South Africa drought and wildlife survival

Andrea Corrado 20205529

Contents

1 Introduction **1**

2 Technical Background **1**

3 Exploratory Data Analysis **3**

4 Climate Modeling **8**

5 Species count Modeling **11**

6 Conclusion **16**

7 Appendices **16**

Abstract

In this short study, we investigate the major climate change phenomenons, such as a rainfall amount and temperature, over the last century. We find a relation between year and temperature and between temperature and rainfall amount which we exploit to make prediction about a future plausible 50 years behavior. Moreover, we focus our attention on the species survival in South Africa in relation to the climate phenomenons. What we find is that the temperature is steeply raising over time and, as it is negatively correlated, the rainfall amount is consequently decreasing. We also find that these phenomenons are associated to the animals individual count evolution in time which, according to the estimated model, is decreasing as well. After the model validation, we provide a possible estimation of the species evolution and notice that they may reach only one individual left, as soon as in 15 years time, on average, which would not allow them to reproduce anymore. This clearly implies the species extinction.

In section 1 we introduce the reader to the topic of interest. In section 2 we briefly introduce the methodology employed in this research. In section 3 we perform a data exploratory analysis on the climate phenomenons. In section 4 we model the climate data and quantify the model uncertainty. In section 5 we model the wildlife individual count in relation to the climate data and provide a possible scenario for the future 50 years time. In section 6 we summarise the finding of this research.

1 Introduction

During the last decades, interest and awareness about the climate change has steeply increased. The Framework Convention on Climate Change (UNFCCC) has, as its ultimate objective, the stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system. In order to allow for ecosystems to adapt naturally, such a level should be achieved within a time frame which would allow the development to proceed in a suitable manner. However, the amount of evidences clearly indicates that this is not happening and that the human society is contributing significantly to worsen the condition the earth has to suffer (Parry, Carter, and Hulme 1996). There have been several studies examining and reporting the potential consequences of these phenomenons. For instance, the temperature raising, changes in rainfall amount and greenhouse gases emissions have been of particular interest. Since last century, we have witnessed an increasing frequency in the number and severity of droughts in particular territories such as South Africa *SA*. Due to the fact that areas such as *SA* suffer from low economic power, a drought may easily drive towards very dramatic consequences. It can affect either agriculture, for a period that lasts up to 6 months, either hydrological systems for a period up to 24 months (Rouault and Richard 2003).

In literature, we can find several different examples of solutions to reduce the impact of droughts on the agriculture field ie. by growing plants restraint to higher temperature or building computational model to predict droughts and take actions before drastic events (Grayson 2013). However, in this study, we wish to focus our attention on the drought consequences on wildlife population evolution. For instance, different studies have shown how animals mortality grows during periods with the highest temperature within the driest areas (Hillman and Hillman 1977). We will expand this idea to more recent data about waterbirds in South Africa (Global Biodiversity Information Facility, n.d.). The data set we are going to present provides the number of individual for different species and genes in each year within the last 45 years. By doing so, we are able to track the evolution of the species during time and to study its relation with the climate phenomenons. However, the data set will not provide the number of new death, new born and the relative causes. It will be our job to build a model which allows us to investigate this particular relation. In particular the data we are going to analyse concern the global temperature and rain fall amount (Group, n.d.) from 1901 up to 2020 which would give a foundation to build on the mathematical models of interest.

Having access to this information, our goal is to build a model which relates the climate change to the number of individual for each species and genes and to predict how, following the current trend, these would evolve over time in a possible 50 years future scenario.

2 Technical Background

Generalized Lienar Model

During the study we will employ the Generalized Linear Model (McCullagh and Nelder 1989) which theory would allow to establish a linear relationship between the quantity of interest $\mathbf{y}$ and a matrix of covariates $\mathbf{X}$ related to each other by a vector of parameters β which values will be determined by the data in input and identified by the *ordinary least square* (OLS) theory. In order to do so, the data matrix $\mathbf{X}$ needs to be of full rank, $rank(\mathbf{X}) = p$, so that we will be able to identify the matrix $(\mathbf{X^T X})^{-1}$ which will be used in the model estimation. This deterministic part will be combined to a stochastic component which models the uncertainty about the random variable realisation and whose distribution $\mathcal{D}(\theta)$ will be pre-identified and super-imposed in the model estimation

$$g(\mathbf{Y}) = \mathbf{X}\beta + \epsilon$$

where we define a distribution $\mathcal{D}$ of interest on the stochastic term ϵ

$$\epsilon \sim \mathcal{D}(\theta)$$

and a link function $g(.)$ which is distribution dependent.

The model will be evaluated by a set of diagnostic on the residuals $\mathbf{Y} - \hat{\mathbf{Y}}$ which provide meaningful insight on the goodness of fitness.

Generalized Additive Model

The *GLM* (McCullagh and Nelder 1989) theory is further extended to allow for non-linear relationship between the predictors $\mathbf{X}$ and the response $\mathbf{Y}$ exploiting the *Generalized Additive Model* theory (Hastie and Tibshirani 1990) while maintaining the model linearity in the parameters. Indeed, the non-linearity is allowed only for the input data $\mathbf{X}$ which will be

transformed by the usage of a basis function function $f()$ to optimally match the response value. The expression is the following

$$g(\mathbf{Y}) = f(\mathbf{X}) + \epsilon$$

where we define a distribution $\mathcal{D}$ of interest on the stochastic term ϵ

$$\epsilon \sim \mathcal{D}(\theta)$$

and a link function $g(.)$ which is distribution dependent. Here, the choice of the function $f()$ is completely arbitrary and up to the researcher. A very common and flexible choice in *GAMs* is to define the function $f(.)$ to be a *spline* (Aerts, Claeskens, and Wand 2002), which can be seen a set of connected piecewise polynomial with additional constraints to allow for a more robust and smooth estimation. Different version of this function have been proposed, we will mainly make use of the *natural spline* for its robustness in the out of sample context and computational efficiency.

Weighted Least Square

GLMs are often estimated by making usage of *OLS* theory. This can be seen as a particular case where the matrix $\mathbf{W}$ employed in the parameters estimation is equal to the identity matrix $\mathbf{W} = \mathbf{I}$. This is not always the case. In fact, during this study we perform different model estimation and will run into different cases where the model diagnostic shows violations of the homoskedasticity assumption. Therefore, to remedy to inadequate diagnostics, we will employ *Weighted Least Square* estimation theory (Hooper 1993). This approach allows us to iteratively estimate a weights matrix $\mathbf{W}$ to be employed in the model estimation as a measure of the importance for each unit $\mathbf{x}_i$ $i = \{1, \dots, n\}$ where n is the total number of observations. The weight definition employed in this study is

$$w_{ii} = \frac{1}{(y_i - \hat{y_i})^2} = \frac{1}{\epsilon_i^2}$$

This definition will allow for different weights in each observations. Those for which the model is unable to represents them in a satisfactory manner will have a large residual, hence a small weight and viceversa.

Local Regression

Local weighted regression is a non parametric framework which allows for an estimation of a response $\mathbf{y}$ as a function of a set of predictors $\mathbf{X}$ plus a stochastic component ϵ on which we posit no assumption

$$\mathbf{y} = f(\mathbf{X}) + \epsilon$$

The model depends on the span hyperparameter s which controls the number of observation to be included in the estimation of each y_i and varies in $[0, 1]$

$$\lceil s \times n \rceil = k$$

where n is the total number of observations. By doing so, for each response unit, we will use the s percentage of the point that are the closest in the p-dimensional space of the covariate according to a distance measure, such the *Euclidean distance*. Clearly, $s = 1$ would lead to the usage of all the data points, hence a global estimation, while a smaller value $s \to 0$ would lead to a more local estimation, with very few points being used.

Mixed Effect models

Mixed effects models theory offers a robust framework to deal with non-independent data. This is particularly useful when we observe data longitudinally, namely follow them in time, or when these are hierarchical, eg we can observe different years for different species-genes pairs. In order to take into account for this structure, we can include in the model the fixed effects, same as *GLMs* and *GAMs* plus the random effects which assumes that the J observed different groups are a sub set of the entire population.

$$y_i = \beta_0 + \alpha_{1,i} Z_{i,j} + \sum_{k=1}^{K} \beta_k \; x_{i,k} + \alpha_{2,i} \; Z_{i,j} \; x_{i,k} + \epsilon_i$$

where β_i governs the fixed effects for all the species-genes pairs and $\alpha_{j,i}$ governs the random effect for each individual species-genes pairs.

