the North Atlantic basin in Table 15.

Data	Lilliefors	Correlation	Breusch–Pagan
Low-SST	0.7416	1.0000	0.0376
High-SST	0.9740	1.0000	0.0002

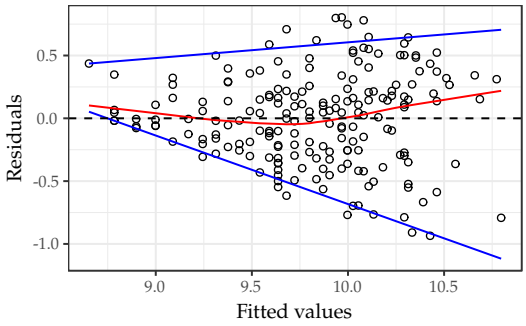
Table 15: List of p-values associated with the statistical hypothesis tests to respectively analyse normality, independence, and homoscedasticity of the residuals on the low-SST and high-SST subsets of the North Atlantic basin

The p-values tell us that the only rejected hypothesis is homoscedasticity for both datasets, which is precisely what we observed on the residual plots.

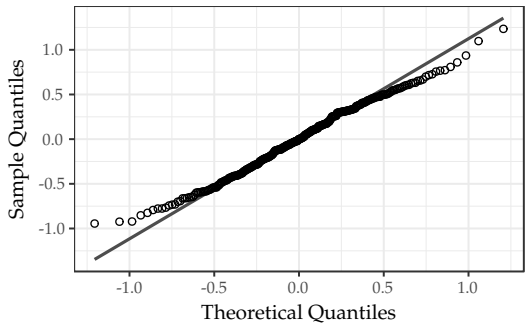
For the Northeast Pacific basin we see a similar picture. The diagnostic plots for the this basin, separating the data by SST class can be observed in Figure 17.



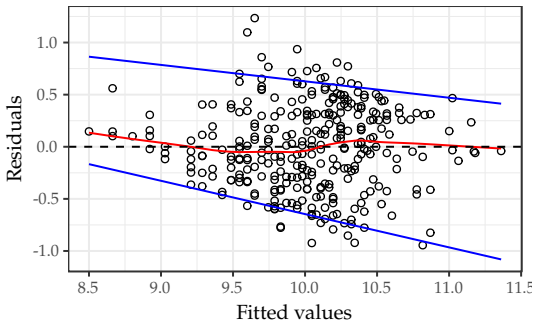
(a) Q-Q plot (low SST)



(b) Residuals vs fitted (low SST)



(c) Q-Q plot (high SST)



(d) Residuals vs fitted (high SST)

Figure 17: Diagnostic plots to analyse the residuals for the Northeast Pacific basin

In the Q-Q plot shown in Figure 17a we can see that the residuals for low-SST year follow a normal distribution with light tails. The Q-Q plot for high-SST years shown in Figure 17c, however, is strongly non-normal. The residual plots in Figure 17b and Figure 17d tell show us quite heteroscedastic residuals, specially for low-SST years.

Again, to confirm this numerically, we explore the results of performing the Lilliefors, correlation, and Breusch–Pagan tests for the Northeast Pacific basin seen in Table 16.

Data	Lilliefors	Correlation	Breusch–Pagan
Low-SST	0.7106	1.0000	0.0000
High-SST	0.0217	1.0000	0.1974

Table 16: List of p-values associated with the statistical hypothesis tests to respectively analyse normality, independence, and homoscedasticity of the residuals on the low-SST and high-SST subsets of the Northeast Pacific basin

In this case, for low-SST years the rejected hypothesis is homoscedasticity, while for the high-SST years normality is the rejected hypothesis. This is precisely what we observed on the residual plots.

Having heteroscedasticity and non-normality tells us we cannot fully rely on the calculated standard errors, and we need to use bootstrap to obtain a more robust linear model to be able to infer statistical properties of the data using linear regression as the underlying model.

4.5 Analysis using bootstrap

We saw in § 4.4 that some of the assumptions for the residual errors ϵ_i in the linear regression do not hold neither for the North Atlantic nor for the Northeast Pacific basins. It is for this reason that we use the bootstrapping methodology explained in § 2.6.1 to resample the observations in order to obtain coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and R^2 robust to failure of the model assumptions.

The particular implementation of Algorithm 1 for our data can be seen in Algorithm 3. The number of simulations performed in the bootstrap algorithm is $R = 500$ for each SST class.

Algorithm 3: Bootstrap applied to linear regression

Data: Hurricane observational data O, with paired variables X, Y; classified by SST class ($C : \{low, high\}$), with n and m observations respectively
Result: Bootstrapped data for coefficient estimates for each SST class

for $class$ ***in*** C **do**
  $O' \leftarrow$ `SubSet` $((x, y) \in O \mid c \equiv class)$
  fit $\leftarrow$ `LinearModel` $(Y' \sim X')$
  $\hat{\beta}_0 \leftarrow$ `GetIntercept` (fit)
  $\hat{\beta}_1 \leftarrow$ `GetSlope` (fit)
  $R^2 \leftarrow$ `GetRSquared` (fit)
  Initialise empty vectors $\vec{\beta}^*_0$, $\vec{\beta}^*_1$, $\vec{R}^{2*}$
  for $i \leftarrow 1$ ***to*** R **do**
    $O'^* \leftarrow$ `ResampleWithReplacement` (O')
    fit$^* \leftarrow$ `LinearModel` $(Y'^* \sim X'^*)$
    $\vec{\beta}^*_0[i] \leftarrow$ `GetIntercept` (fit*)
    $\vec{\beta}^*_1[i] \leftarrow$ `GetSlope` (fit*)
    $\vec{R}^{2*}[i] \leftarrow$ `GetRSquared` (fit*)
return $(\vec{\beta}^*_{0,low}, \vec{\beta}^*_{1,low}, \vec{R}^{2*}_{low})$ & $(\vec{\beta}^*_{0,high}, \vec{\beta}^*_{1,high}, \vec{R}^{2*}_{high})$

This resampling algorithm is performed both for the PDI(lifetime) regression model as well as the inverse lifetime(PDI) model.

