Climate Change, Biodiversity Loss*

A replication analysis

Julia Kim

April 2, 2024

Abstract

Abstract here.

Table of contents

---

1 Introduction

Throughout, our investigation is conducted using the open-source statistical programming language `R` (R Core Team 2023), with functionalities from `tidyverse` (Wickham et al. 2019), `knitr` (Xie 2023), `here` (Müller 2020) and `kableExtra` (Zhu 2021).

The 1973 Endangered Species Act (ESA) under the U.S. Fish & Wildlife Service (USFWS) empowered the federal government of the United States to protect species from extinction (USFWS 2020). Simply by listing a species as endangered, the government would promptly respond by implementing protections and regulations aimed at conserving the species and its habitat (USFWS 2020). Cost-benefit analysis would be practically omitted, and legal avenues would be opened for other developmental projects to be postponed or cancelled, and millions of dollars in expenditures to be incurred (Weitzman 1992). Moreover, the listed species would be eligible to receive funds for their recovery, with the aim of reducing their endangerment to levels that would qualify their removal from the list (Weitzman 1992). The costs of this type of environmental protection are thus considerable, growing faster than any other expenditure of comparable size in the American economy (Weitzman 1992). To illustrate, the fiscal budget of 2023 for the conservation of endangered wildlife exceeded $111 million USD for just over 300 projects (USFWS 2020). Understanding the factors that influence the listing of a species on the ESA list therefore deserves serious attention.

Evolution in genetics mean it is now possible to measure the evolutionary distinctiveness of species, by determining its genetic distance to its nearest relatives (Moore et al. 2022).

The paper proceeds as follows.

2 Data

2.1 Source

The paper and data used for replication are obtained from "Noah's Arc in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation Costs in the United States" (Moore et al. 2022), published in the *Journal of the Association of Environmental and Resource Economists*. Their analysis sought, in part, to explore how the probability of a species being listed under the ESA was influenced by its conservation status, utility value to humans, and species uniqueness. Our replication seeks to address the validity of these three findings and to discuss the projected increases in species listing due to climate change.

2.2 Methodology

The original dataset is made publicly available by Moore et al. (2022). It composed of $n = 64,583$ observations, with each row representing a unique species and each column indicating a specific variable. As part of our reproduction, we removed variables not of interest to our analysis and clarified the names of variables to make them easier to work with. A further discussion as to the variables of actual interest, their source of data, measurement, and the methodology employed to clean them, follow.

2.2.1 Endangerment of Species

The taxon and conservation status of species were obtained from NatureServe, an authoritative source of biodiversity data throughout the United States (NatureServe 2024). The database is dynamic, being continuously updated by botanists, zoologists and ecologists, with inputs from scientists at heritage and conservation programs (NatureServe 2024). The assessment of species is revised on a periodic basis, for we use the most recent assessment conducted between 1985 and 2019.

NatureServe regroups species into nine taxon or classes of living things, of which there are five vertebrates species – amphibians, birds, fishes, mammals, reptiles – and four non-vertebrate species – plants, invertebrates, fungi and protists. Such a classification is determined by experts, who collect information from field surveys, taxonomic treatments and other scientific publications (NatureServe 2024). Additionally, NatureServe uses a protocol tp assign a conservation status to a species on the basis of ten factors, regrouped into three categories (Faber-Langendoen et al. 2009). The most important is rarity, composing population size and extent. The other two are the anthropogenic threats to the species and the long- and short-term trends in population size or range extent (Faber-Langendoen et al. 2009). Data collected from primary literature and the field are used to assign scores for individual factors, which are then translated into one of five final conservation status ranks (Faber-Langendoen et al. 2009). Using definitions provided by NatureServe (2024), these five rankings are explained in Table 1.

Table 1: Definition of Conservation Status Ranks used by NatureServe.

Rank	Category	Definition
G1	**Critically Imperiled**	At very high risk of extinction, due to very restricted range, extreme rarity, very severe threats or other factors.
G2	**Imperiled**	At high risk of extinction due to restricted range, few occurrences, severe threats or other factors.
G3	**Vulnerable**	Rare or local, due to restricted range, few occurences, recent or widespread threats or other factors.
G4	**Apparently Secure**	Uncommon, but not rare, with some cause for long-term concern.
G5	**Secure**	Common, widespread and abundant.