To evaluate the contribution of the random effects j, we can estimate its variance contribution w.r.t. the overall variance

$$\frac{\sigma_j^2}{\sum_{l=1}^{L} \sigma_l^2}$$

Ordinary Differential Equation

To track the evolution of a phenomenon over time, it is natural to think about *Oridnary Differential Equation (ODE)* theory (Gbur 2011) which allows us to

keep track of the change of a quantity in continuous time

$$\frac{d}{dt}x = f(x,t)$$

given initial condition $f(x, t = 0) = x_0$

Levenberg-Marquardt algorithm

Very often, the function defined within an ODE is characterized by a set of parameters β. In empirical studies, we usually have access to data realization of a set of covariates and a response variable $\{(y_i, \mathbf{x}_i,)\}_{i=1}^n$. In this context, we want to find the best parameter estimates $\hat{\beta}$ so that $y = f(\mathbf{x}, \hat{\beta})$ best fits the data of interest. The *Levenberg-Marquardt* algorithm (Moré 1978) is an optimization technique which provides an useful and efficient solution to the problem. In particular, in this context we wish to minimize the objective function which we define as the sum of residuals squared scaled by the estimated variance at each point

$$\chi^2(\beta) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{\sigma_{y_i}} \right)^2$$

The *LM* algorithm makes smart combination of *Gauss-Newton* and *Gradien Descend* theory. In brief, the method smartly adapt the learning rate λ at each iteration t according to the magnitude of the objective function at the previous iteration $t-1$. By doing so, we remedy to the fact that the *GN* needs a starting values close to the true value and we speed up the convergence time.

Sensitivity Analysis

Mathematical modeling is often subject to very strong assumption which may be difficult to evaluate. In order to provide a robust framework, it is necessary to validate the model under perturbations on the findings. This is why, we will exploit *Local Sensitivity Analysis* theory (Morio 2011) to evaluate the robustness of the estimated models. The method let us quantify the system response variations to parameter changes about a nominal value one-at a time by deriving the sensitivity equation

$$\frac{d}{dt}\frac{\delta \mathbf{u}}{\delta q_j} = \frac{d\mathbf{f}}{d\mathbf{u}}\frac{\delta \mathbf{u}}{\delta q_j} + \frac{\delta \mathbf{f}}{\delta q_j}$$

and by defining $\frac{\delta \mathbf{u}}{\delta q_j} = s$, the sensitivity, we get

$$\frac{d}{dt}s = \frac{d\mathbf{f}}{d\mathbf{u}}s + \frac{\delta \mathbf{f}}{\delta q_j}$$

which will be added to the system of *ODE* and numerically integrated as usual.

3 Exploratory Data Analysis

Within this exploratory data analysis section, we wish to give a short introduction to the data of interest by illustrating some characteristics. The climate data we have available concern the rainfall amount and global temperature.

In the first place, we investigate the rainfall data set. We have access to the information between January 1901 up to December 2020. For simplicity, we decide to consider the annual granularity and inspect the resulting distribution.

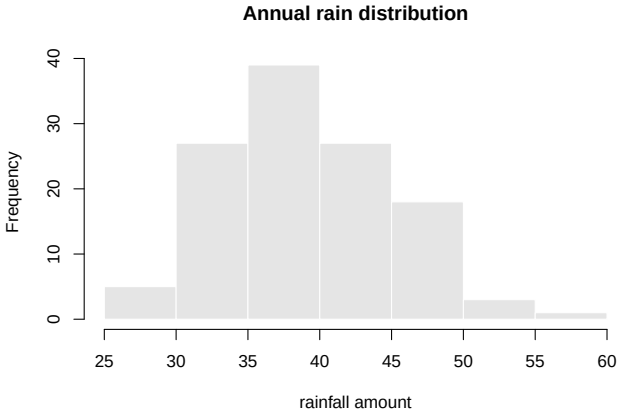


Figure 1: Yearly rainfall distribution

From figure 1 we can observe that the data distribution is somewhat bell-shaped. However, we can easily notice that the distribution is also right skewed. Clearly, due to the nature of phenomenon, the rainfall amount is left bounded by 0, $x \in \mathbb{R}_+$, meaning that we can rarely observe very extreme event such as very heavy precipitation but we cannot observe a negative amount of rainfall. We are able to quantify the amount of skewness by computing the data sample moments ratio known as the *Skewness* index which result to be 0.42, namely a right-skewed distribution, agreeing with the histogram depicted.

Afterwards, we investigate the temperature data. As for the rain data set, we have availability of the information from 1901 up to 2020. In figure 2 we depict the average annual temperature distribution.

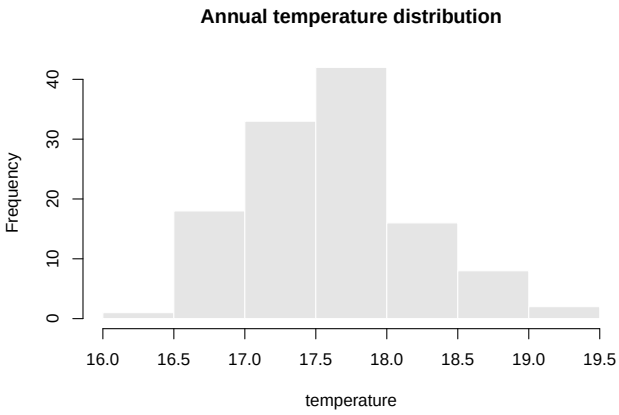


Figure 2: Annual temperature distribution

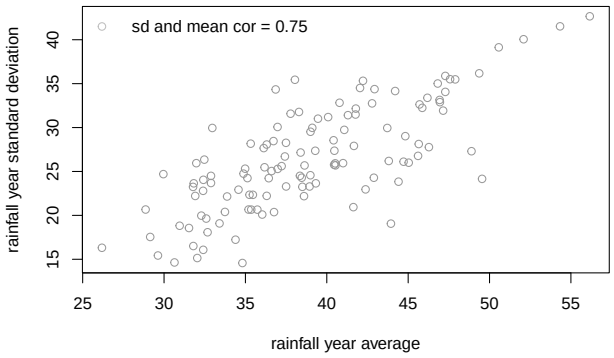


Figure 3: Rain average and standard deviation relation. The phenomenons present a correlation of $\hat{\rho} = 0.75$

We can clearly observe a considerable right skeweness indicating more extreme events such as sweltering days than very cold days. The phenomenon now takes values in the real domain, $x \in \mathbb{R}$, and presents a skewness index of 0.43, again, a right-skewsness indicator.

Climate change and its related phenomenon is believed not to depend strictly on the annual average value but on its distribution as well. For instance, we can think to a year which has had the same amount of rainfall as the previous one, but the concentration of the rainfall may assume very different values. We may observe several days with very low precipitation as well as very few days with extreme precipitations. This would result in the same average value (Alexander 1995). This is why, in figure 3 we present the relationship between the two summary statistics, mean and standard deviation, per year. We can clearly see that these are closely and positively related, meaning that for an higher average rainfall amount, we can expect higher variability in the phenomenon, with an estimated correlation of $\hat{\rho}_{sd,mean} = 0.75$. Therefore, for simplicity of the study, we have decided to use the average values as a proxy measure of the entire rainfall phenomenon.

On the other hand, this does not happen for the temperature value. In particular, in figure 4 can observe how the relation between the average annual temperature, figure a), and the annual increase, figure b), seem to be randomly related to the annual standard deviation. Therefore, we will investigate both of them.

In order to keep track the evolution of the phenomenon over time and to give an idea of the overall trend we make use of non parametric model estimation, meaning that we do not assume anything but a relation between $\mathbf{x}$ and $\mathbf{y}$ governed by a function $f(.)$. The method employed is the locally weighted regression discussed in section 2 for a certain value of the span s. In particular:

- 100% of the neighbor data to take into account the global context
- 50% of the neighbor data to allow for a more local structure

By investigating the results in figure 5 we can notice how the global fit to the rain data shows a constant trend up to the 1950 which evolves in a slightly linear decreasing trend onward, plot a). On the other hand, the local fit shows an overall constant trend up to the 1980s which evolves in a steeper decrease for the later years, b). Nevertheless, both the models suggest an evolution of the amount of rainfall which negatively changes over time.

Afterwards, we fit a the same two models for the temperature data. Here the relation with time is much clearer. For instance, we can observe how the global fit shows a linearly increasing trend over time, with constant velocity, c). In contrast, the local fit shows an increasing trend with different intensities in different periods allowing for non linear relationship d). Nevertheless, the overall trend does not change dramatically. The annual average temperature has clearly been increasing over the last century.

