South Africa drought and wildlife survival

Andrea Corrado 20205529

Contents

Contents

Abstract

1 Introduction

During 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 human society is contributing significantly to worsen the condition the earth has to suffer [**?**]. There have been several studies examining and reporting the potential consequences of these phenomenon. For instance, the temperature raising, changes in rainfall amount and greenhouse gases emissions have been of particular interest. Since last century, we have witnessed a increasing frequency in the number and severity of droughts in particular territories such as South Africa SA. Due to the fact that area such as SA suffer from low economic power, a drought may drive towards dramatic consequences. It can affect either agriculture, for a period that lasts up to 6 months, either hydrological up to 24 months [**?**].

In literature, we can find several different examples of methods to reduce the impact of droughts on the agriculture field ie. by growing plants restraint to higher temperature or computational model to predict droughts and take actions before drastic events [**?**]. However, in this study, we wish to focus our attention on the drought consequences on wildlife population evolution. For instance, different studies have shown as animals mortality grows during periods with the highest temperature within the driest areas [**?**]. We will expand this idea to more recent data about waterbirds in South Africa [**?**]. This data set 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 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 the relation with other factors. In particular the data we are going to analyse concern the global temperature and rain fall amount [**?**] 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 predict how, following the current trend, these would evolve over time and the potential consequences these could drive towards.

In section 2 we give a brief introduction to the methodology used within this research project. In section 3 we present the climate data with an exploratory data analysis. In section 4 we model the climate data and provide prediction on a possible future scenario. In section 5 we model the species count in relation to the climate data. In section **??** we employ the estimated model within a specific specie and provide a practical example. In section 7 we discuss the finding of the project.

2 Technical Background

2.1 Generalized Lienar Model

For the purpose of the analysis we conduct within this study,we are going to employ different methodologies. In particular, we will related the temperature data to the rain fall amount. In order to do so, we will employ the Generalized Linear Model [**?**] which theory would allow to establish a linear relationship between the quantity of interest (rain) $\mathbf{y}$ and a matrix of covariates $\mathbf{X}$ (temperature) 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 its inverse 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 validate by a set of diagnostic on the residuals $\mathbf{Y} - \hat{\mathbf{Y}}$ which provide meaningful insight on the goodness of fitness.

2.2 Generalized Additive Model

The *GLM* [?] 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 [?] 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 is completely free and up to the researcher. A very common and flexible choice in *GAMs* is to define the function $f(.)$ to be a spline *spline* [?], 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.*

2.3 Weighted Least Square

GLMs are often estimated making use 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 [?]. This approach allows us to iteratively estimate a weight matrix $\mathbf{W}$ to be employed in the model estimation as a measure of the importance for each unit $\mathbf{x}_i$ $i = \{1, \ldots, N\}$ 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.

2.4 Local Regression

2.5 Mixed Effect models

2.6 Ordinary Differential Equation

To track the evolution of a phenomenon over time, it is natural to think about *Oridnary Differential Equation (ODE)* theory [?] which allows us to keep track of the change of a quantity in continuous time

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

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

2.7 Levenberg-Marquardt algorithm

Very often, the function defined within an ODE is characterized by a set of parameters β. In empirical studies, we often 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 β so that $y = f(\mathbf{x}, \beta)$ best fits the data of interest. The *Levenberg-Marquardt* algorithm [?] 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.

2.8 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 [**?**] to evaluate the robustness of the estimated models.

3 Exploratory Data Analysis

3.1 Rainfall amount

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.

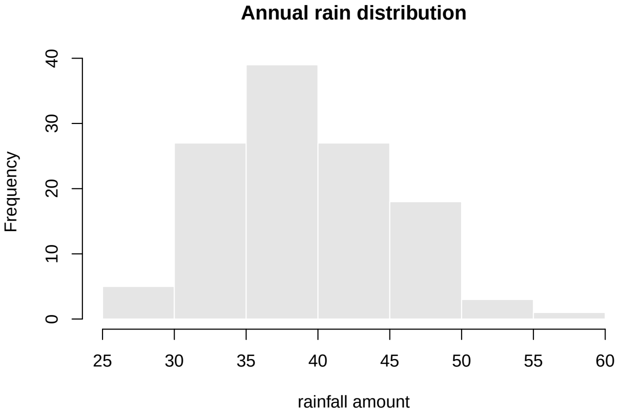


Figure 1: Yearly rainfall distribution

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, 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 **??** we depict the average annual temperature distribution.

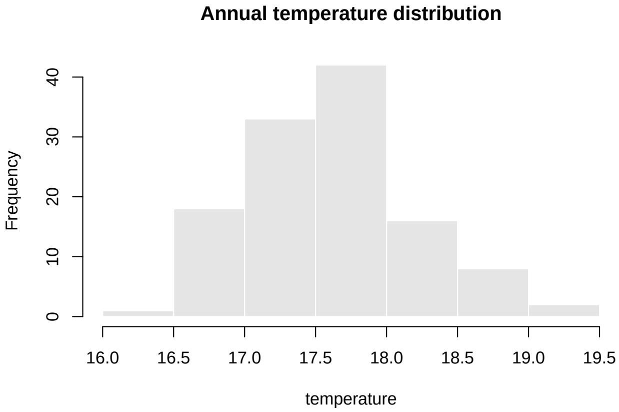


Figure 2: Annual temperature distribution

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 present a skewness index of 0.43.

Climate change and its related phenomenon is believed not to depend strictly on the annual average value but on its distribution. 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 [**?**]. This is why, in figure **??** 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.

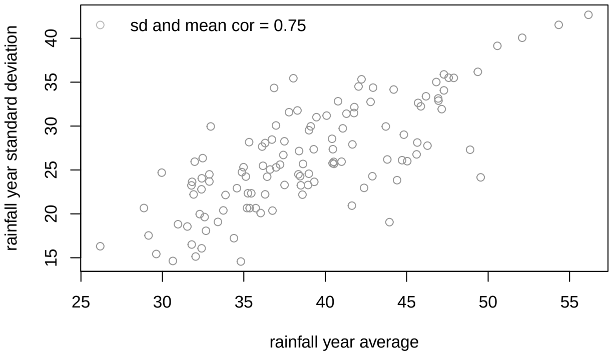


Figure 3: Rain average and standard deviation relation. The phenomenons present a correlation of $\hat{\rho} = 0.75$

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

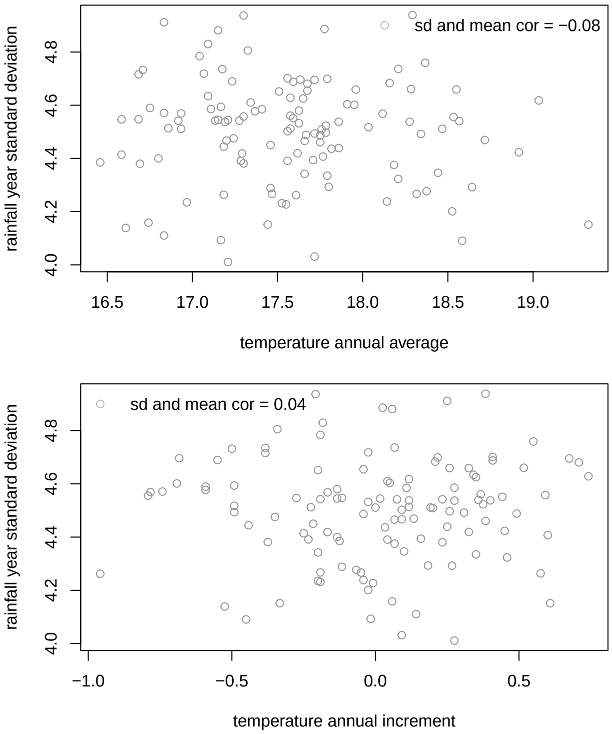


Figure 4: Evolution in time of rainfall average value and rainfall standard deviation with a non parametric with.