To obtain the bootstrapped coefficient estimates $\hat{\beta}^*_0$, $\hat{\beta}^*_1$, R^{2*}, and and their associated standard errors from the resulting vector data $\vec{\beta}^*_0$, $\vec{\beta}^*_1$, and $\vec{R}^{2*}$, one can simply calculate the mean and the standard deviation of their distributions:

$$\hat{\theta}^* = \frac{1}{R}\sum_{i=1}^{R}\theta_i, \quad \widehat{\mathrm{se}}(\hat{\theta}^*)^2 = \frac{1}{R}\sum_{i=1}^{R-1}\left(\theta_i - \hat{\theta}^*\right)^2. \tag{28}$$

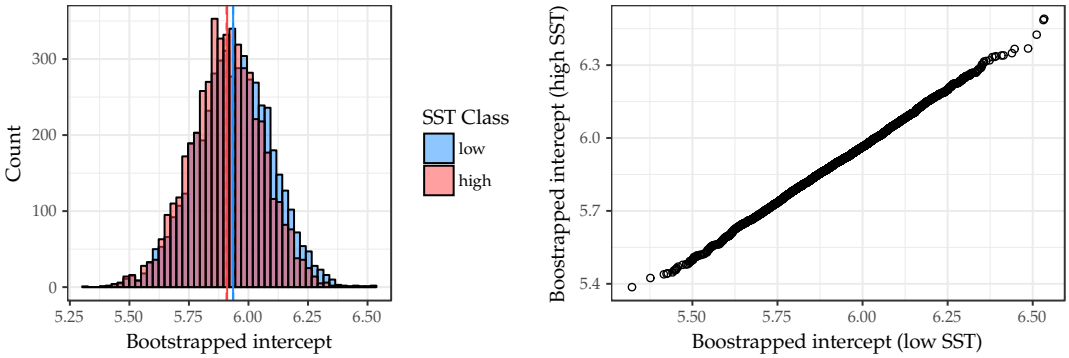
This is possible because the bootstrapped data for estimating a coefficient θ (either β_0,

β_1, or R^2) follows a normal distribution:

$$\vec{\theta}^* \sim \mathcal{N}(\hat{\theta}^*, \widehat{\text{se}}(\hat{\theta}^*)^2). \tag{29}$$

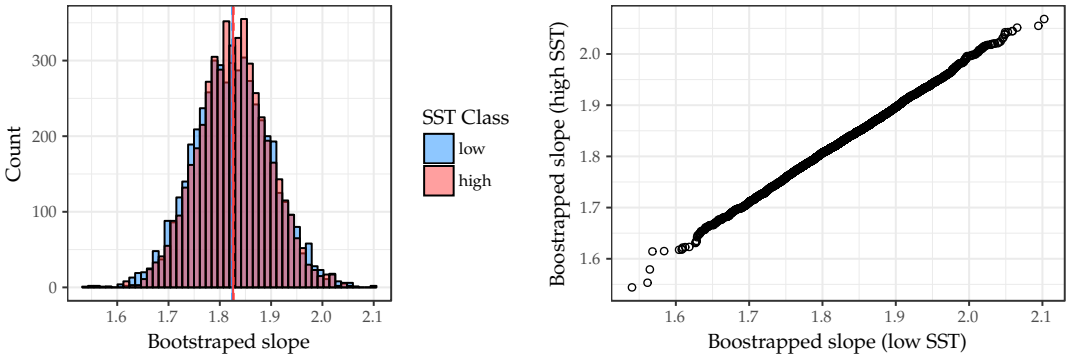
We can see this for the North Atlantic basin data in Figure 18, where we plot the histograms of the bootstrapped intercept and slopes (both for low-SST and high-SST years), as well as a Q-Q plot to compare both distributions. Notice we only show these plots are only for the PDI(lifetime) regression model; for the inverse regression, the results and conclusions are comparable.

From Figure 18a and Figure 18c we can see how the bootstrap ensures the normality in the coefficients; this is also confirmed by the Q-Q plots in Figure 18b and Figure 18d, that show that the coefficients for different SST class are from the same distribution family (in this case, a normal). This ensures the assumptions in linear regression hold.



(a) Histogram of the bootstrapped intercept

(b) Q-Q plot to compare the intercepts



(c) Histogram of the bootstrapped slope

(d) Q-Q plot to compare the slopes

Figure 18: Resampled slopes and intercepts obtained by bootstrapping for the North Atlantic basin data for the PDI(lifetime) regression model. The dashed lines represent the coefficient estimates obtained via bootstrap, while the solid lines the estimates obtained via OLS

The coefficient estimates obtained for each of the four resulting regression models for the North Atlantic data are shown in Table 17. When compared to Table 11 (coefficients

obtained using OLS), one sees that the nominal values, as well as their standard error are almost identical.

This is by no means a bad thing. The main point of using bootstrap to resample the observations of hurricane occurrences was to have a robust theory to ensure the assumptions required for the linear model hold.

X	Y	SST class	$\hat{\beta}_0^*$	$\hat{\beta}_1^*$	R^{2*}
lifetime	PDI	Low	5.91 ± 0.17	1.84 ± 0.08	0.67 ± 0.03
		High	5.90 ± 0.15	1.83 ± 0.07	0.61 ± 0.03
PDI	lifetime	Low	-1.44 ± 0.15	0.36 ± 0.02	0.67 ± 0.03
		High	-1.15 ± 0.14	0.34 ± 0.01	0.62 ± 0.03

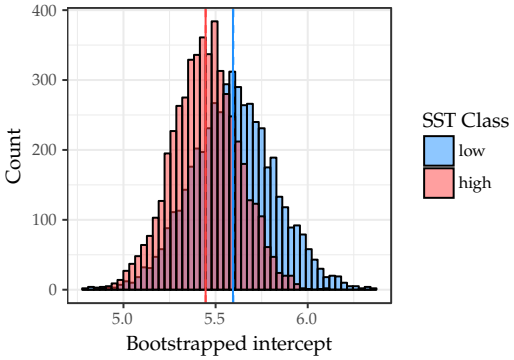
Table 17: Linear regression coefficients obtained performing bootstrap on the North Atlantic basin data

The values of the test statistics calculated using the coefficient estimates obtained using bootstrap for this basin are shown in Table 18. The results, save some small discrepancies in the lifetime(PDI) regression model, are comparable to those displayed in Table 12.