Then, we investigate the annual temperature increment. In figure e) with span size $s = 1$ we can see a flat trend up to the last half century, and then an increasing trend, indicating more positive values than negative ones. In figure f) with span size $s = .5$ we

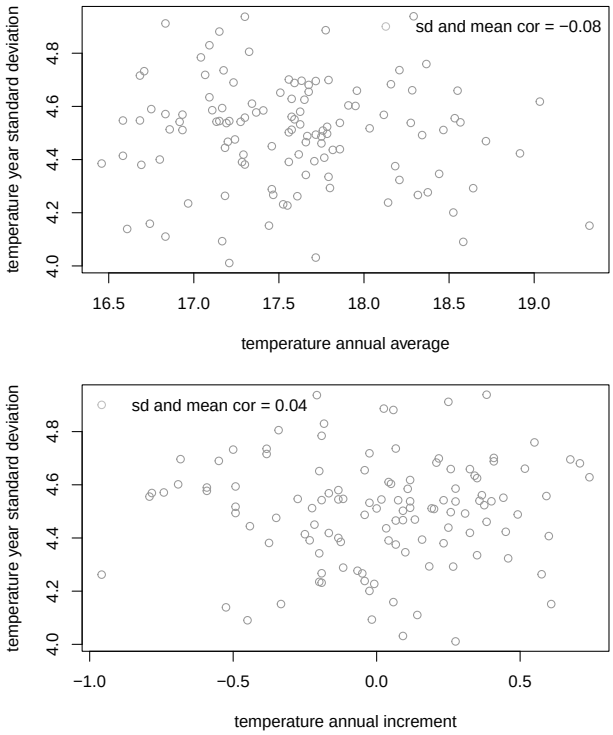


Figure 4: a) annual average temperature vs temperature standard devation; b) annual temperature increase vs temperature standard deviation

can see an initial slightly increasing trend, followed by a fairly negative period during the 50s, followed by a clearly steep increasing trend from the 60s onward. The results are approximately the same, the growth in tempeature over the last decades is clearly positive.

	year	rain	temp	temp_sd
year	1.00	-0.06	0.82	-0.09
rain	-0.06	1.00	-0.33	-0.23
temp	0.82	-0.33	1.00	-0.08
temp_sd	-0.09	-0.23	-0.08	1.00

Table 1: Climate correlation table

In order to give an overall idea of the data relationship we are investigating we would like to present the pairs plot showing each pair combination plus the estimated correlation matrix $\mathbf{R}$ in table 1. As mentioned earlier, in figure 6 we can notice the strong relationship between temperature and year $\hat{\rho}_{t,y}$ = 0.82 and a weaker relationship between rain and year $\hat{\rho}_{r,y}$ = -0.06 . However, we can see a clear relationship between year and rain $\hat{\rho}_{r,t}$ = -0.33 which may be worth further investigation. It is also interesting to notice that the temperature standard deviation is negatively related to the rainfall amount $\hat{\rho}_{r,t_{sd}}$ = -0.23, meaning that for a year with low average rainfall, we can expect slightly higher variation in the average temperature.

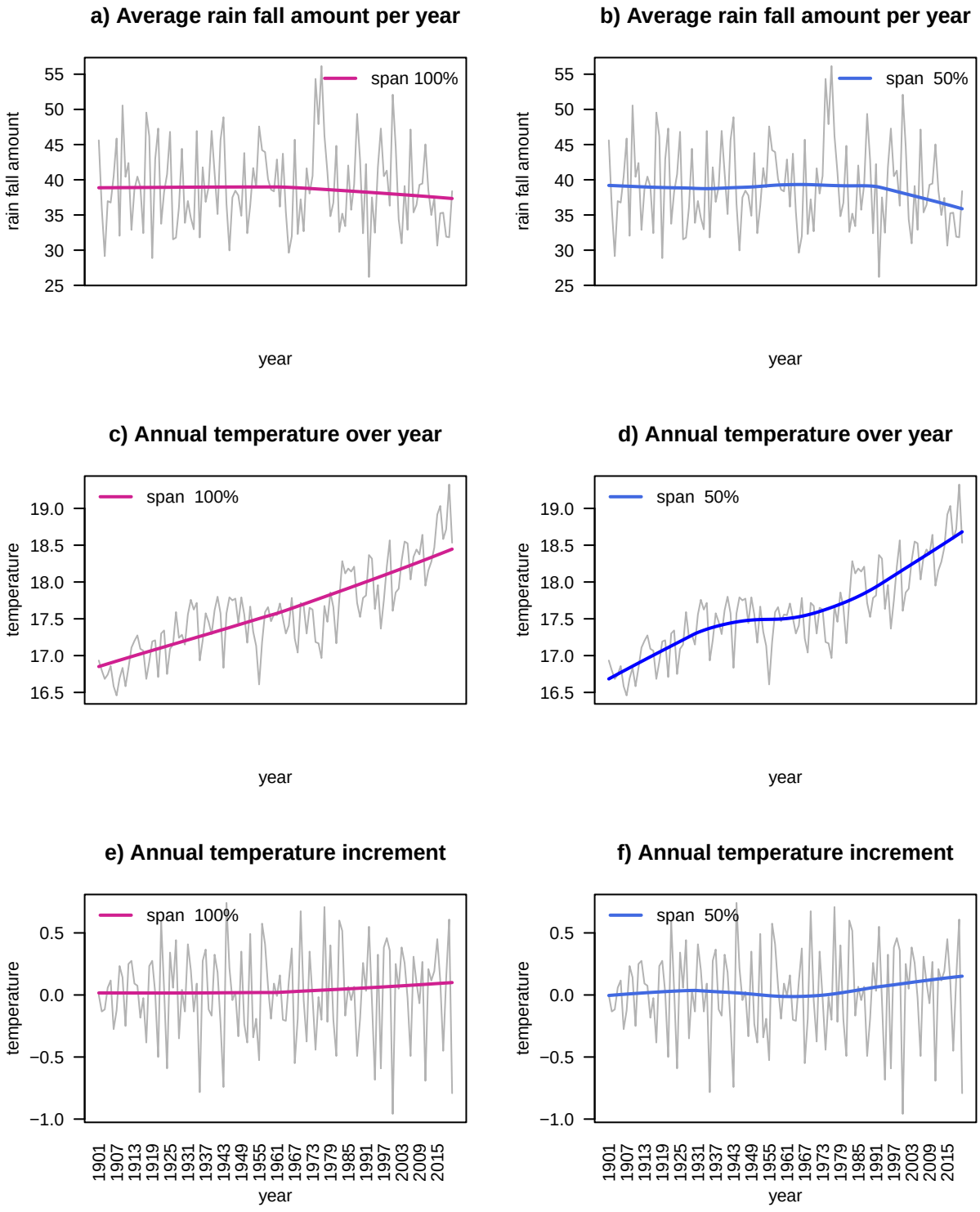


Figure 5: Local weighted regression fit to the data: a) b) annual rainfall amount, c) d) annual temperature, e) f) annual temperature increment temperature increment

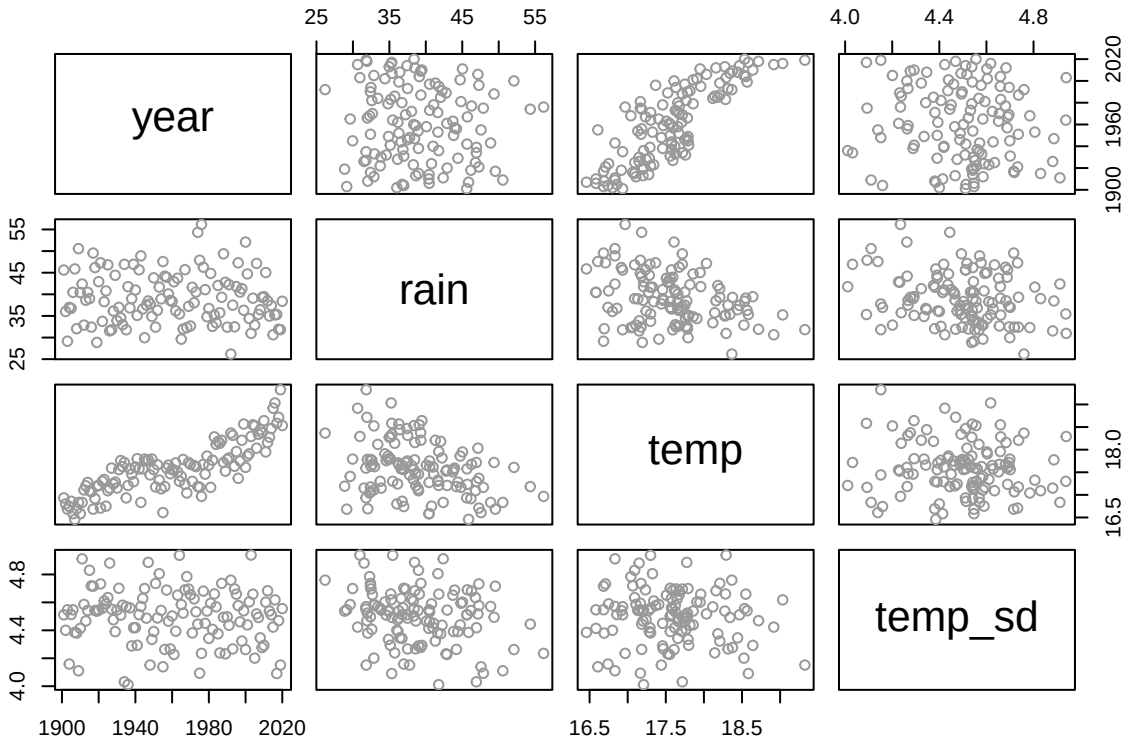


Figure 6: Year, Rain, Temperature scatterplot

4 Climate Modeling

Tempeature evolution

We are now going to propose a very simple physical model to describe the temperature behavior. Ascertained that temperature heavily evolves over time, we present a very simple ordinary differential equation given the temperature τ and time t

$$\frac{d}{dt}\,\tau = \gamma\,\tau$$

which posit a linear evolution of the phenomenon over time. We have seen in the previous section that the relation of interest may not be linear. However, we know that non-linear models may perform wiggly in extrapolation context, therefore we will assume, for simplicity, a linear relationship. Moreover, we add initial condition equal to the average temperature values of the first 5 years. The reason behind this choice is that, within this project, we do not have access to older data and during the exploratory data analysis in section 3 we have seen how the phenomenon is highly variable. Therefore, we choose the mean as initial condition

$$f(\tau, t=0) = \tau_{1:5} = 16.8$$

Given the aforesaid model, we are interested on the values the parameter γ might take. In particular, we want to tune it to find the best match to the observed data and we are going to do so exploiting the Levenberg-Marquardt algorithm discussed in section 2. The algorithm applied to the data set of interest returns a point estimate of $\hat{\gamma}$ =7.6528×10^{-4} and an estimated standard error of $\hat{SE}(\hat{\gamma})$ =2.43×10^{-5}, resulting in the 95% confidence interval

$$P(\gamma \in [0.00071764, 0.00081291]) = 0.95 \qquad (1)$$

which does not contain the null value, hence statistically significant at a 5% level. Overall, we can claim that, according to the model of interest, the global temperature increases over time by a factor within the region in (1). We graphically present the estimated model in figure 7. We can clearly notice how the model may underfit the data. This is due to the fact that we have assumed a linear relationship for robustness in the extrapolation context.