In order to give 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 $\mathbf{x}$ governed by a function $f(.)$. The method employed is the locally weighted regression discussed in section 2 for a certain span s.

Therefore, to track the evolution of the phenomenon we fit the model for different values of the span. In particular:

- 100% of the neighbor data to take into account the global context (violet)

- 50% of the neighbor data to allow for a more local structure (blue)

By investigating the results in **??** 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. On the other hand, the local fit shows a constant trend up to the 1980s which evolves in a steeper decrease for the later year. Nevertheless, both the models suggest an evolution of the amount of rainfall which changes over time.

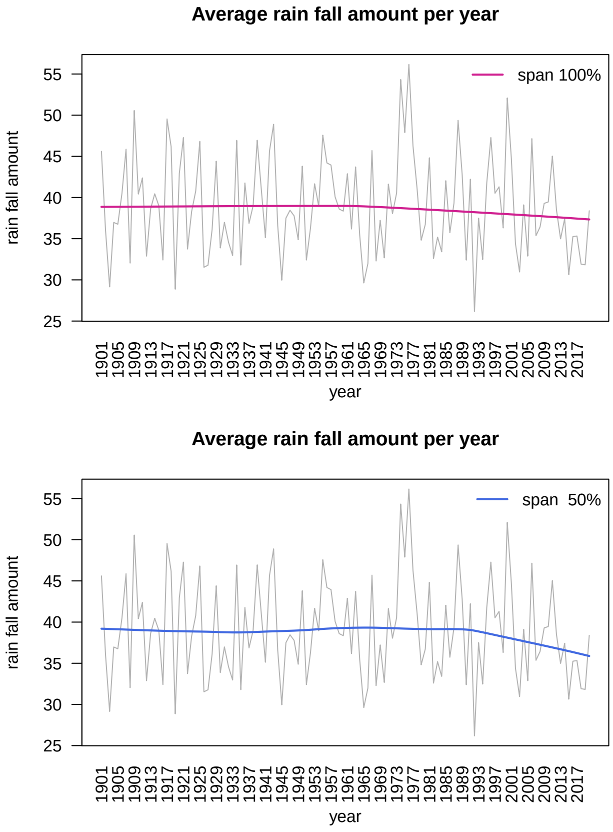


Figure 5: Annual rainfall

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. On the other hand, the local fit shows an increasing trend with different intensities in different period allowing for non linear relationship. Nevertheless, the overall trend does not change dramatically. Annual temperature has clearly been increasing over the last century.

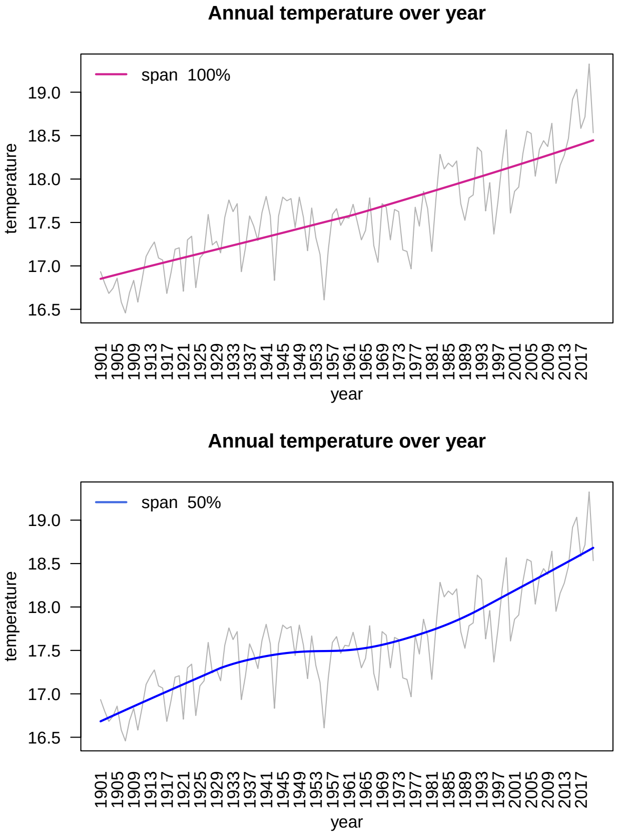


Figure 6: Annual average temperature

In figure **??** we depict what the annual increment is for both span values of $s = \{50, 100\}$. In figure n) with span size $s = 100$ 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 (b) with span size $s = 50$ we 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.

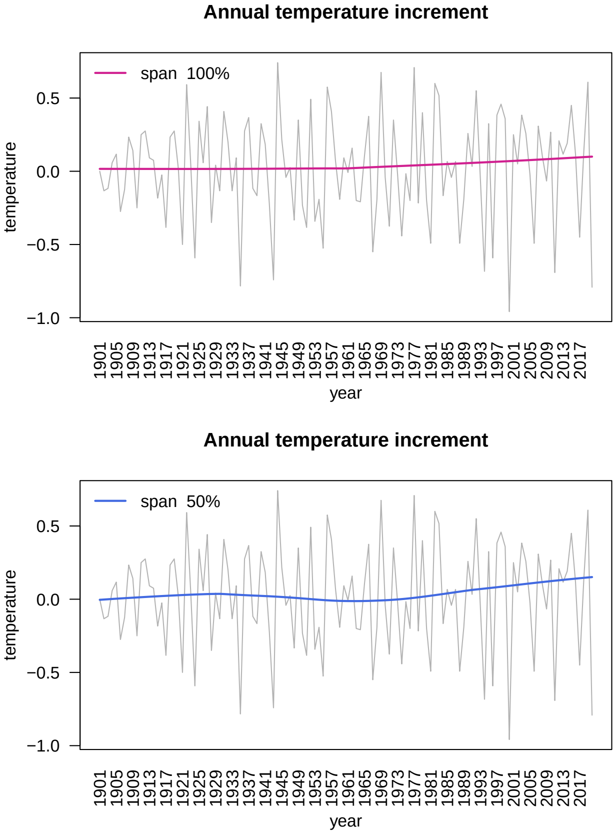


Figure 7: Annual temperature increment

	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: limate correlation table

In order to give an overall idea of the the data we are investigating we would like to present the pairs plot showing each pair combination plus the estimated correlation matrix $\mathbf{R}$ in table **??**. As mentioned earlier, in figure **??** 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 temperature, we can expect higher variation in the average temperature.