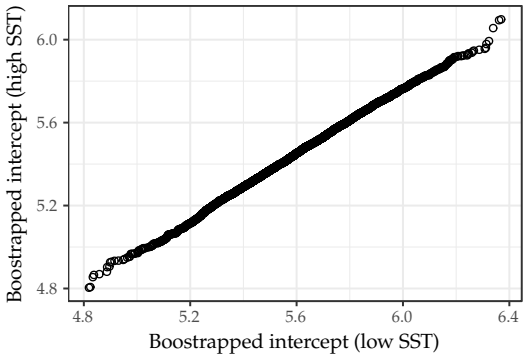
X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.007	0.007	0.054	0.031	0.065	0.095
PDI	lifetime	0.289	0.027	0.048	1.404	1.324	2.727

Table 18: Value of the studied statistics for North Atlantic basin data set using bootstrap

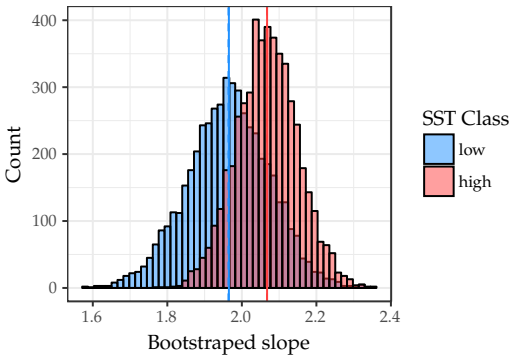
For the Northeast Pacific basin data, we also observe normality in the bootstrapped coefficients, as can be seen in Figure 19. A major difference, that we did not see in Figure 18 is that the distributions associated to low-SST and high-SST years present very distinct shapes, but this is a result of the inherent properties of the data; it represents the same numerical difference we already saw in the OLS coefficient estimates shown in Table 13), but from a graphical point of view.



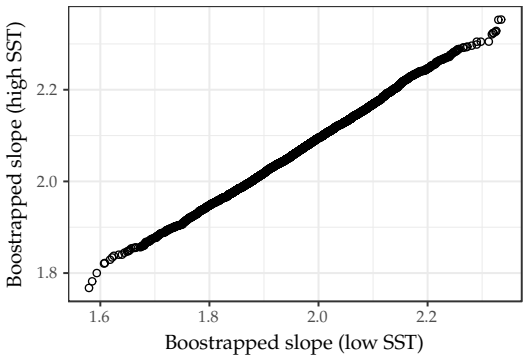
(a) Histogram of the bootstrapped intercept



(b) Q-Q plot to compare the intercepts



(c) Histogram of the bootstrapped slope



(d) Q-Q plot to compare the slopes

Figure 19: Resampled slopes and intercepts obtained by bootstrapping for the Northeast Pacific basin data for the PDI(lifetime) regression model. The dashed lines represent the coefficient estimates obtained via bootstrap, while the solid lines the estimates obtained via OLS

The coefficient estimates obtained for each of the four resulting regression models for the Northeast Pacific data are shown in Table 19. When compared to Table 13 (coefficients obtained using OLS), one sees that the nominal values, as well as their standard error are almost identical, just as happened for the North Atlantic data.

X	Y	SST class	$\hat{\beta}_0^*$	$\hat{\beta}_1^*$	R^{2*}
lifetime	PDI	Low	5.59 ± 0.24	1.97 ± 0.12	0.60 ± 0.05
		High	5.44 ± 0.18	2.07 ± 0.08	0.57 ± 0.04
PDI	lifetime	Low	-0.84 ± 0.17	0.30 ± 0.02	0.60 ± 0.05
		High	-0.56 ± 0.14	0.28 ± 0.01	0.57 ± 0.04

Table 19: Linear regression coefficients obtained performing bootstrap on the Northeast Pacific basin data

The values of the test statistics calculated using the coefficient estimates obtained using bootstrap for this basin are shown in Table 20. The results are comparable and almost identical to those displayed in Table 14.

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.152	0.103	0.030	0.509	0.728	1.237
PDI	lifetime	0.286	0.028	0.027	1.266	1.266	2.532

Table 20: Value of the studied statistics for Northeast Pacific basin data set using bootstrap

4.6 Analysis using permutation tests

In § 4.3 and § 4.5 we noticed that the value of some of the $T^{(i)}$ statistics seem a bit high, both when performing a standard (OLS) regression analysis or a bootstrap-powered regression analysis. The problem is there is no way to quantify how big these statistics can be from theory.

It is for this reason that we should perform a permutation test on the data following the methodology explained in § 2.6.2. This would allow us to properly quantify the statistical significance of evidence against the hypothesis that storms of equal lifetime should, in theory, have the same wind speed and PDI, and have the same joint distribution, regardless of the SST.

The particular implementation of Algorithm 2 for our data can be seen in Algorithm 4. The number of simulations performed in the permutation test algorithm is $R = 1000$.

In the algorithm, $\text{fit}_{low/high}$ is an object that contains the coefficient estimates of the linear model. For the sake of robustness in the coefficient estimates calculated on each permutation, `LinearModel()` allows, as an option, to perform the estimations using bootstrap; the number of bootstrap simulations in these cases is $R' = 500$.