A breakdown of the conservation status by taxon is shown in Figure 1, which is a replication of Figure 1B in the paper by Moore et al. (2022). To improve the visualisation of the result, we have coloured each status ranking, according to the conventional colours employed by NatureServe (2024). Observe 60% of species have an unknown status, being almost entirely of the invertebrate (41%), plants (40%) or fungi (16%) categories. This distribution of missing data is informative, as it suggests a lack of scientific attention to these three non-vertebrate species that appear not to be highly prioritised in the ranking process (Moore et al. 2022). In terms of the data that are known, reptiles appear to have the greatest proportion of known secure species, whilst mammals have the greatest proportion of known critically imperiled species.
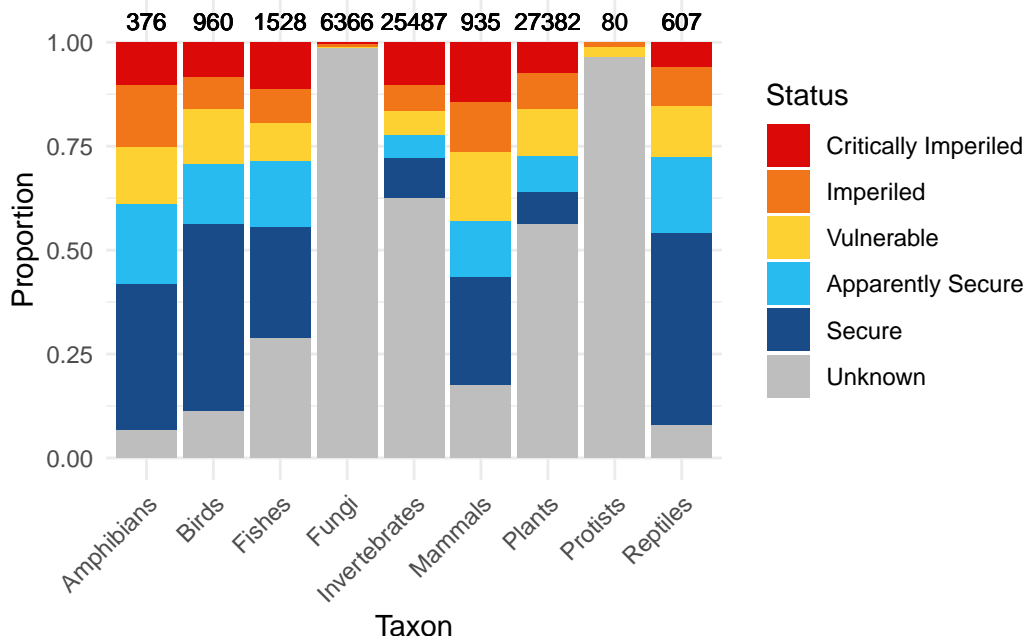
Figure 1: Distribution of assessed conservation status using NatureServe ranking system. Number above the bars gives the total number of species per taxon.

2.2.2 Google Ngrams for Species

An n-gram is defined as a sequence of n words in some particular order: a 2-gram, for instance, is a two-word sequence like "*Balaena mysticetus*", whilst a 3-gram is three-word sequence like "*Lithobates areolatus circulosus.*" An on-line search engine, the Google Books Ngram Viewer uses Optical Character Recognition (OCR) to provide the frequency of such n-grams in a corpus of over 25 million digitised books published over the course of more than 500 years (Google 2020). These books are published in eight languages; in particular, the English corpus of 2019 consists over over 16 million of such books published between 1470 and 2019 (Michel et al. 2011).

As argued by Moore et al. (2022), the rich variety of ways in which a species may provide utility – through cultural significance, commercial value, scientific interest, and so forth – should likely directly influence the frequency with which the species is written about over time. This is consistent with intuition and broader literature in applied computing, with authors positing that elements most influential on a society and culture should naturally be reflected in the written word of those who are a part of that society and culture (Knight and Tabrizi 2016; Aisopos et al. 2016). As a result, following the workflow of Moore et al. (2022), this analysis adopts common and scientific name n-grams as an additional estimate of the utility value of species to humans. Potential limitations of this measure are provided in Section 5.5.

To determine the n-gram frequencies for species names, Moore et al. (2022) performed case-insensitive searches in Google Books' English corpus of 2019 for all common and scientific names present in the NatureServe and ESA archives between 1800 and 2016. They assigned a frequency of 0 across all years whenever a name failed to return valid data. Here, the purpose of treating common and scientific name n-grams separately owes to the fact that common names pose a few challenges. In particular, a species may be designated colloquially by multiple common names or may lack a common name altogether (Cheese 2021). Conversely, a single common name may be used for multiple different species (e.g., a "millipede" referring to any of the 10,000 species in the anthropod class Diplopoda) or may have additional, unrelated meanings or uses (e.g., "British soldier" referring to an army serviceman

of the United Kingdom or to the lichen *Cladonia cristatella*) (Cheese 2021). To account for such complicating factors, common and scientific names were thus treated separately.