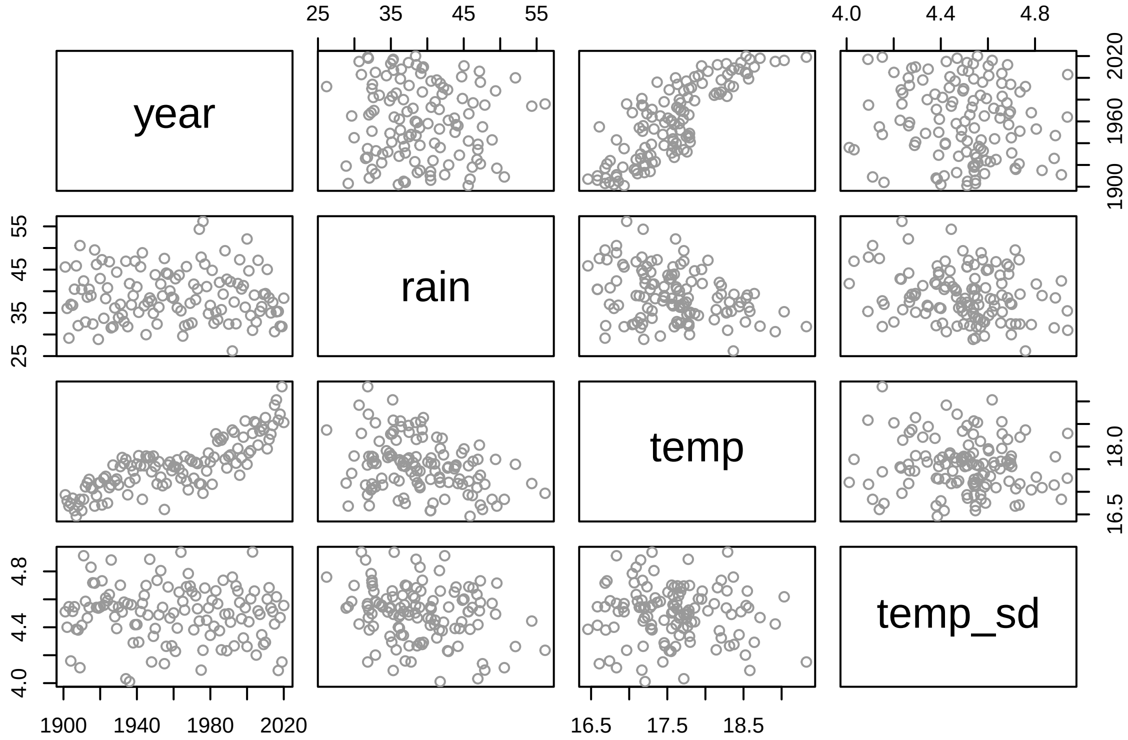


Figure 8: Year, Rain, Temperature scatterplot

4 Climate Modeling

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 different 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 model 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 year. 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 refsec:eda we have seen how the phenomenon is highly variable. Therefore, we choose the mean as initial condition

$$f(\tau, t = 0) = \tau_0 = 16.8$$

Given the aforesaid model, we are interested of 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 $\times\ 10^{-4}$ and an estimated standard error of $\hat{SE}(\hat{\gamma})$ =2.43 $\times\ 10^{-5}$, resulting in the 95% confidence interval

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

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 9. 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.

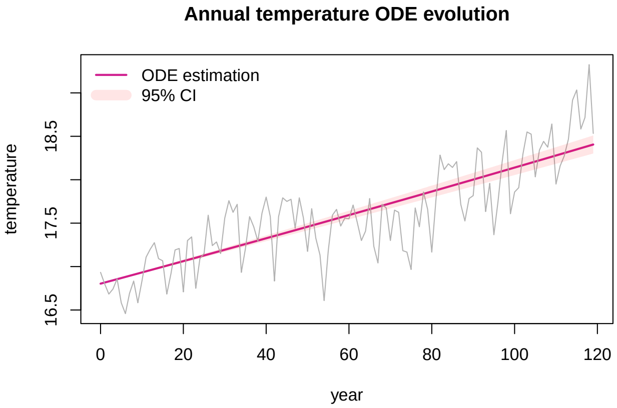


Figure 9: Annual temperature modeled by Ordinary Differential Equation

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 are interested in what would happen in a 50 years time following the current trend. According to the estimated model and depicted in figure 10, the global temperature may continue to rise up to reaching 19 Celcius average degree in 50 years time, with a reasonably tight confidence interval.

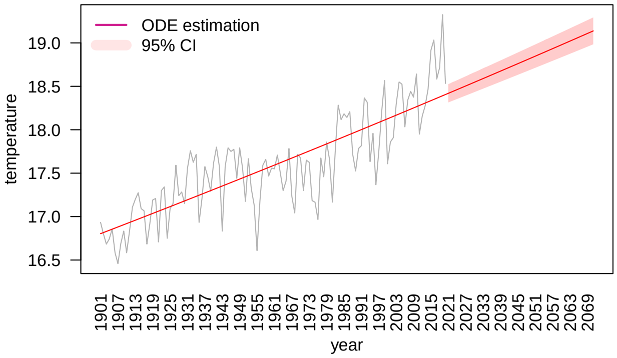


Figure 10: 50 out of sample years temperature prediction

A fundamental requisite for a mathematical model is the 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 allow us to derive the sensitivity equation as

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

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 time the estimated standard error away from the point estimate $\gamma \pm 5 \times \hat{SE}(\hat{\gamma})$. In figure 11 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 value 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 state how the model of interest is robust to fairly small perturbations of the input data. Thus, we can rely its inference.