Algorithm 4: Permutation test to compare two populations using linear regression

Data: Hurricane observational data O, with paired variables X, Y; classified by SST class ($C : \{low, high\}$), with n and m observations respectively
Result: p-values defined under the null hypothesis

$O_{low} \leftarrow$ `SubSet` $((x, y) \in O \mid c \equiv low)$
$O_{high} \leftarrow$ `SubSet` $((x, y) \in O \mid c \equiv high)$ `// Notice that` $O_{low} \cap O_{high} = O$
$\text{fit}_{low} \leftarrow$ `LinearModel` $(Y_{low} \sim X_{low})$
$\text{fit}_{high} \leftarrow$ `LinearModel` $(Y_{high} \sim X_{high})$
$T \leftarrow$ `GetStatistic` $(\text{fit}_{low}, \text{fit}_{high})$
count = 0
for $i \leftarrow 1$ ***to*** R **do**
    $O' \leftarrow$ `Permute` (O)
    $O^*_{low} \leftarrow$ `SubSet` $((x, y)_i \in O', \forall i \in [1, n])$
    $O^*_{high} \leftarrow$ `SubSet` $((x, y)_i \in O', \forall i \in [n + 1, n + m])$ `//` $O^*_{low} \cap O^*_{high} = O'$
    $\text{fit}^*_{low} \leftarrow$ `LinearModel` $(Y^*_{low} \sim X^*_{low})$
    $\text{fit}^*_{high} \leftarrow$ `LinearModel` $(Y^*_{high} \sim X^*_{high})$
    $T^* \leftarrow$ `GetStatistic` $(\text{fit}^*_{low}, \text{fit}^*_{high})$
    if $T^* > T$ **then**
        count $\leftarrow$ count + 1
return *p-value* $\leftarrow$ *count* / R

Notice that we only illustrate one test statistic in Algorithm 4; this is just to avoid overcomplicating the basic concept behind the permutation test. Naturally, in our case,

we calculate all the test statistics proposed in section § 4.2:

$$T^{(1)} = \left|\hat{\beta}_{0,h} - \hat{\beta}_{0,l}\right|, \quad T^{(2)} = \left|\hat{\beta}_{1,h} - \hat{\beta}_{1,l}\right|, \quad T^{(3)} = \left|R_h^2 - R_l^2\right|, \tag{26 bis}$$

$$T^{(4)} = \frac{\left|\hat{\beta}_{0,h} - \hat{\beta}_{0,l}\right|}{\widehat{\text{se}}(\hat{\beta}_{0,h} - \hat{\beta}_{0,l})}, \quad T^{(5)} = \frac{\left|\hat{\beta}_{1,h} - \hat{\beta}_{1,l}\right|}{\widehat{\text{se}}(\hat{\beta}_{1,h} - \hat{\beta}_{1,l})}, \quad T^{(6)} = T^{(4)} + T^{(5)}. \tag{27 bis}$$

In Table 21 and Table 22 we can see the p-values associated to each test statistic under the null hypothesis, using the standard OLS method and the bootstrap method, respectively, to calculate the coefficient estimates of the regression models for the North Atlantic basin data.

As we can see, neither of the tests rejects the null hypothesis, both using OLS and bootstrap as the underlying calculation for the model coefficient estimates. This indicates a strong evidence in favour of the null hypothesis being true.

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.164	0.184	0.749	0.160	0.187	0.173
PDI	lifetime	0.901	0.991	0.750	0.900	0.992	0.968

Table 21: List of p-values of the standard (OLS) permutation test for the North Atlantic basin data

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.160	0.187	0.778	0.132	0.162	0.142
PDI	lifetime	0.925	0.987	0.757	0.922	0.995	0.977

Table 22: List of p-values of the bootstrap-powered permutation test for the North Atlantic basin data

Similarly, for the Northeast Pacific basin, neither of the tests rejects the null hypothesis, both using OLS (Table 21) and bootstrap (Table 22) as the underlying calculation for the model coefficient estimates. This indicates a strong evidence in favour of the null hypothesis being true.

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.233	0.232	0.354	0.243	0.245	0.247
PDI	lifetime	0.629	0.480	0.329	0.622	0.475	0.549

Table 23: List of p-values of the standard (OLS) permutation test for the Northeast Pacific basin data

X	Y	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$	$T^{(5)}$	$T^{(6)}$
lifetime	PDI	0.250	0.238	0.365	0.230	0.231	0.229
PDI	lifetime	0.717	0.533	0.376	0.721	0.551	0.632

Table 24: List of p-values of the bootstrap-powered permutation test for the Northeast Pacific basin data

5 Geographical analysis

5.1 Geographical variables

During the analysis performed in § 3 and § 4 we only used geographical information about the hurricane occurrences for the calculation of the basin-wide averaged SST.

From the raw HURDAT2 data sets, we can extract some relevant geographical data about each tropical-cyclone that could help obtain some insight or physical reason behind the displacement of the joint distribution of PDI and lifetime of the storms between low-SST and high-SST years occurrences.

We focus on the geographical genesis location of a tropical-cyclone, as well as its death location. This information is easy to get from the individual longitude and latitude tracks for each hurricane.

Apart from this, we calculate the total travelled distance, or path length, of each hurricane by means of the *spherical law of cosines*:

$$d(p_1, p_2) = \cos^{-1}\left[\sin(\phi_1) \cdot \sin(\phi_2) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \cos(\Delta\lambda)\right] \cdot R_{\mathrm{E}}, \qquad (30)$$

where ϕ and λ respectively represent latitude and longitude in radians, and R_{E} represents the Earth's radius in meters.

In Table 25 one can see the structure of the cleaned data illustrate these newly calculated geographical variables.

storm.id <chr>	storm.name <chr>	n.obs <int>	storm.duration <dbl>	...	first.lat <dbl>	last.lat <dbl>	first.long <dbl>	last.long <dbl>	distance <dbl>
AL011966	ALMA	42	907200	···	12.7	42.0	-84.0	-70.5	4129705
AL021966	BECKY	9	194400	···	32.4	45.5	-57.8	-58.5	1678458
AL031966	CELIA	36	777600	···	19.1	52.0	-59.5	-57.0	5635625
AL041966	DOROTHY	37	799200	···	31.0	53.5	-41.0	-38.5	2864172
AL051966	ELLA	26	561600	···	10.0	24.3	-35.0	-68.4	3917529
AL071966	GRETA	26	561600	···	13.7	28.0	-48.4	-71.7	2977350

Table 25: Excerpt of the North Atlantic data set, focusing on the geographical variables

Similarly to the unified HURDAT2 & HadISST1 data sets (see § 3.4), these geographically enhanced data sets are available in the `HurdatHadISSTData` package:

- `tc.pdi.geog.natl` – Data set for the North Atlantic basin.
- `tc.pdi.geog.epac` – Data set for the Northeast Pacific basin.
- `tc.pdi.geog.all` – Data set for both basins.