We have now available a model which let us understand the evolution of the temperature over time and may lead us to future prediction. In particular, we

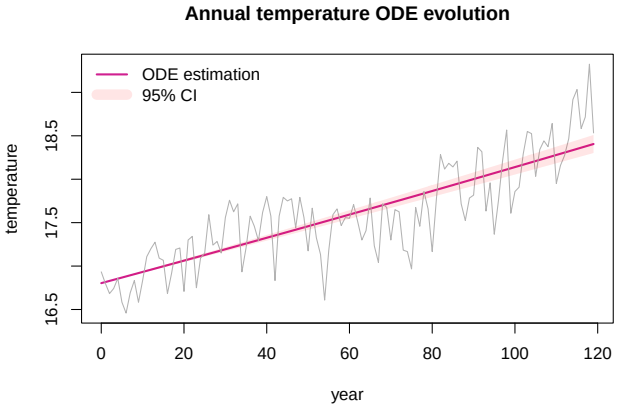


Figure 7: Annual temperature modeled by Ordinary Differential Equation

are interested in what would happen in a 50 years time following the current trend. According to the estimated model and depicted in figure 8, the global temperature may continue to rise up to reaching 19 Celcius average degree in 50 years time, with a reasonably tight confidence interval. Moreover, here we can appreciate how stable the prediction is, due to the linearity assumption.

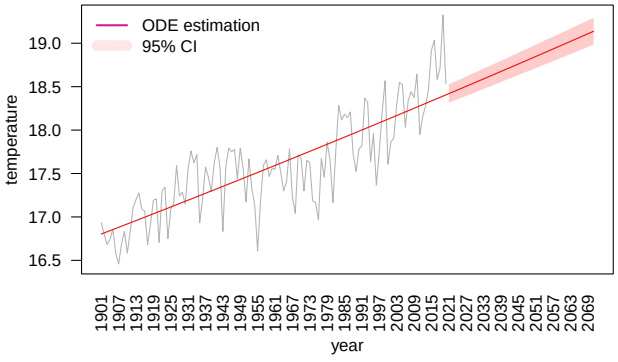


Figure 8: 50 out of sample years temperature prediction

A fundamental requisite for a mathematical model is its robustness with respect to its input. Here, the data is empirically observed and the parameter γ is estimated from it. The question we want to address is whether we can rely on this parameter and whether the model output would change if the parameters in input changed, ie for small perturbations in the data. In order to tackle this question, a common and reliable choice is local sensitivity analysis discussed in section 2. Given the model

$$f(\tau) = \gamma\,\tau$$

the method allows us to derive the sensitivity equation as

$$\begin{aligned}\frac{d}{dt}s &= \frac{d}{d\tau}f\ s + \frac{d}{d\gamma}f \\ &= \gamma s + \tau\end{aligned} \tag{2}$$

Therefore, we will need to deal with the system of ordinary differential equation

$$\begin{cases}\frac{d}{dt}\tau = \gamma\tau \\ \frac{d}{dt}s = \gamma s + \tau\end{cases} \tag{3}$$

As done previously, we numerically integrate the system of *ODEs* and inspect the result. In particular, we want to investigate what would happen with different values for the input parameter γ. In this setting, we try to evaluate the model output for values which are 5 times the estimated standard error away from the point estimate, $\hat{\gamma} \pm 5 \times \hat{SE}(\hat{\gamma})$. In figure 9 we can observe the system output in log-scale. In the first place, plotting the temperature values τ against time t in log-scale, we do not observe any appreciable difference for such an extreme shift of the parameter considered. In the second plot, we depict the sensitivity value s against time t and, as before, we cannot notice any significant difference. This analysis clearly states how the model of interest is robust to fairly small perturbations of the input data. Thus, we can rely on its inference.

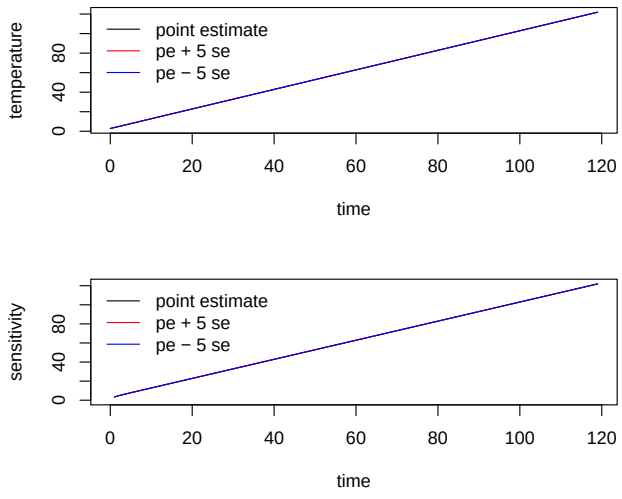


Figure 9: Evolution of temperature in time with input parameter $\pm$ 5 $\hat{SE}(\hat{\gamma})$; b) evolution of γ parameter in time $\pm$ 5 $\hat{SE}(\hat{\gamma})$

Rain-temperature model

Once ascertained the robustness of the temperature model, we are ready to investigate the relationship between the latter and the rainfall amount. In order to do so, we estimate a linear regression model exploiting the *GLM* theory presented in section 2, where we add a non-linear component derived by the spline theory leading to a *GAM* which allows us to write

$$\textbf{rain} = f(\tau) + \epsilon \quad \text{where } \epsilon \sim N(\mu, \sigma^2)$$

Here, the spline degree has been empirically chosen by making usage of the BIC, a penalized likelihood criteria, which is optimized for $K = 2$. From the results inspection, the first trial is not satisfactory as the model presents some deficiencies. In particular, the homoskedasticity assumption is violated, meaning that the variance is not constant over the range of values for the input space. In order to address the problem, we perform an iterative weighted least square estimation. By doing so, we drastically reduce the parameters estimated standard error, as shown in table 2, and gain a considerable 29% reduction in the estimated BIC values.

	OLS	WLS
$\hat{SE}(\hat{\beta}_0)$	1.72897239	0.00008380
$\hat{SE}(\hat{\beta}_{1,\tau})$	3.65154399	0.00015488
$\hat{SE}(\hat{\beta}_{2,\tau})$	2.96582256	0.00001818

Table 2: OLS and WLS parameters estimated standard error

The model parameters results are summarised in table 3. Unsurprisingly, from those we can observe how both the estimated spline coefficients have negative sign with very tiny estimated standard error. Therefore, we can appreciate how, up to 8 decimal precision, the returned P-value is 0, meaning that we reject the null hypothesis $H_0 : \ \beta_i = 0$ in favour of the alternative $H_1 : \ \beta_i \neq 0$ for $i = \{0, 1, 2\}$. This means that, according to the estimated model, the relation between temperature and rainfall amount is negative. With higher temperature we can expect to observe lower amount of rainfall and vice versa.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
$\hat{\beta}_0$	41.23907927	0.00008380	492090.20447958	0.00000000
$\hat{\beta}_{1,\tau}$	-7.90378231	0.00015488	-51030.55292079	0.00000000
$\hat{\beta}_{2,\tau}$	-8.53033020	0.00001818	-469180.40392581	0.00000000

Table 3: Weighted Least Square model estimation for temperature and rainfall amount

We now need to evaluate the robustness of the model with respect to the parameters in input. As we did previously, we will now try two configurations where we shift the *linear model* parameters estimates by 5 times their standard deviation, $\hat{\beta}_i \pm 5\hat{SE}(\hat{\beta}_i)$. The results are depicted in figure 10. Here we can notice that, despite of the fact that the different scenarios lead to different intensities, the estimated confidence regions intersect each other, leading to an overall similar trend. Therefore, we can claim that the model is robust with respect to the data and parameters in input and that, considered the worldwide temperature change, the rainfall amount is consequently decreasing over time.