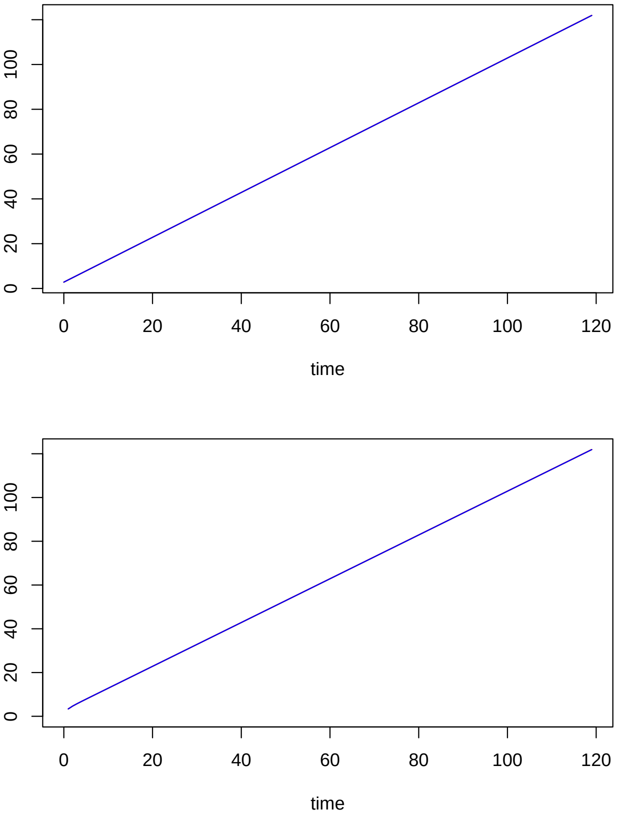


Figure 11: Evolution of temperature in time with input parameter ± 5 $\hat{SE}(\hat{\gamma})$; b) evolution of γ parameter in time ± 5 $\hat{SE}(\hat{\gamma})$

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 2 which allows us to write

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

The first trial is not satisfactory as the model presents some deficiency. In particular, from the left plot in figure **??** we can observe the presence of heteroskedasticity, meaning that the variance is not constant over the range of values for the value in input. 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 a 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 **??**. 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 configuration where we shift the *linear model* parameters estimates by 5 times their standard deviation $\hat{\beta}_i \pm \hat{SE}(\hat{\beta}_i)$. The results are depicted in figure **??**. 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.

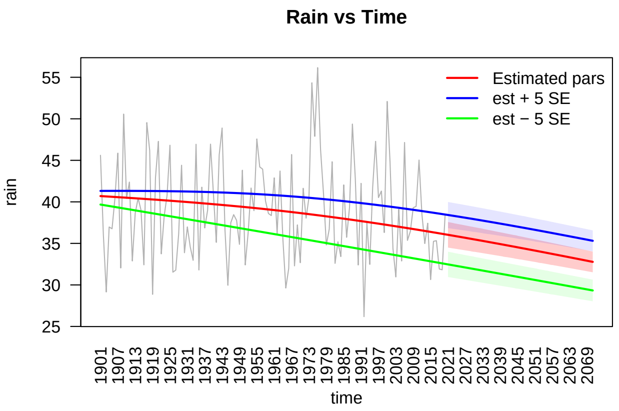


Figure 12: Evolution of rainfall w.r.t. temperature with input parameter ± 5 $\hat{SE}(\hat{\gamma})$

5 Species count Modelling

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 specie 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 might provide robust result, from $n_i = 30$ onward, as per figure **??**.

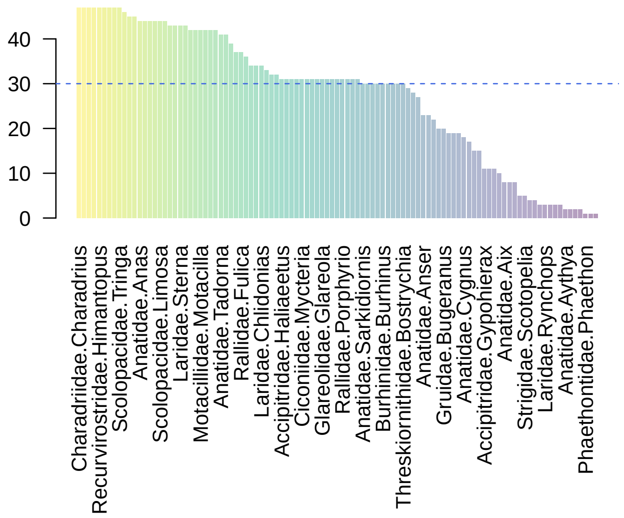


Figure 13: Years observation per species

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 appreaciable large, we provide either mean (red line) and median (blue line) value for robust result. Despite being on different scale, both the results shows similar global trend. The waterbirds individual count has been decreasing over time. This behavior is depicted in figure **??**.

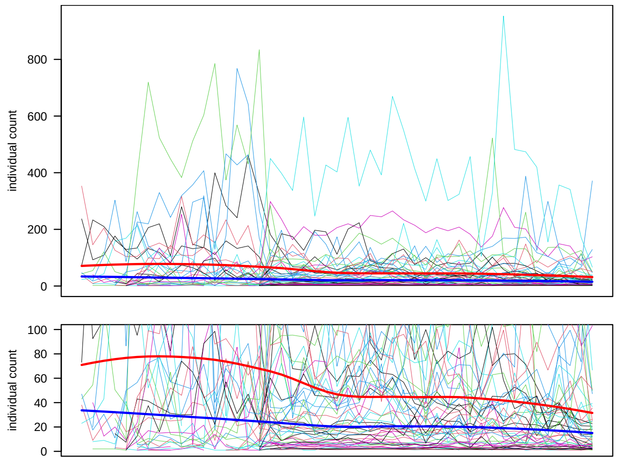


Figure 14: Yearly individual count per species plus mean and median observation

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 ??. As the data takes value on a discrete and positive set, it is reasonable to assume a *Poisson* distribution of the residuals

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

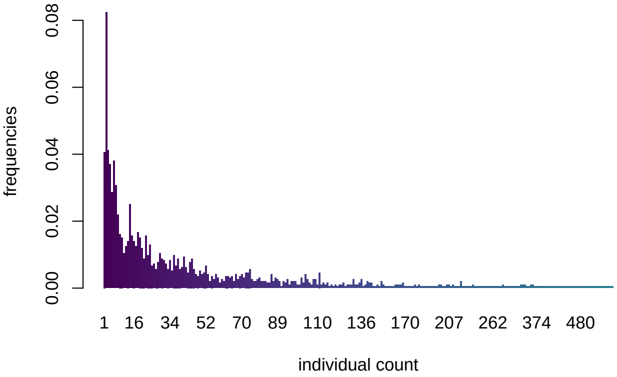


Figure 15: Distribution for animal counts

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 account for 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

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

And positing a distribution on the random terms $\alpha_i \sim N(\mu_{\alpha_i}, \sigma^2_{\alpha_i})$ indicates a random intercept which contribution depends on the $\mathbf{Z} \in \mathbb{N}$ matrix which indicates the species belonging

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

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

The spline degree has been 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 empirically shown to be $K = 5$ with the following variables: *average rainfall* and *average temperature.*