5.2 Analysis of the path length

In Figure 20 and Figure 21 we can see the bivariate lognormal distributions of the path length d and lifetime of the storms for the North Atlantic and Northeast Pacific basins separating storms by SST class.

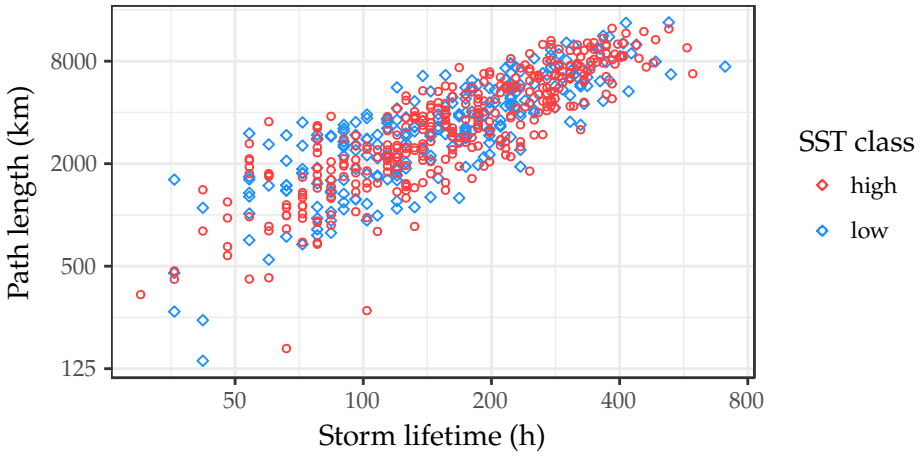


Figure 20: Bivariate lognormal distribution $f(d, \text{lifetime})$ of the hurricane observations for the North Atlantic basin

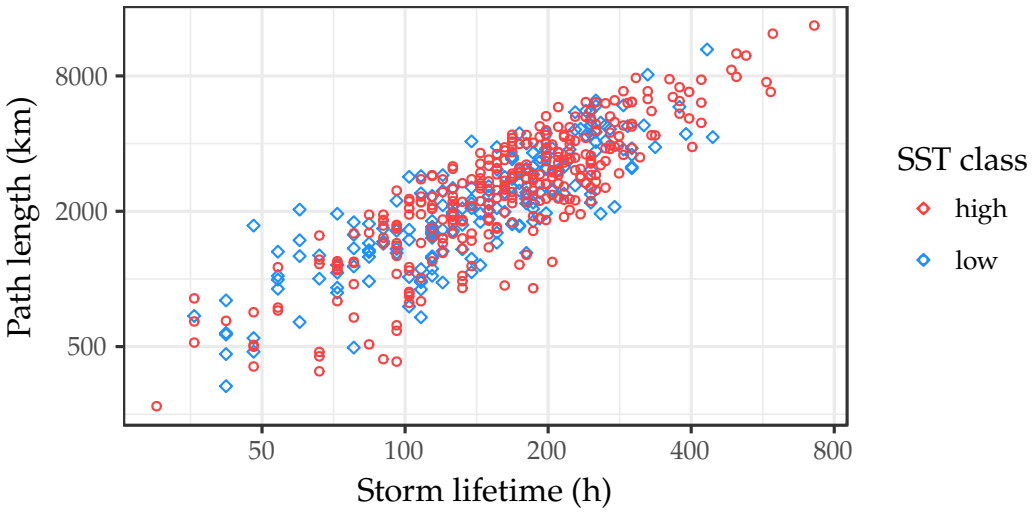


Figure 21: Bivariate lognormal distribution $f(d, \text{lifetime})$ of the hurricane observations for the Northeast Pacific basin

Contrarily to the procedure of analysing the marginals of this joint distribution that we performed on § 4.1.2, we want study the mean focus on the forward speed of the

hurricanes. Notice that this speed is different to the sustained surface wind speed. Naturally, this mean forward speed is calculated as

$$\langle v_f \rangle = \frac{d}{\text{lifetime}}. \tag{31}$$

We think this variable may be an intermediary variable to relate the storm path length and its PDI, and expect a displacement to higher speeds for high-SST years.

In Figure 22 we show a histogram of the mean forward speed for the North Atlantic basin, while in Figure 23 we show the same histogram for the Northeast Pacific basin.

As it can be seen, there seems to be no general trend on the behaviour of the mean forward between the North Atlantic and Northeast Pacific basins.