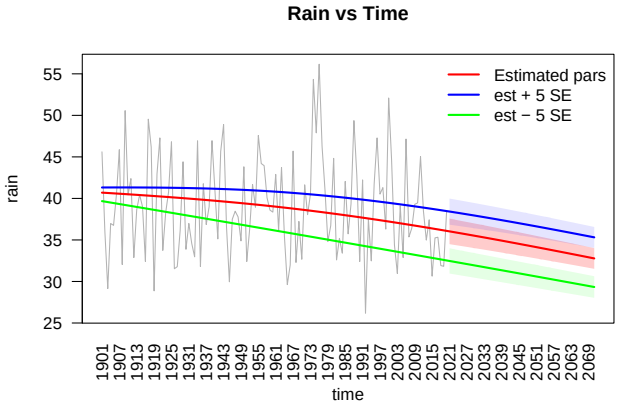


Figure 10: Evolution of rainfall w.r.t. temperature with input parameter $\pm$ 5 $\hat{SE}(\hat{\gamma})$

5 Species count Modeling

Data Pre Processing

In the previous section we have presented the data and respective results on the climate phenomenons including temperature and rainfall information from 1901 up to 2021. In this section, we are going to present the waterbirds species count data over time with the goal to identify the relation between its evolution and the covariates explored earlier. In particular, the data set of interest provide the species individual count per year from 1975 to 2021. Within this data set we observe monthly variation which, for simplicity and consistency with the climate data, will be averaged over the different years. For different species and genes, we have different number of observation and, as we observe up to 103 species and genes combinations, we decide to keep only those whose numerousness can provide robust results, from $n_i = 30$ onward, as per figure 11.

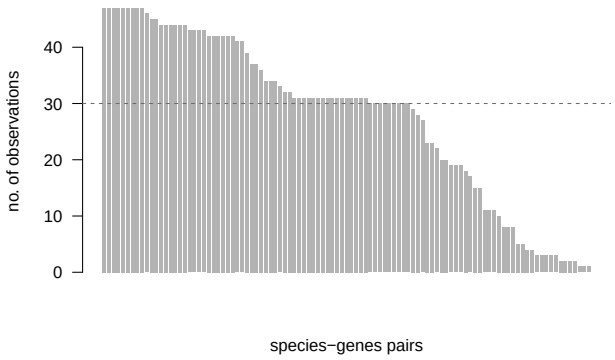


Figure 11: Yearly number of observations per species-genes pair

Once selected the species of interest, we are interested in the overall trend evolution in time. Beside a descriptive analysis, we provide a non-parametric fit made possible by a local weighted regression to give an idea of the overall trend. As the values variance is appreciable large, we provide either mean (red line) and median (blue line) value for robust results. Despite being on different scales, both the results shows similar global trend. The waterbirds individual count has been decreasing over time. This behavior is depicted in figure 12.

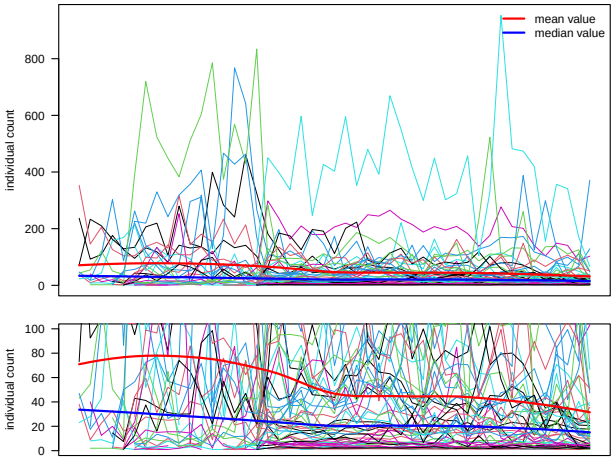


Figure 12: Yearly individual count per species-genes pairs plus mean (red) and median (blue)

Once ascertained the decreasing trend, the next step is to relate the individual count data to the climate information presented earlier. In the first place, we inspect the response distribution, as shown in figure 13. As the data takes value on a discrete and positive

set, it is reasonable to assume a *Poisson* distribution of the residuals

$$\epsilon \sim Poi(\lambda)$$

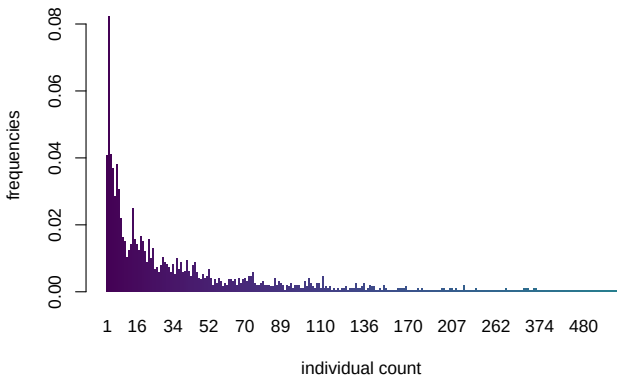


Figure 13: Distribution for animal counts

Data Modeling

We now want to estimate a model describing the evolution of the population w.r.t. the climate input feature. In order to do so, we need to employ mixed-effect models, discussed in section 2, since its theory allows us to include data from different classes, namely their species, and follow them longitudinally taking into account the within correlation.

The model to be estimated is of the form

$$\mathbf{count}_i = \alpha_{1,i}\ \mathbf{Z}_i + f(\mathbf{X}) + \epsilon$$

where $\epsilon \sim Poi(\lambda)$ and $\alpha_{1,i}$ represents the random intercept. The same can be said for the random slope related to the temperature, In fact

$$f(\tau) = \sum_{k=1}^{K+M} \beta_k\ g_k(\tau) + \alpha_{2,i}\ Z_{ij}\ \tau$$

And positing a distribution on the random terms $\alpha_i \sim N(\mu_{\alpha_i}, \sigma^2_{\alpha_i})$ indicates a random intercept and temperature slope which contribution depends on the $\mathbf{Z} \in \mathbb{N}^{n \times p}$ matrix, where p is the number of species-genes pairs. The matrix indicates the species belonging

$$z_{ij} = \begin{cases} 1 \text{ if observation i is in species j} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

thus varying the intercept of the model according to the species of interest.

The spline degree has been empirically chosen by making usage of the BIC measure which aims to maximize the likelihood while taking into account a penalization for the number of parameters estimated. The optimal value has shown to be $K = 5$ with the following variables: *average rainfall* and *average temperature*. The inclusion of *temperature standard deviation* has not improved the model significantly, hence it has been excluded.

Since we are including different predictors in the model which vary on different scales, the model estimation may be unstable, therefore, we decide to apply a pre-processing step which standardize the variables to have mean zero and variance one

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

From the model results presented in table 4, we can observe how the coefficients related to the *year*, *rain*, and τ marginal terms $\hat{\beta}_{i,year}$, $\hat{\beta}_{rain}$, $\hat{\beta}_{\tau}$ are mainly positive. However, the model interaction terms, $\hat{\beta}_{i,year \times \omega}$ are mainly negative. Moreover, the majority of the estimated standard error is very small compared to the point estimate. This returns an high value of the $z - score$ which implies the coefficient to be statistically significant. All the estimated coefficients are part of a global non-linear function of the predictors which have been standardized. Hence, in order to comment on the results, we need to consider the features simultaneously. We will do it later in the report, after the model will have been validated, by showing its practical application to the data of interest.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	2.829537	0.198497	14.254789	0.000000
$\hat{\beta}_{1,y}$	0.008287	0.043617	0.189990	0.849317
$\hat{\beta}_{2,y}$	0.425417	0.057715	7.371046	0.000000
$\hat{\beta}_{3,y}$	0.045776	0.036597	1.250793	0.211010
$\hat{\beta}_{4,y}$	0.722263	0.114729	6.295396	0.000000
$\hat{\beta}_{5,y}$	-0.807545	0.045591	-17.712793	0.000000
$\hat{\beta}_{\tau}$	0.361911	0.056974	6.352164	0.000000
$\hat{\beta}_{rain}$	0.208523	0.025575	8.153318	0.000000
$\hat{\beta}_{1,y\times\tau}$	0.120815	0.039645	3.047392	0.002308
$\hat{\beta}_{2,y\times\tau}$	-0.817581	0.057604	-14.193191	0.000000
$\hat{\beta}_{3,y\times\tau}$	-0.368144	0.042014	-8.762385	0.000000
$\hat{\beta}_{4,y\times\tau}$	-1.072883	0.104063	-10.309945	0.000000
$\hat{\beta}_{5,y\times\tau}$	-0.084810	0.045199	-1.876369	0.060605
$\hat{\beta}_{1,y\times rain}$	0.056719	0.026509	2.139634	0.032384
$\hat{\beta}_{2,y\times rain}$	-0.435746	0.038606	-11.287067	0.000000
$\hat{\beta}_{3,y\times rain}$	-0.614404	0.041138	-14.935070	0.000000
$\hat{\beta}_{4,y\times rain}$	-0.755862	0.069450	-10.883571	0.000000
$\hat{\beta}_{5,y\times rain}$	0.070351	0.059929	1.173905	0.240433

Table 4: Mixed effects model of the individual count on the temperature and rainfall data results

Model Diagnostics

For the model of interest, we have assumed the error distributed as a *Poisson*, therefore, in the fitted vs residuals inspection, we run into a different behavior. In particular, in figure 14 we can observe how the majority of the fitted values are very small number close to zero with very few extreme values. Moreover, we can also observe how the variances grows with the magnitude of the fitted values. This behavior is expected and this diagnostic produces a satisfactory results, hence the model is appropriate to describe the data.