Since we are including different predictors in the model, in order to get stable results we apply a preprocessing step which standardize the variables to have mean zero and variance one

$$Z = \frac{X - \mu}{\sigma}$$

From the model results presented in table 4, we can observe how the coefficients related to the *year* marginal term $\hat{\beta}_{i,year}$ are mainly negative. Moreover, the coefficients related to its interaction terms $\hat{\beta}_{i,year \times \omega}$ are mainly negative as well. On the other hand, we can easily identify positive coefficients for the marginal terms $\hat{\beta}_\tau$ and $\hat{\beta}_r$. However, their magnitude is relatively small compared to the terms effect they interact with $\hat{\beta}_{i,year,temp}$ and $\hat{\beta}_{i,year,rain}$, therefore, we expect an overall negative effect.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	3.14	0.20	15.84	0.00
$\hat{\beta}_{1,y}$	-0.35	0.03	-11.01	0.00
$\hat{\beta}_{2,y}$	0.16	0.03	4.85	0.00
$\hat{\beta}_{3,y}$	-0.04	0.08	-0.54	0.59
$\hat{\beta}_{4,y}$	-1.00	0.04	-24.94	0.00
$\hat{\beta}_{\tau}$	0.12	0.03	3.80	0.00
$\hat{\beta}_{r}$	0.23	0.02	11.50	0.00
$\hat{\beta}_{1,y\times\tau}$	0.20	0.03	6.40	0.00
$\hat{\beta}_{2,y\times\tau}$	-0.52	0.04	-13.65	0.00
$\hat{\beta}_{3,y\times\tau}$	-0.44	0.08	-5.52	0.00
$\hat{\beta}_{4,y\times\tau}$	0.21	0.04	5.41	0.00
$\hat{\beta}_{1,y\times r}$	0.01	0.02	0.36	0.72
$\hat{\beta}_{2,y\times r}$	-0.80	0.03	-22.94	0.00
$\hat{\beta}_{3,y\times r}$	-0.78	0.05	-14.36	0.00
$\hat{\beta}_{4,y\times r}$	-0.01	0.05	-0.11	0.91

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

For the model of interest, we have assumed the error distributed as a *Poisson* distribution, therefore, in the fitted vs residuals inspection, we run into a different behavior. In particular, in figure **??** 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 diagnostic produce satisfactory results, hence the model is appropriate to describe the data.

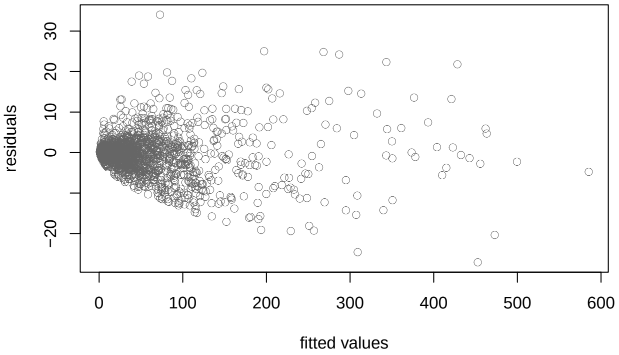


Figure 16: Distribution for animal counts

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

- for the random intercept the assumption does not hold as we can notice a considerable discrepancy between the observed and theoretical quantiles **??**. However, the intercept related variance is $\hat{\sigma}_0 = 1.3586$, which counts for the majority variance as we can see in figure **??**. For instance, in figure **??** 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 which confidence interval does not contain the null values.
- on the other hand, for the random slope, the empirical distribution agrees with the theoretical one in the central area, but is shows deficiencies in the tail regions, figure **??**. Its contribution to the overall variance is much lower than the previous term but still significant, approximately 9%, figure **??**. With regard to the single term values, we can observe a much tighter range of variation, approximately [±1]. Nevertheless, many species slope is estimated to be singificantly different from zero, hence worth including it.