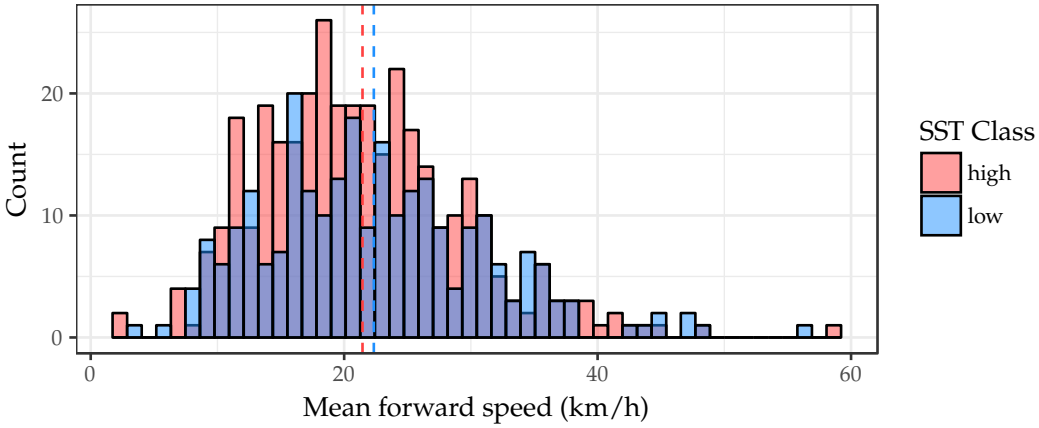


Figure 22: Mean forward speed histogram for the North Atlantic basin

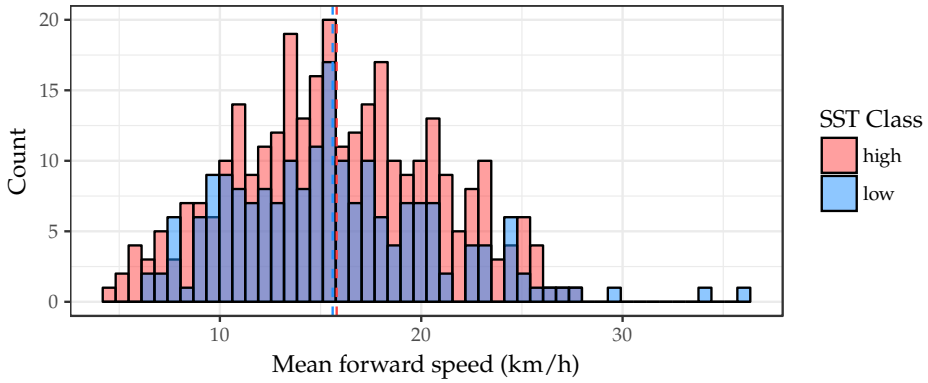


Figure 23: Mean forward speed histogram for the Northeast Pacific basin

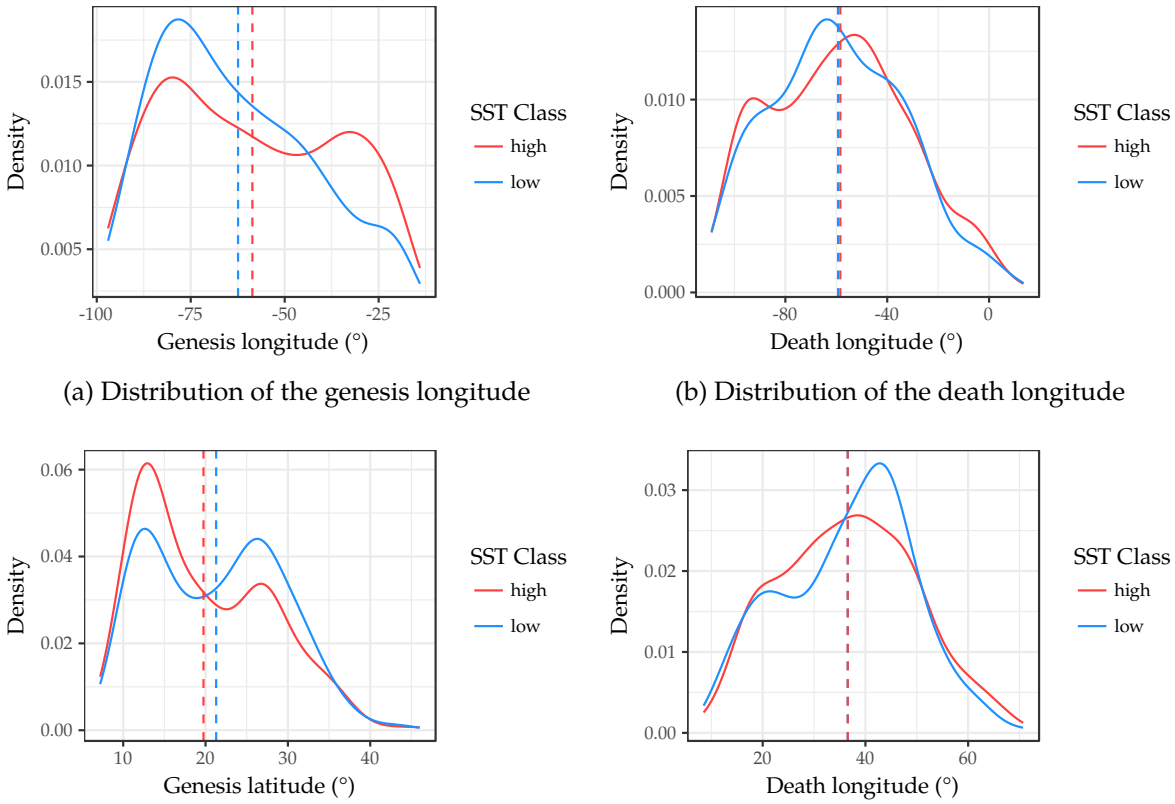
5.3 Analysis of the location

A particularly important variable in this analysis is, obviously, the location of genesis of the storms, and possibly their location of death.

In this part we do an exploratory comparison of the difference in position of genesis and death of tropical-cyclones between low-SST and high-SST years.

In Figure 24 we can see the distributions of genesis and death positions (longitude and latitde) of the storms occurred on the North Atlantic basin, while in Table 26 we can see the expected value of the distributions.

The results suggest show the major difference between low-SST and high-SST years is the location of the genesis of the hurricanes, as it seems to be displaced to the South-East. However, there is no significant difference in the death location.