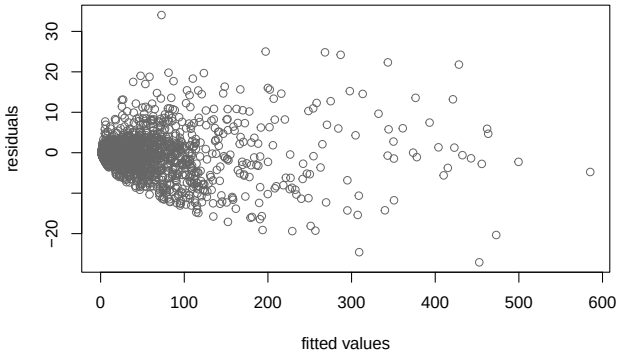


Figure 14: Distribution for animal counts

With regard to the random effects, we have assumed a *Normal* distribution for each of these. In figure 15 we can inspect the results and notice the following:

- for the random intercept the assumption does not hold much as we can notice a considerable discrepancy between the observed and theoretical quantiles, figure 15 a). However, the intercept related variance is $\hat{\sigma}_0^2 = 1.84571$, which counts for the majority of the total variance, approximately 63%, as we can see in table 5. For instance, in figure 16 we can observe a very wide range of variation for the random intercept estimation with very few species around the zero and all the other with 95% confidence interval which does not contain the null values, hence statistically significant.
- on the other hand, for the random slope, the empirical distribution agrees with the theoretical one in the central area, but it shows deficiencies in the tail regions, figure 15 b). Its contribution to the overall variance is much lower than the previous, approximately 2%, table 5, but still worth it since it allows for different effects on different species. With regard to the single term values, we can observe a much tighter range of variation, approximately $[\pm 1]$. Nevertheless, many species-genes slopes are estimated to be singificantly different from zero, hence worth including them.

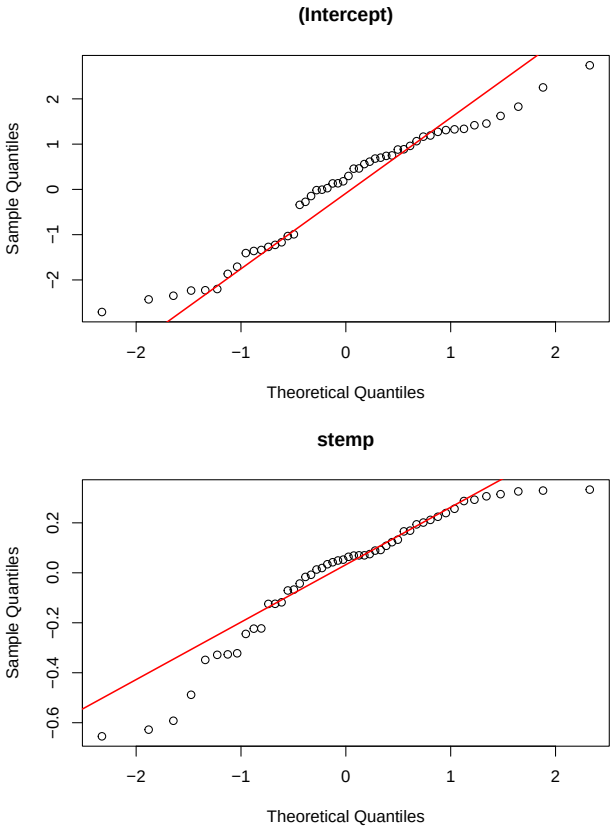


Figure 15: a) Intercept random terms qqplot; b) temperature slope andom terms qqplot

	intercept	temperature	residuals
σ_i^2	1.84571	0.06743	1.0000
σ_i^2/σ_Y^2	0.6336	0.0231	0.3433

Table 5: Random effect variance contribution

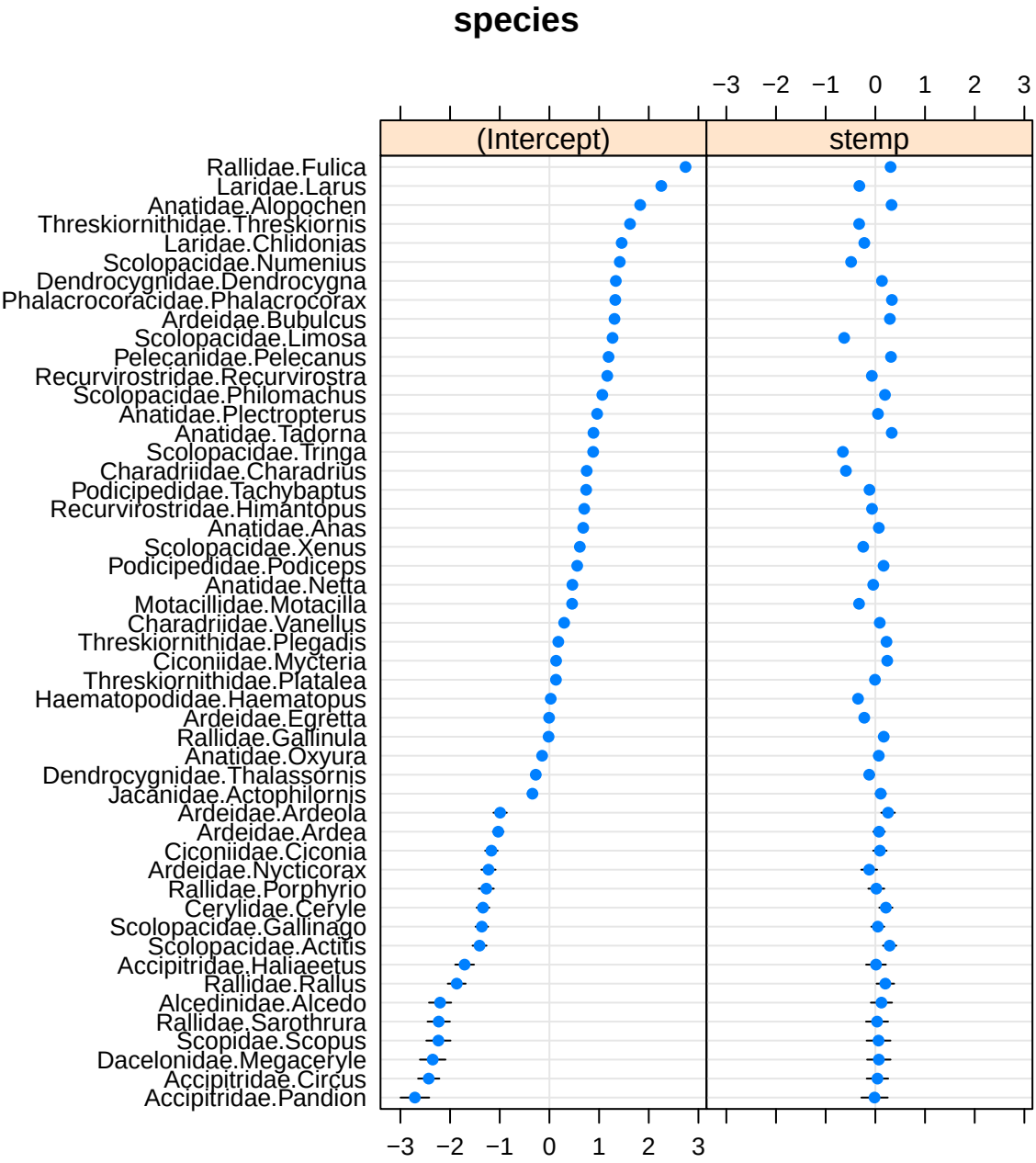


Figure 16: Individual random effect for each species

Future scenario prediction

So far, we have estimated a random effect model to allow for the prediction of the response variable, individual count, by making usage of non-linear relationships with the climate phenomenon.

We now want to present these result for each of the species-genes pairs and compare them with the observed values. Moreover, we provide estimated confidence interval for the next 50 years and what might happen to these species, following the current trend. All the results we are going to describe are depicted in the figures in the appendix. In particular, we can observe that for some of these, the model fits very well the observed data and its trend evolution, see species-genes pairs $\{4, 6, 19, 25, 36, 38\}$, while for others, the fit is not suitable at all, see species-genes $\{24, 39, 45\}$. Nevertheless, excluded the most extreme cases, we can see that the model predicted trends agree with the overall wildlife evolution. Here, it is clear that as time goes on, temperature raises and rainfall decreases, the overall trend, despite some non-linear behavior, is clearly decreasing. In particular, for each of these species, we can notice how easy it is, according to the estimated model, to reach the context where only one individual will be left, in the lower bound CI, which would not allow for the species reproduction and, consequently, survival. According to the estimated model, which has shown to capture the overall data structure, this might happen as soon as in 15 years time.