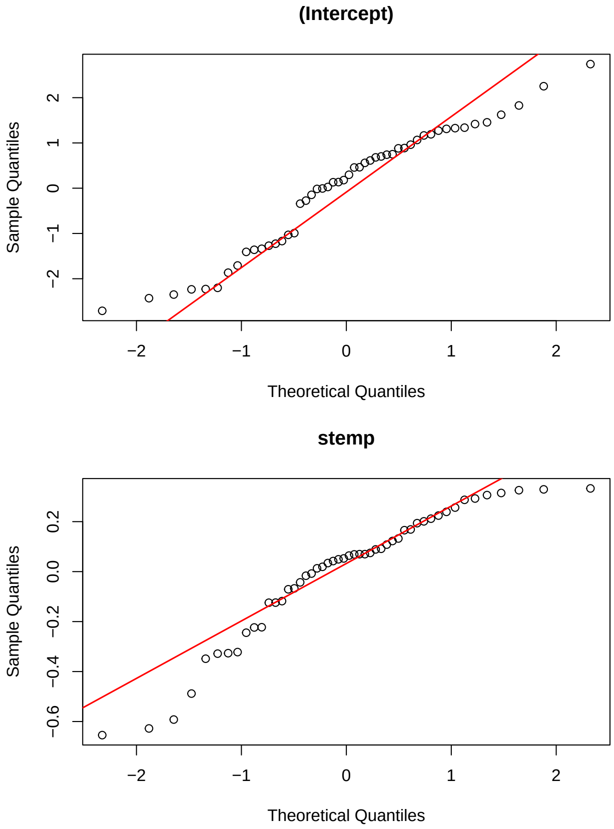


Figure 17: Random effects terms quantile - quantile plot

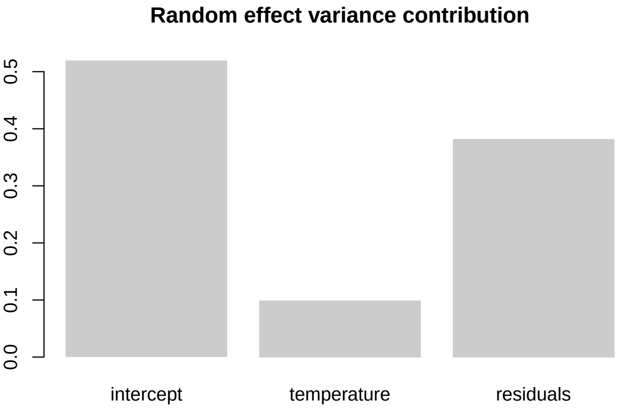


Figure 18: Random effects variance contribution

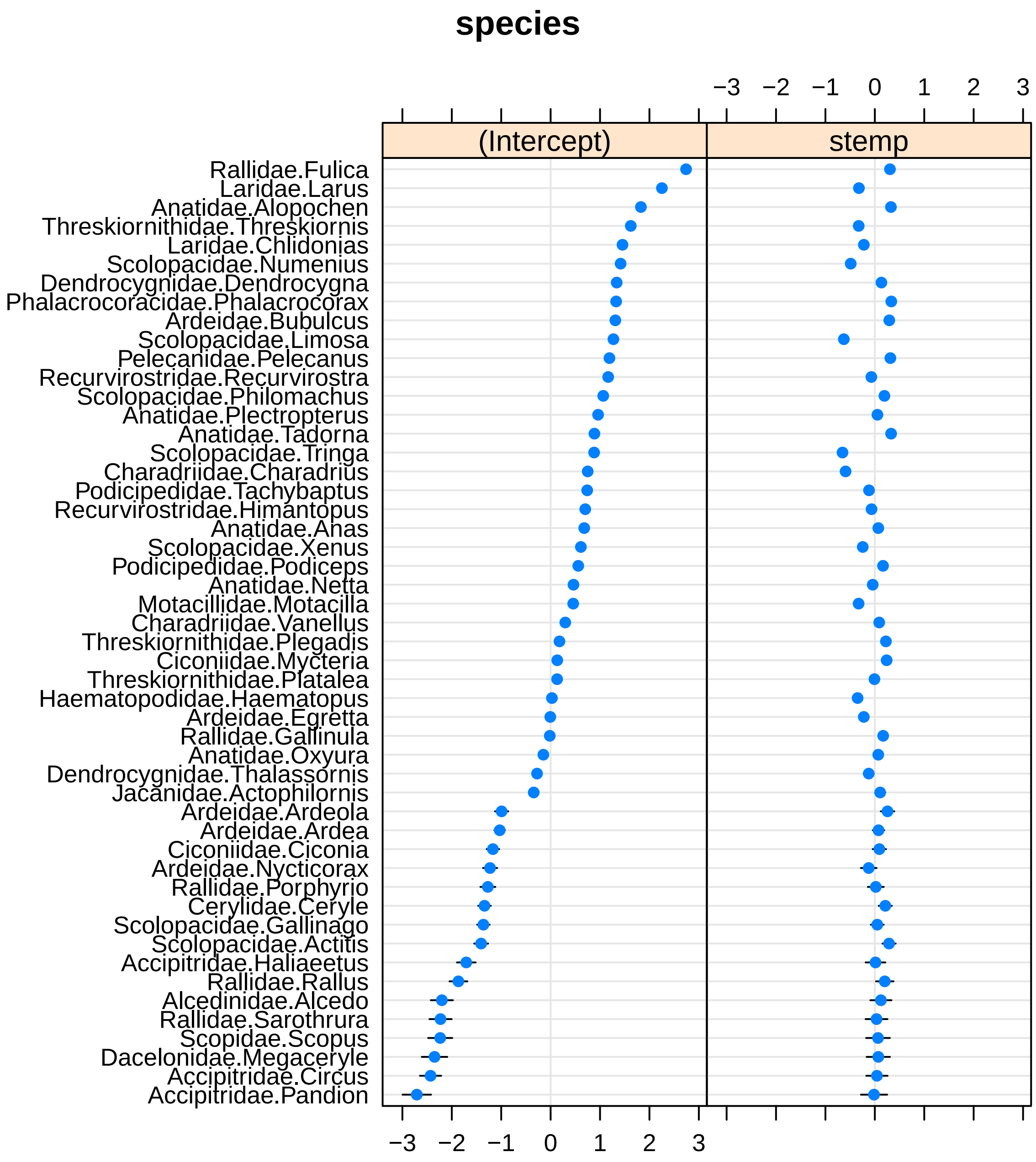
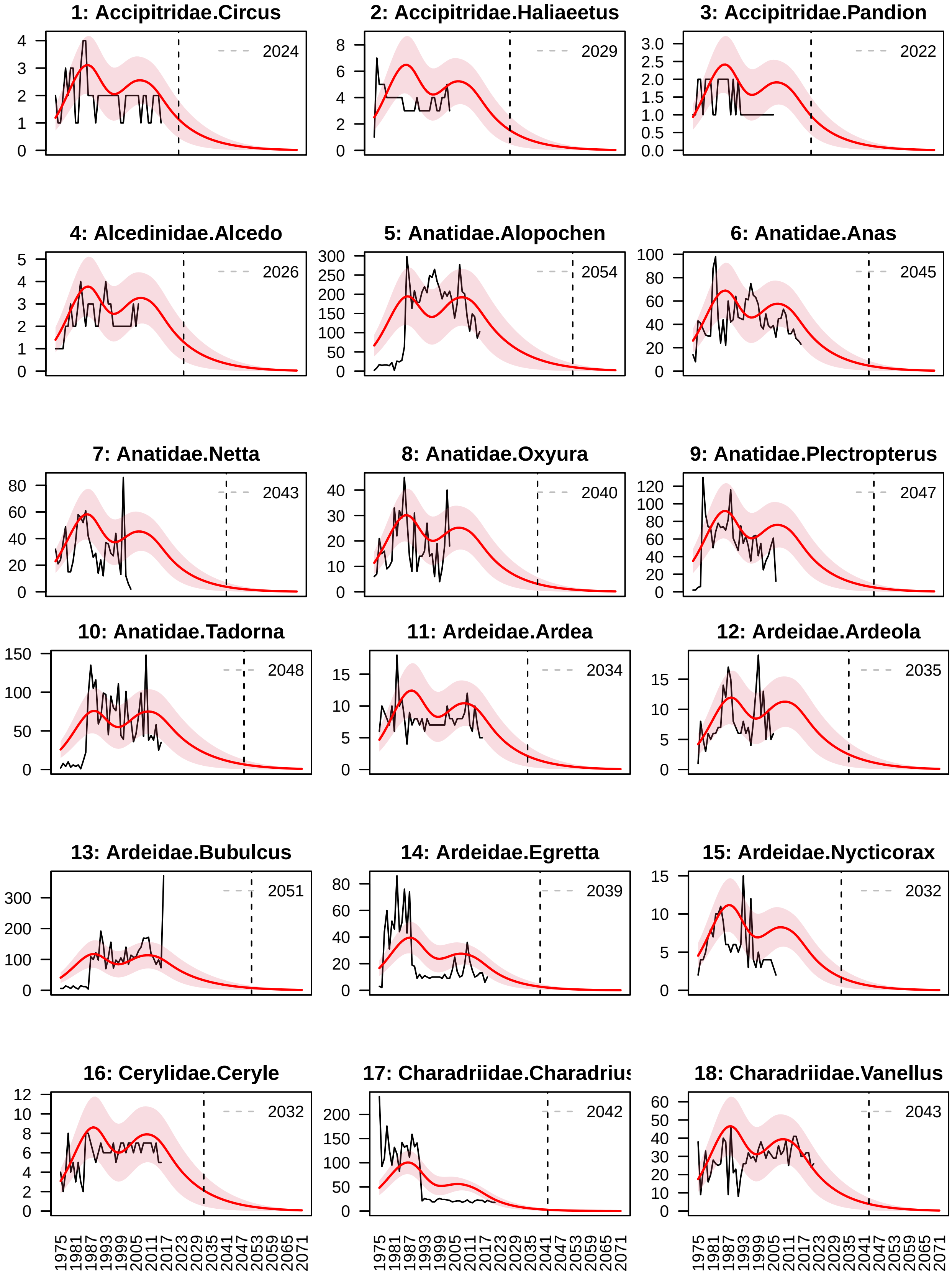