(a) Distribution of the genesis longitude

(b) Distribution of the death longitude

(c) Distribution of the genesis latitude

(d) Distribution of the death latitude

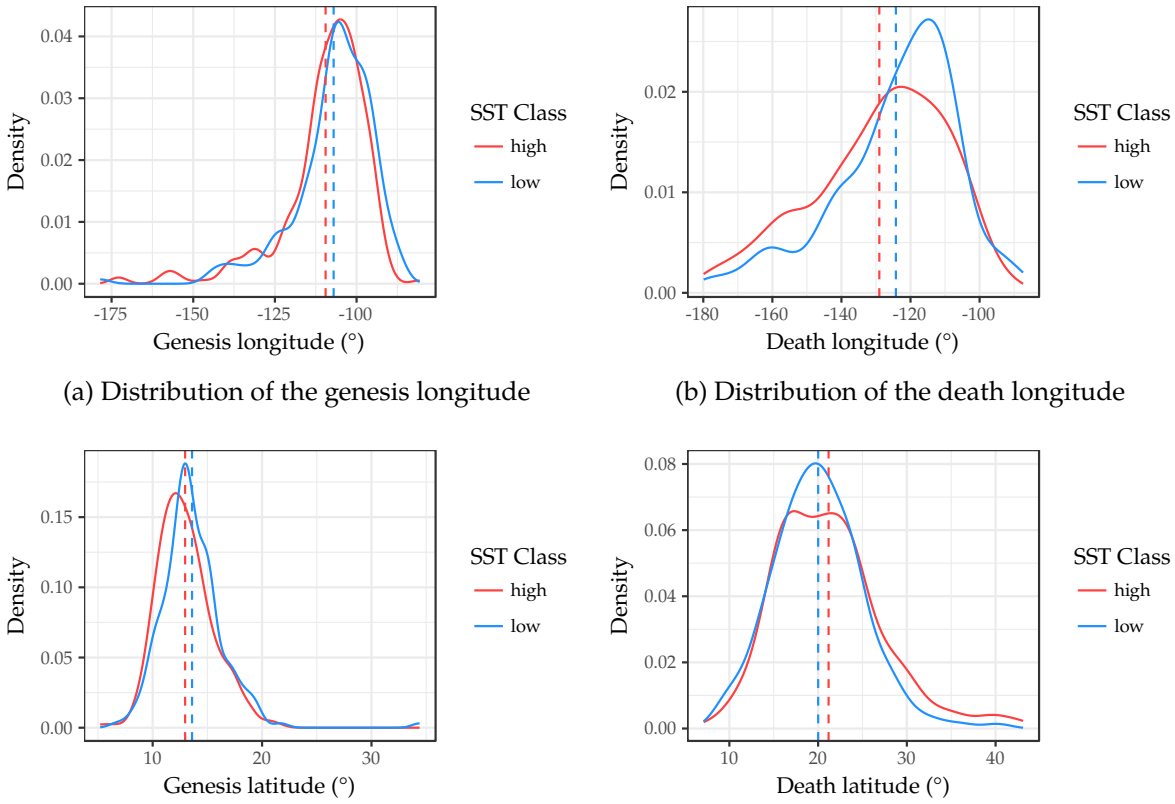
Figure 24: Spatial distributions of the geographical position variables of storms for the North Atlantic basin

SST Class	$\bar{\lambda}_{\text{gen}}$	$\bar{\phi}_{\text{gen}}$	$\bar{\lambda}_{\text{death}}$	$\bar{\phi}_{\text{death}}$
Low	-59.35 ± 22.94	20.84 ± 7.97	-59.37 ± 24.09	33.08 ± 12.94
High	-58.68 ± 23.44	19.47 ± 7.74	-59.39 ± 26.57	34.72 ± 13.31

Table 26: Summary of the expected values of the geographical position variables of storms for the North Atlantic basin

For the Northeast Pacific we have a similar scenario. In Figure 25 we can see the distributions of genesis and death positions (longitude and latitde) of the storms, while in Table 27 we can see the expected value of the distributions.

The results suggest show the major difference between low-SST and high-SST years is the location of the genesis of the hurricanes, as it seems to be displaced to the South-East. Contrarily to the North Atlantic, in the Northeast Pacific, there seems to be a slight displacement in the death position for high-SST years to the North-East.



(a) Distribution of the genesis longitude

(b) Distribution of the death longitude

(c) Distribution of the genesis latitude

(d) Distribution of the death latitude

Figure 25: Spatial distributions of the geographical position variables of storms for the Northeast Pacific basin

SST Class	$\bar{\lambda}_{\text{gen}}$	$\bar{\phi}_{\text{gen}}$	$\bar{\lambda}_{\text{death}}$	$\bar{\phi}_{\text{death}}$
Low	-107.90 ± 11.98	13.74 ± 2.60	-123.28 ± 17.34	19.35 ± 4.71
High	-111.12 ± 14.92	13.13 ± 2.64	-129.18 ± 20.11	20.64 ± 6.15

Table 27: Summary of the expected values of the geographical position variables of storms for the Northeast Pacific basin

6 Conclusions

In this work we have thoroughly explored and analysed the theoretical foundation of the linear regression under the ordinary least squares method. In particular, we saw both graphically and numerically how some of the assumptions on the random error, such as homoscedasticity and normality, do not hold for the joint distribution of PDI and storm lifetime for the North Atlantic and Northeast Pacific basins.

To solve these problems, we used bootstrap to resample the observations in order to provide a more accurate and robust regression analysis than the one provided by the OLS method.

Last but not least, we proposed a statistical test to compare low-SST and high-SST years by performing a permutation test. This allowed us to quantify the statistical significance of evidence against the hypothesis that storms of equal lifetime have the same PDI and same joint distribution, regardless of the SST. The results provide strong evidence that this hypothesis is indeed true, as none of the performed tests rejected the null hypothesis.

Our conclusions are compatible with the view of tropical cyclones as an activation process, in which, once the event has started, its intensity is kept in critical balance between attenuation and intensification (and so, higher SST does not trigger more intensification).

An open question, nonetheless, is why the increase of tropical-cyclone lifetime with SST triggers an increase in wind speed as a by-product.

The results of a simple exploratory analysis of the geographical properties of the tropical-cyclones show that the longer lifetimes for high-SST are mainly due to a shift to South-East of the tropical-cyclones genesis point, although further analysis is needed.

The steps to follow would be to perform a hierarchical clustering of the location of genesis and death of storms using the aggregate information provided by the PDI, lifetime, location, and path length of each storm to have a deeper understanding of the difference between hurricane occurrences in low-SST and high-SST years.

References

[1] P. J. Webster et al. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science (New York, N.Y.)*, 309(5742):1844–6, 2005. ISSN: 1095-9203. URL: http://www.ncbi.nlm.nih.gov/pubmed/16166514.