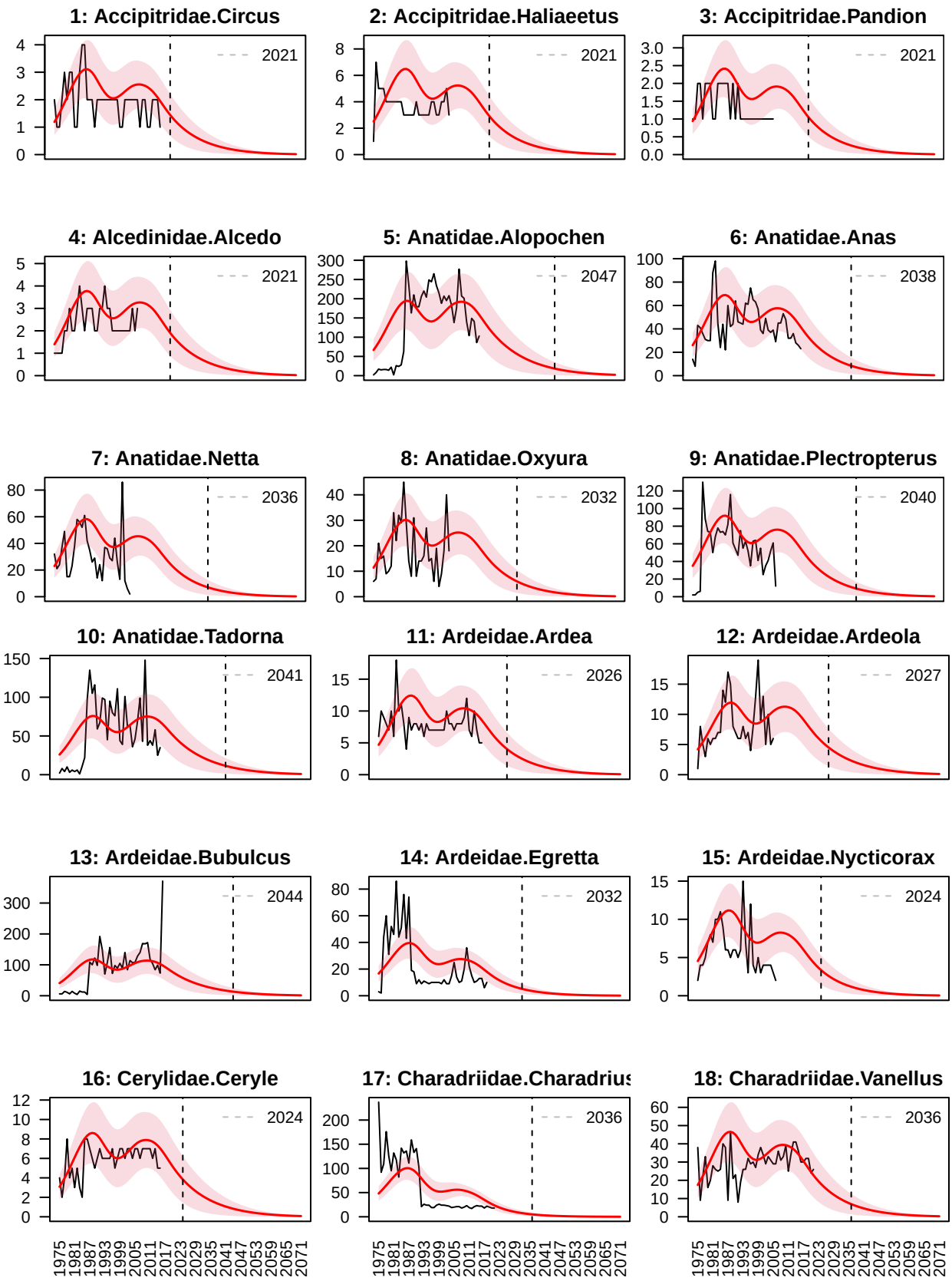
6 Conclusion

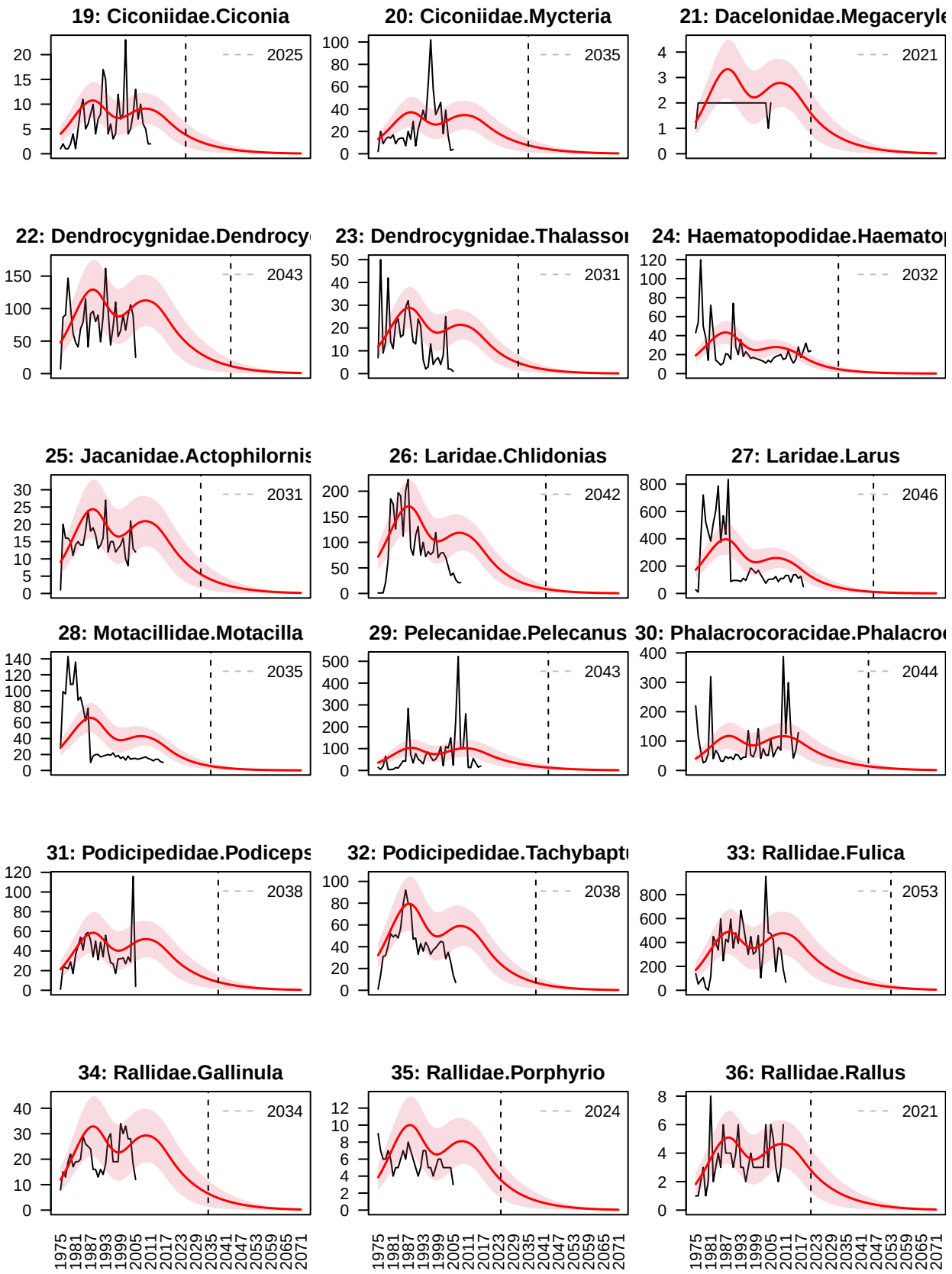
During this short study, we have expressed our interest in the climate change phenomenons and its dramatic consequences to the wildlife species survival. In particular, in the first place, we have explored the rainfall amount and temperature trend evolution over the last century, from 1901 up 2020. The data investigated have clearly shown how the time is related to the temperature change and how the temperature is related to the rainfall amount. This has allowed us to model the temperature as an ODE and computationally optimize the parameter which governs the relation. The estimated confidence region has been estimated to be $P(\gamma \in [0.00071764, 0.00081291]) = 0.95$, which does not contain the null value and clearly states the temperature raising. Afterwards, the estimated model has been employed to produce estimation for rain and temperature within the following 50 years, up to 2071. This data has been very useful to provide a robust fit of the rainfall amount in relation to the temperature. The result, as expected, have indicated a negative relationship between the two phenomenons. With higher temperature we can expect lower amount of rain, and vice versa. As these computational models mainly depends on the data in input, it is fundamental to run a sensitivity analysis to quantify the uncertainty. Both the models, have shown to be robust and produce a fairly overall similar results up to a perturbation of 5 times the estimated standard error coefficients.