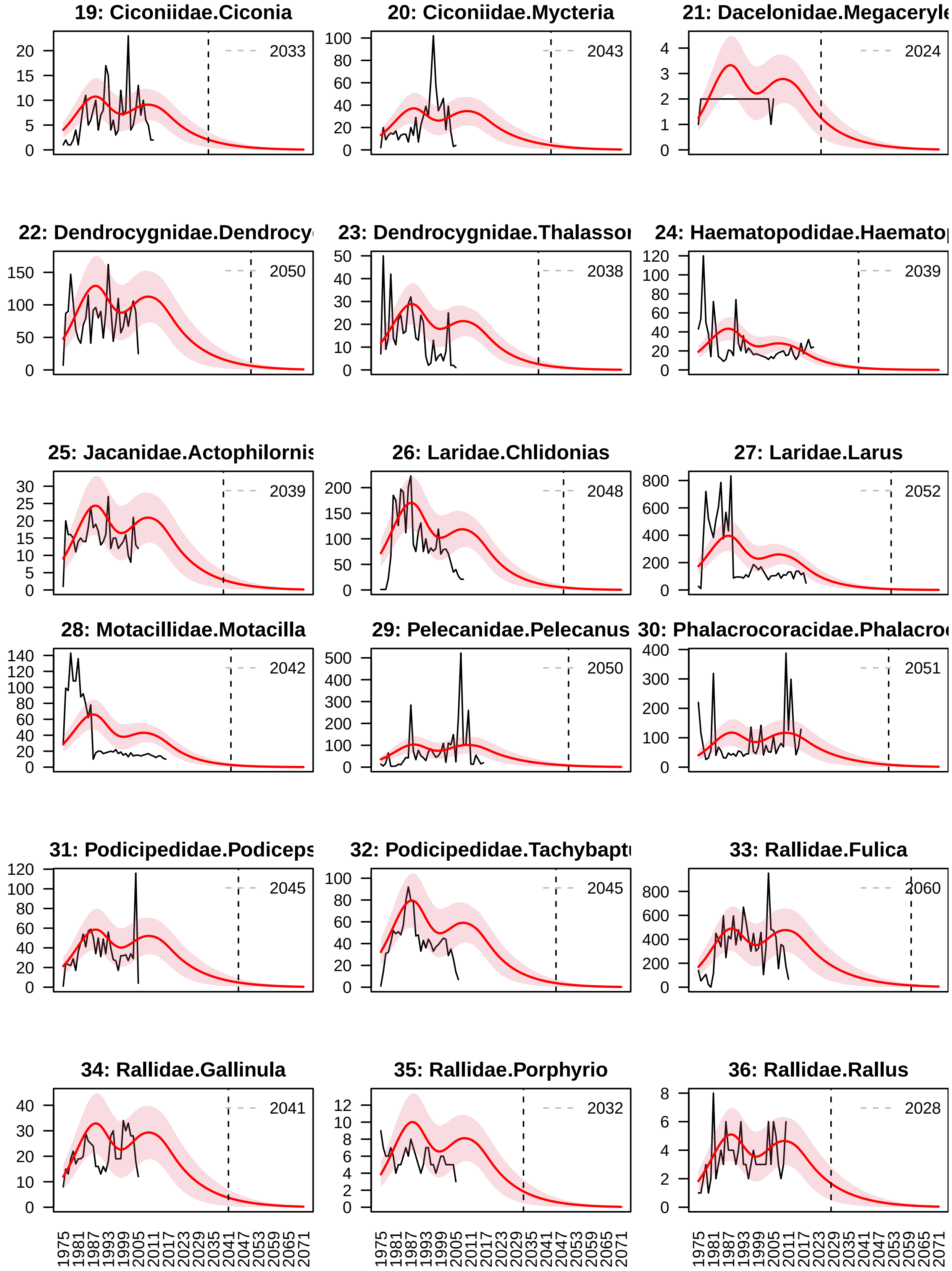
Figure 19: Individual random effect for each specie