As described by Moore et al. (2022), n-gram frequencies were then aggregated to the species level. More specifically, for unlisted species, common and scientific n-gram frequencies were taken to be their respective averages from 1950 to 2016. For listed species, they were calculated as their respective average frequencies from 1950 to 10 years *prior* to the date of the listing decision. This was done to minimise the bias to the n-gram values following any publicity generated by the listing decision (Moore et al. 2022). Owing to how small the raw n-gram frequencies were, all frequencies were then z-score normalised: this was accomplished by calculating the difference between each raw frequency and the corresponding n-gram mean and dividing the whole by the corresponding n-gram standard deviation. At last, to further account for the challenges posed by common names, as discussed previously, any species whose standardised common name n-gram had a frequency > 10 times its corresponding standardised scientific n-gram frequency were discarded in the dataset (Moore et al. 2022).

A plot of the mean standardised common and scientific n-gram frequencies per taxon is shown in Figure 2. This figure shows that scientific name n-grams are typically correlated with their corresponding common name n-grams. For birds, fishes and mammals, both types of n-grams are well above the standardised mean, whilst for the non-vertebrates – comprising fungi, invertebrates and plants –, both types fall below the mean. Altogether, it appears that the utility value of a species is directly reflected in the frequency with which the species is written about *both* in colloquial and scientific settings, with vertebrate species exhibiting greater utility value than non-vertebrate species. That scientific name n-grams are much greater in magnitude than their corresponding common name n-gram may be especially driven by the preferred usage of scientific names in technical literature.
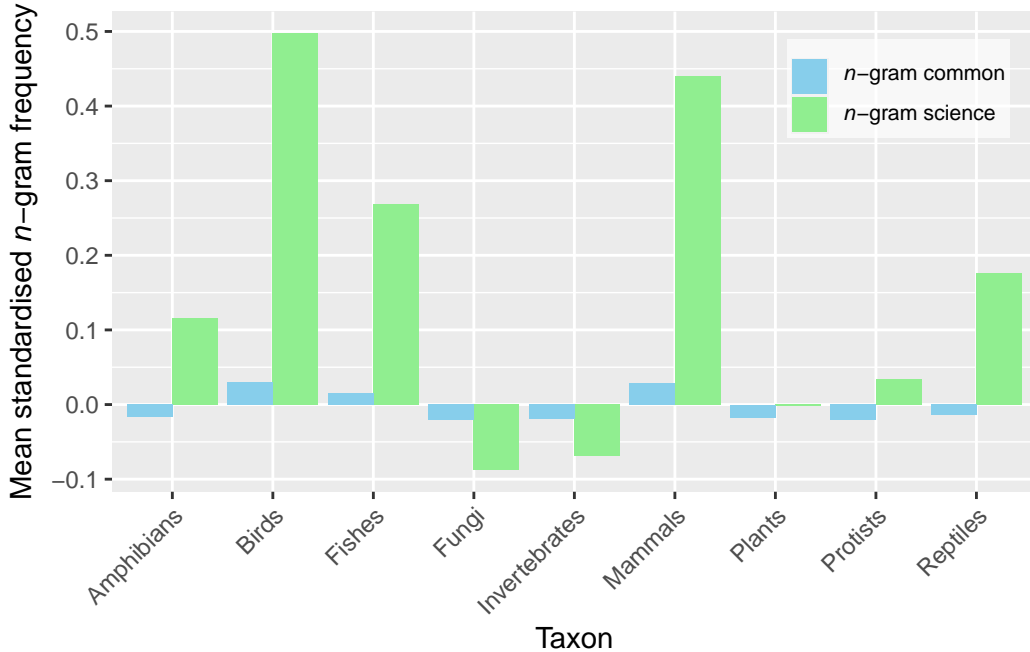

Figure 2: Distribution of the mean standardised common and scientific n-gram frequencies per taxon.

2.2.3   Species uniqueness

Species uniqueness measures how evolutionarily isolated a species is on its family (or phylogenetic) tree (Redding et al. 2008). A highly unique species is expected to have split off long ago from its

nearest living relative and, hence, to have evolved measurably different genetic features and functions compared to any other member in its phylogenetic tree (Redding et al. 2008). On a phylogenetic tree will generally appear members of the species' clade, which is to say the group of species that are descendants of a common ancestor (Moore et al. 2022). To illustrate, the Siberian tiger and North American brown bear are not regarded as evolutionarily unique: each has a number of closely-related species that are highly genetically similar and in little danger of going extinct (Metrick and Weitzman 1996). At the opposite extreme are the platypus and narwhal that are so evolutionarily unique as to each form a monotypic genus: they are the only genetically distinct representatives of their entire genus and are very distantly related to their nearest relatives in other genera (Metrick and Weitzman 1996). Note that a genus (or, in plural, genera) simply refers to a taxonomic classification of organisms between the levels of species and family.