[2] Á. Corral, A. Ossó and J. E. Llebot. Scaling of tropical-cyclone dissipation. *Nature Physics*, 6(9):693–696, 2010. ISSN: 1745-2473. URL: http://www.nature.com/doifinder/10.1038/nphys1725.

[3] K. A. Emanuel. An Air-Sea Interaction Theory for Tropical Cyclones. Part I: Steady-State Maintenance. *Journal of the Atmospheric Sciences*, 43(6):585–605, 1986. ISSN: 0022-4928. URL: https://doi.org/10.1175/1520-0469(1986)043%3C0585:AASITF%3E2.0.CO;2.

[4] K. A. Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436(7051):686–688, 2005. ISSN: 0028-0836. URL: http://www.nature.com/doifinder/10.1038/nature03906.

[5] K. Trenberth. Uncertainty in Hurricanes and Global Warming. *Science*, 308(5729):1753–1754, 2005. URL: http://science.sciencemag.org/content/sci/308/5729/1753.full.pdf.

[6] K. A. Emanuel. Tropical Cyclones. *Annual Review of Earth and Planetary Sciences*, 31(1):75–104, 2003. ISSN: 0084-6597. URL: http://www.annualreviews.org/doi/10.1146/annurev.earth.31.100901.141259.

[7] G. James et al. *An Introduction to Statistical Learning*. 2006. ISBN: 9780387781884. arXiv: arXiv:1011.1669v3. URL: http://books.google.com/books?id=9tv0taI8l6YC.

[8] P. Dalgaard. *Introductory Statistics with R*. Statistics and Computing. Springer New York, New York, NY, 2008. ISBN: 978-0-387-79053-4. URL: http://link.springer.com/10.1007/978-0-387-79054-1.

[9] H. W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62(318):399, 1967. ISSN: 01621459. URL: https://www.jstor.org/stable/2283970?origin=crossrefhttps://www.jstor.org/stable/1911963?origin=crossref.

[10] N. M. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011. ISSN: 9789673631575. DOI: doi:10.1515/bile-2015-0008.

[11] T. S. Breusch and A. R. Pagan. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47(5):1287, 1979. ISSN: 00129682. URL: https://www.jstor.org/stable/1911963?origin=crossref.

[12] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997. ISBN: 9780511802843. URL: http://ebooks.cambridge.org/ref/id/CBO9780511802843.

[13] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Springer US, Boston, MA, 1993. ISBN: 978-0-412-04231-7. URL: http://link.springer.com/10.1007/978-1-4899-4541-9.

[14] P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, volume 53 of number 9 in *Springer Series in Statistics*. Springer-Verlag, New York, 2005, page 666. ISBN: 0-387-20279-X. arXiv: arXiv:1011.1669v3. URL: http://link.springer.com/10.1007/b138696.

[15] F. B. Butar and J.-w. Park. Permutation Tests for Comparing Two Populations. *Journal of Mathematical Sciences and Mathematics Education*, 3(2):19–30, 2008.

[16] D. Polko-Zając. Application of the permutation test for comparing regression models. In *10th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, 2016.

[17] National Hurricane Center. URL: http://www.nhc.noaa.gov/.

[18] HURDAT Re-Analysis Project. URL: http://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html.

[19] C. Landsea and J. Franklin. Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. *Monthly Weather Review*, 141(10):3576–3592, 2013. ISSN: 0027-0644. URL: http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-12-00254.1.

[20] North Atlantic Hurricane Basin (1851–2016) Comparison of Original and Revised HURDAT. URL: http://www.aoml.noaa.gov/hrd/hurdat/comparison_table.html.

[21] S. Delgado, C. W. Landsea and H. Willoughby. Reanalysis of the 1954–63 Atlantic Hurricane Seasons. *Journal of Climate*, 31(11):4177–4192, 2018. ISSN: 0894-8755. URL: http://journals.ametsoc.org/doi/10.1175/JCLI-D-15-0537.1.

[22] C. Landsea, J. Franklin and J. Bevin. The revised Atlantic hurricane database (HURDAT2), 2014. URL: http://www.nhc.noaa.gov/data/hurdat/hurdat2-format-atlantic.pdf.

[23] C. Landsea et al. The revised Northeast and North Central Pacific hurricane database (HURDAT2), 2016. URL: http://www.nhc.noaa.gov/data/hurdat/hurdat2-format-nencpac.pdf.

[24] NCAR Climate Data Guide Content with Tag: SST - sea surface temperature. URL: https://climatedataguide.ucar.edu/variables/ocean/sst-sea-surface-temperature.

[25] N. A. Rayner. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, 108(D14):4407, 2003. ISSN: 0148-0227. URL: http://doi.wiley.com/10.1029/2002JD002670.

[26] Weather and climate change - Met Office. URL: http://www.metoffice.gov.uk/.

[27] Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST). URL: http://www.metoffice.gov.uk/hadobs/hadisst/.

[28] Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST.2). URL: http://www.metoffice.gov.uk/hadobs/hadisst2/.

[29] HadISST data format instructions. URL: http://www.metoffice.gov.uk/hadobs/hadisst/data/Read_instructions_sst.txt.

[30] A. Hernández. Tropical-Cyclones on GitLab. URL: https://gitlab.com/aldomann/tropical-cyclones-plus.

[31] A. Hernández. Tropical-Cyclones (HurdatHadISSTData Package) on GitLab. URL: https://gitlab.com/aldomann/hurdat-hadisst-data.

[32] J. L. McBride. *Observational analysis of tropical cyclone formation*. Colorado State University, 1979. URL: http://nla.gov.au/nla.cat-vn1301003.

[33] P. Phillips. Understanding spurious regressions in econometrics. *Journal of Econometrics*, 33(3):311–340, 1986. ISSN: 03044076. URL: http://linkinghub.elsevier.com/retrieve/pii/0304407686900011.

[34] N. T. Thomopoulos. *Statistical Distributions*. Springer International Publishing, Cham, 2017, pages 1–12. ISBN: 978-3-319-65111-8. URL: http://link.springer.com/10.1007/978-3-319-65112-5.