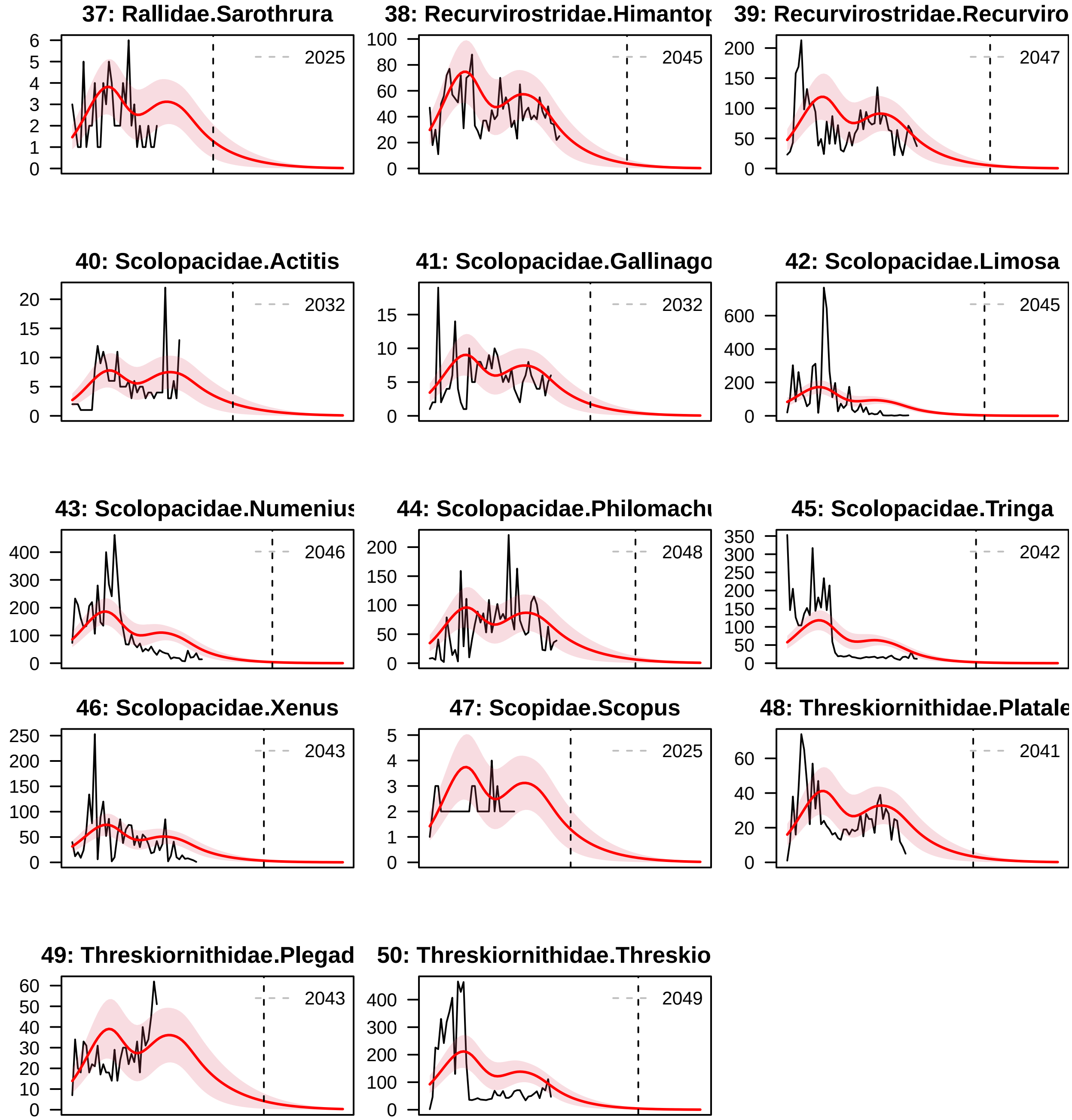
6 Species application

In the previous section, 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 figure **??**. In particular, we can observe that for some of these, the model fits very well the observed data and its trend evolution, see species-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. In particular, for each of these species, we can notice how easy it is, according to the estimated model, to reach one individual left, in the lower bound, which would not allow for the species survival. According to the estimated model, which has shown to capture the overall data structure, this might happen as soon as in ten years time.


1: Accipitridae.Circus
0
1
2
3
4
2024
2: Accipitridae.Haliaeetus
0
2
4
6
8
2029
3: Accipitridae.Pandion
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2022
4: Alcedinidae.Alcedo
0
1
2
3
4
5
2026
5: Anatidae.Alopochen
0
50
100
150
200
250
300
2054
6: Anatidae.Anas
0
20
40
60
80
100
2045
7: Anatidae.Netta
0
20
40
60
80
2043
8: Anatidae.Oxyura
0
10
20
30
40
2040
9: Anatidae.Plectropterus
0
20
40
60
80
100
120
2047
10: Anatidae.Tadorna
0
50
100
150
2048
11: Ardeidae.Ardea
0
5
10
15
2034
12: Ardeidae.Ardeola
0
5
10
15
2035
13: Ardeidae.Bubulcus
0
100
200
300
2051
14: Ardeidae.Egretta
0
20
40
60
80
2039
15: Ardeidae.Nycticorax
0
5
10
15
2032
16: Cerylidae.Ceryle
0
2
4
6
8
10
12
2032
17: Charadriidae.Charadrius
0
50
100
150
200
2042
18: Charadriidae.Vanellus
0
10
20
30
40
50
60
2043
1975
1981
1987
1993
1999
2005
2011
2017
2023
2029
2035
2041
2047
2053
2059
2065
2071


19: Ciconiidae.Ciconia
0
5
10
15
20
2033
20: Ciconiidae.Mycteria
0
20
40
60
80
100
2043
21: Dacelonidae.Megaceryle
0
1
2
3
4
2024
22: Dendrocygnidae.Dendrocy
0
50
100
150
2050
23: Dendrocygnidae.Thalassor
0
10
20
30
40
50
2038
24: Haematopodidae.Haemato
0
20
40
60
80
100
120
2039
25: Jacanidae.Actophilornis
0
5
10
15
20
25
30
2039
26: Laridae.Chlidonias
0
50
100
150
200
2048
27: Laridae.Larus
0
200
400
600
800
2052
28: Motacillidae.Motacilla
0
20
40
60
80
100
120
140
2042
29: Pelecanidae.Pelecanus
0
100
200
300
400
500
2050
30: Phalacrocoracidae.Phalacro
0
100
200
300
400
2051
31: Podicipedidae.Podiceps
0
20
40
60
80
100
120
2045
32: Podicipedidae.Tachybapt
0
20
40
60
80
100
2045
33: Rallidae.Fulica
0
200
400
600
800
2060
34: Rallidae.Gallinula
0
10
20
30
40
2041
35: Rallidae.Porphyrio
0
2
4
6
8
10
12
2032
36: Rallidae.Rallus
0
2
4
6
8
2028
1975
1981
1987
1993
1999
2005
2011
2017
2023
2029
2035
2041
2047
2053
2059
2065
2071


37: Rallidae.Sarothrura
0 1 2 3 4 5 6
2025
38: Recurvirostridae.Himantop
0 20 40 60 80 100
2045
39: Recurvirostridae.Recurviro
0 50 100 150 200
2047
40: Scolopacidae.Actitis
0 5 10 15 20
2032
41: Scolopacidae.Gallinago
0 5 10 15
2032
42: Scolopacidae.Limosa
0 200 400 600
2045
43: Scolopacidae.Numenius
0 100 200 300 400
2046
44: Scolopacidae.Philomachu
0 50 100 150 200
2048
45: Scolopacidae.Tringa
0 50 100 150 200 250 300 350
2042
46: Scolopacidae.Xenus
0 50 100 150 200 250
2043
47: Scopidae.Scopus
0 1 2 3 4 5
2025
48: Threskiornithidae.Platale
0 20 40 60
2041
49: Threskiornithidae.Plegad
0 10 20 30 40 50 60
2043
50: Threskiornithidae.Threskio
0 100 200 300 400
2049

7 Conclusion

8 Appendices

References

[1] T. R. C. M L Parry and M. Hulme, "What is a dangerous climate change? ," *Global Environmental Change*, vol. 6, no. I, 1996.

[2] M. R. . Y. Richard, "Intensity and spatial extension of drought in South Africa at different time scales," *African Journals Online*, vol. 29, no. 4, 2003.

[3] M. Grayson, "Agriculture and drought," *Nature Outlook*, vol. 501, no. 7468, 2013.

[4] J. C. H. et. al, "Mortality of wildlife in Nairobi National Park, during the drought of 1973-1974," *African Journal of Ecology*, vol. 15, no. 1-18, 1977.

[5] G. G. B. I. Facility, "Free and open access to biodiversity data." www.gbif.org/occurrence/.

[6] W. B. Group, "Climate change knowledge portal." https://climateknowledgeportal.worldbank.org/download-data.

[7] P. McCullagh and J. Nelder, *Generalized Linear Models*. International series of monographs on physics, Chapman and Hall, 1989.

[8] R. T. Trevor Hastie, *Generalized Additive Models*. 1990.

[9] M. W. M. Aertsa, G. Claeskensb, "Some theory for penalized spline generalized additive models," *Journal of Statistical Planning and Inference*, vol. 103, no. 455-470, 2002.

[10] P. M. Hooper, "Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models," *Journal of the American Statistical Association*, vol. 103, no. 88:421, 2012.

[11] G. Gbur, *Mathematical methods for optical physics and engineering*. 2011.

[12] M. J.J., *The Levenberg-Marquardt algorithm: Implementation and theory*. Lecture Notes in Mathematics, Springer, 1978.

[13] J. Morio, "Global and local sensitivity analysis methods for a physical system," *European Journal of Physics*, vol. 32, no. 6, 2011.

[14] W. Alexander, "Floods, droughts and climate change," *South African Journal of Science*, vol. 91, 1995.