Two estimates of species uniqueness are employed in this analysis. The first is the number of species within a genus, a metric originally used by Metrick and Weitzman (1996). As discussed previously, a large number of species in a single genus should imply a smaller degree of species uniqueness, and vice-versa. A benefit to this measure is that, because genera were historically defined based on obvious similarities in physical traits, it may capture aspects of distinctiveness that are more visually evident to humans (Moore et al. 2022). Another benefit is that data on the number of species in genera are widely available for almost every species ($n = 55,406/64,583$) in our dataset. For our analysis, the data were obtained from two automated databases, the Integrated Taxonomic Information Service (ITIS) and the National Centre for Biotechnology Information (NCBI), using the `taxize` package in R (Chamberlain and Szocs 2013). Note ITIS and NCBI provide central databases that curate taxonomic information from submissions made by taxonomy services, specialists, and researchers working on primary literature in the field (ITIS et al. 2023; Schoch et al. 2020). These experts collect and analyse morphological and molecular data from thousands of species, and compare their values to references in relevant phylogenetic literature, in order to develop robust and consistent taxonomic classifications (Smith et al. 2006). Once taxonomic placement for each species is relatively certain, researchers may then approximate the number of species in each genera by counting (Smith et al. 2006).

The second measure to describe species uniqueness is evolutionary distinctiveness (ED), which has received the most widespread use of all uniqueness measures (Gumbs et al. 2018). ED describes the species' relative contribution to the total evolutionary history of its clade (Moore et al. 2022). A more accurate proxy than genus size, ED is computed by first assigning a value to each branch in the species' phylogenetic tree, equal to the branch's length (measured in millions of years) divided by the number of species subtending it (Isaac et al. 2012). The ED of a species is then the sum of these values for all branches from which the species is descended (Isaac et al. 2012). For our analysis, ED data were obtained by published ED scores for mammals (Gumbs et al. 2018), amphibians (Isaac et al. 2012; Safi et al. 2013; Gumbs et al. 2018), birds (Jetz et al. 2014; Gumbs et al. 2018) and reptiles (Gumbs et al. 2018), all obtained from the Evolutionary Distinctive Globally Endangered (EDGE) of Existence program (London 2008). To obtain ED scores, Isaac et al. (2012), Gumbs et al. (2018), Safi et al. (2013) and Jetz et al. (2014) assembled phylogenetic trees for extant mammalia, amphibians, birds or reptiles that were maximally inclusive (i.e., including all relevant species) and maximally resolved (given the available data). The authors typically built on previously established phylogenies, adding additional clades, dating the nodes with Markov processes, and combining molecular data and taxonomic constraints for data-deficient species to more comprehensively account for uncertainty. Thereafter, they estimated the species' contribution to species uniqueness using the ED algorithm described by Isaac et al. (2012). At the time of the analysis, no data were yet available for fish, fungi, invertebrates or protists, and too few data ($n = 319/19,092$) were yet known for the plants in our study to be considered for analysis.

We compare these two measures in Section 6, plotting ED scores against genera size on species, for which ED scores were available. Although the relationship tends in the expected direction, whereby

species in larger genera tend to be less evolutionarily distinct, only a very small proportion of the variance ($R^2 = 3.6$) in the ED variable can be explained by the genus size variable. Consistent with Moore et al. (2022), this shows that genus size is a only very weak measure of the genetic distinctiveness of a species.

2.2.4 Listing of Species

The dependent variable of interest in this analysis is the `listed` variable, set to 1 if the species was listed for protection as of August 2020 under the ESA and to 0 otherwise. An excellent overview of the process of listing a species is described by Metrick and Weitzman (1996): it begins when the species is proposed by the USFWS as a "candidate" for protection. During its period of candidacy, USFWS collects data from internal and external sources to assess whether the species warrants listing. If sufficient scientific evidence exists, then the USFWS proceeds to submit a formal proposal under the Federal Register and makes requests for comments from the public. After this period of public surveying, USFWS comes a final decision, which officially determines whether the species is listed under the ESA.

Brief summary statistics of the total number and proportion (%) of listed species per taxon is presented in Table 2. The four non-vertebrates categories contain either no listed species (fungi, protists) or a relatively small proportion of listed species (invertebrates and plants). Instead, listing decisions are significantly biased towards the vertebrate categories, for which the average proportion is 7.3%– about three times as high as the highest proportion among the non-vertebrate categories.