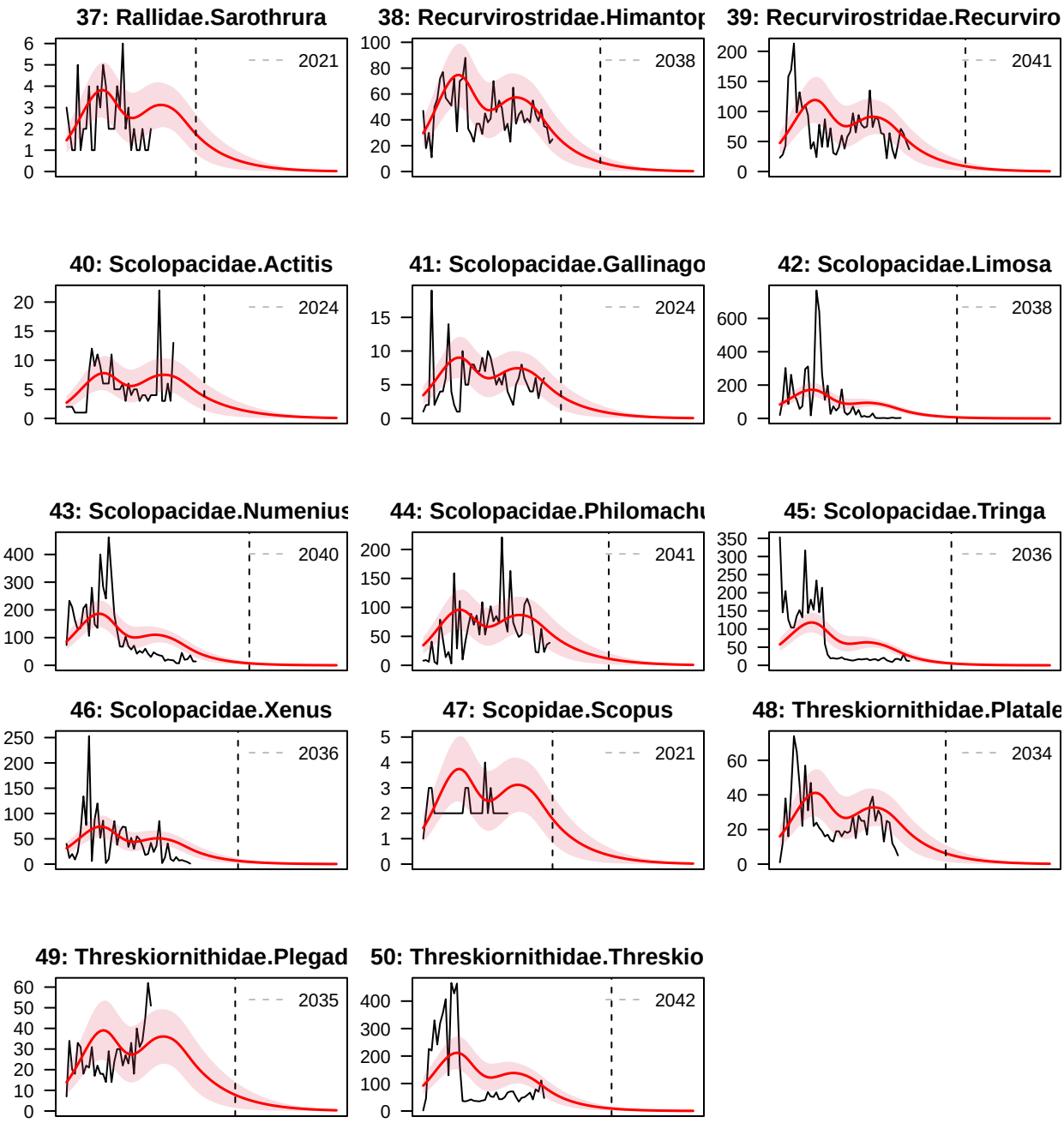
Afterwards, we have focused our attention on many different South Africa waterbirds species-genes pairs and their evolution over time from 1975 up to 2020. Not all of them were observed every year and, since there were several available, we have decided to keep only those which numerousness could provide robust estimations and results, namely those which had more than 30 observations. In contrast to the $rain - temperature$ model, here we have had to deal with multiple input feature which take values on different scales, therefore, we have performed a standardization on the predictors for a reliable estimation process. In order to allow for the multiple species-genes modelling, we have exploited the mixed-effects model theory, which takes into account for the intra-class correlation. This has allowed to us posit a random effect on either the intercept term and the temperature feature, so that each species-genes pairs has had its own behavior. As previously, the model has been validated by a set of diagnostics on the residuals and the random-effect terms.

In the end, the model has been applied to the data of interest and we have been able to see that in most of the cases, the result approximated the reality in a satisfactory manner. We have also noticed that, in the long term behavior prediction, these species individual count trend is clearly decreasing. This means that, according to the circumstances we have been being in the last 45 years, these species may not survive long. The unit threshold was hit by the lower confidence bound in 15 years time, on average. This means that, according to the estimated model, very soon there will be only one individual left for the species-genes animal and it won't be able to reproduce itself anymore, hence quickly leading to the species-genes extinction.

7 Appendices


1: Accipitridae.Circus
2021
2: Accipitridae.Haliaeetus
2021
3: Accipitridae.Pandion
2021
4: Alcedinidae.Alcedo
2021
5: Anatidae.Alopochen
2047
6: Anatidae.Anas
2038
7: Anatidae.Netta
2036
8: Anatidae.Oxyura
2032
9: Anatidae.Plectropterus
2040
10: Anatidae.Tadorna
2041
11: Ardeidae.Ardea
2026
12: Ardeidae.Ardeola
2027
13: Ardeidae.Bubulcus
2044
14: Ardeidae.Egretta
2032
15: Ardeidae.Nycticorax
2024
16: Cerylidae.Ceryle
2024
17: Charadriidae.Charadrius
2036
18: Charadriidae.Vanellus
2036
1975 1981 1987 1993 1999 2005 2011 2017 2023 2029 2035 2041 2047 2053 2059 2065 2071


19: Ciconiidae.Ciconia
2025
20: Ciconiidae.Mycteria
2035
21: Dacelonidae.Megaceryle
2021
22: Dendrocygnidae.Dendrocy
2043
23: Dendrocygnidae.Thalassor
2031
24: Haematopodidae.Haemato
2032
25: Jacanidae.Actophilornis
2031
26: Laridae.Chlidonias
2042
27: Laridae.Larus
2046
28: Motacillidae.Motacilla
2035
29: Pelecanidae.Pelecanus
2043
30: Phalacrocoracidae.Phalacroc
2044
31: Podicipedidae.Podiceps
2038
32: Podicipedidae.Tachybapt
2038
33: Rallidae.Fulica
2053
34: Rallidae.Gallinula
2034
35: Rallidae.Porphyrio
2024
36: Rallidae.Rallus
2021
1975 1981 1987 1993 1999 2005 2011 2017 2023 2029 2035 2041 2047 2053 2059 2065 2071


37: Rallidae.Sarothrura
2021
38: Recurvirostridae.Himantop
2038
39: Recurvirostridae.Recurviro
2041
40: Scolopacidae.Actitis
2024
41: Scolopacidae.Gallinago
2024
42: Scolopacidae.Limosa
2038
43: Scolopacidae.Numenius
2040
44: Scolopacidae.Philomachu
2041
45: Scolopacidae.Tringa
2036
46: Scolopacidae.Xenus
2036
47: Scopidae.Scopus
2021
48: Threskiornithidae.Platale
2034
49: Threskiornithidae.Plegad
2035
50: Threskiornithidae.Threskio
2042

References

Aerts, Marc, Gerda Claeskens, and Matthew P Wand. 2002. "Some Theory for Penalized Spline Generalized Additive Models." *Journal of Statistical Planning and Inference* 103 (1-2): 455–70.

Alexander, WJR. 1995. "Floods, Droughts and Climate Change." *South African Journal of Science* 91 (8): 403–8.

Gbur, Gregory J. 2011. *Mathematical Methods for Optical Physics and Engineering*. Cambridge University Press.

Global Biodiversity Information Facility, GBIF -. n.d. "Free and Open Access to Biodiversity Data." www.gbif.org/occurrence/.

Grayson, Michelle. 2013. "Agriculture and Drought." *Nature* 501 (7468): S1–1.

Group, World Bank. n.d. "Climate Change Knowledge Portal." https://climateknowledgeportal.worldbank.org/download-data.

Hastie, Trevor J, and Robert J Tibshirani. 1990. "Generalized Additive Models, Volume 43 Of." *Monographs on Statistics and Applied Probability* 15.

Hillman, Jesse C, and Alison KK Hillman. 1977. "Mortality of Wildlife in Nairobi National Park, During the Drought of 1973–1974." *African Journal of Ecology* 15 (1): 1–18.

Hooper, Peter M. 1993. "Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models." *Journal of the American Statistical Association* 88 (421): 179–84.

McCullagh, Peter, and JA Nelder. 1989. "Generalized Linear Models II." Chapman; Hall, London.

Moré, Jorge J. 1978. *The Levenberg-Marquardt Algorithm: Implementation and Theory*. Springer.

Morio, Jérôme. 2011. "Global and Local Sensitivity Analysis Methods for a Physical System." *European Journal of Physics* 32 (6): 1577.

Parry, Martin L, Timothy R Carter, and Mike Hulme. 1996. "What Is a Dangerous Climate Change?" *Global Environmental Change* 6 (1): 1–6.

Rouault, Mathieu, and Yves Richard. 2003. "Intensity and Spatial Extension of Drought in South Africa at Different Time Scales." *Water Sa* 29 (4): 489–500.