Table 2: Summary statistics table of the number of species listed under the ESA by taxon.

Taxon	Listed	Prop. (%)
Amphibians	17	4.5
Birds	93	9.7
Fishes	128	8.4
Fungi	0	0.0
Invertebrates	212	0.8
Mammals	78	8.3
Plants	687	2.5
Protists	0	0.0
Reptiles	35	5.8

3 Model

We perform a logistic regression to determine the relative importance of these species characteristics in the listing decision.

3.1 Model set-up

3.1.1 Model justification

4 Results

5 Discussion

5.1 Findings

Since listing a species is the crucial first step in its protection, Moore et al. (2022) seek to gain insights into the determinants of the government's decision to list the species under the ESA. Our paper has successfully replicated three of their major findings:

(1) The probability of listing changes with conservation status, decreasing monotonically from the most to the least endangered.

(2) There is some evidence that the probability of listing of a species is associated with its utility value to humans. In particular, imperiled vertebrates – including reptiles, fish, birds, amphibians, and mammals – are much more likely to be listed than plants and invertebrates. However, listing probability appears to be significantly affected by neither scientific nor common name *n*-grams.

(3) There is conflicting evidence as to whether the probability of listing is related to the distinctiveness of a species. Whilst species in smaller genera are more likely to be listed, there is no evidence that the species' phylogenetic diversity plays a role in the listing decision.

5.2 Case study

5.3 Ethical Implications

5.4 Accounting for Bias

5.5 Limitations

Although the ESA mandates the listing all species at risk of extinction, capacity and budgetary constraints often slow down the the listing process, leading to a backlog of species awaiting assessment (Alexander 2010). In some cases, USFWS will designate a species as "warranted but precluded" to acknowledge that listing the species is necessary but of lesser priority than other species in need of greater protection (Alexander 2010).

A limitation of the *n*-gram measure is that it captures only English-language media and includes only material published in books, excluding popular media such as amagazines, websites or newspapers.

Limitations to using ngram as a proxy for species value

5.6 Future Research

In addition, USFWS assigns a "conflict code", indicating whether or not the USFWS has determined that protection of species conflicts with economic development.

6 Appendix

7 Additional data details

References

Aisopos, Fotis, Dimitrios Tzannetos, John Violos, and Theodora Varvarigou. 2016. "Using n-Gram Graphs for Sentiment Analysis: An Extended Study on Twitter." In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 44–51. https://doi.org/10.1109/BigDataService.2016.13.

Alexander, Kristina. 2010. "Warranted but Precluded: What That Means Under the Endangered Species Act (ESA)." Washington, DC: Congressional Research Service.

Chamberlain, Scott, and Eduard Szocs. 2013. "Taxize - Taxonomic Search and Retrieval in r." *F1000Research* 2: 191. https://f1000research.com/articles/2-191/v2.

Cheese, Tyler. 2021. "What's in a Name? Problematic Names in the World of Wildlife." *Canadian Geographic.* Canadian Geographic. https://canadiangeographic.ca/articles/whats-in-a-name-problematic-names-in-the-world-of-wildlife/.

Faber-Langendoen, Don, Larry Master, Jennifer Nichols, Kristin Snow, Adele Tomaino, Roxanne Bittman, Geoffrey Hammerson, Bonnie Heidel, Leah Ramsay, and Bruce Young. 2009. "NatureServe Conservation Status Assessments: Methodology for Assigning Ranks." Arlington, VA: NatureServe.

Google. 2020. "Google Books Ngram Viewer." https://storage.googleapis.com/books/ngrams/books/datasetsv3.html.

Gumbs, Rikki, Claudia L. Gray, Oliver R. Wearn, and Nisha R. Owen. 2018. *PLoS ONE* 13 (4): 1–19.

Isaac, Nick J. B., David W. Redding, Helen M. Meredith, and Kamran Safi. 2012. "Phylogenetically-Informed Priorities for Amphibian Conservation." *PLoS ONE* 7 (8): 1–8.

ITIS, S. Alexander, A. Hodson, D. Mitchell, D. Nicolson, T. Orrell, and D. Perez-Gelabert. 2023. "The Integrated Taxonomic Information System." In *Catalogue of Life Checklist (Version 2023-10-31)*, edited by O. Bánki, Y. Roskov, M. Döring, G. Ower, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, et al. ITIS. https://doi.org/10.48580/dfgnm-4ky.

Jetz, Walter, Gavin H. Thomas, Jeffrey B. Joy, David W. Redding, Klaas Hartmann, and Arne O. Mooers. 2014. "Global Distribution and Conservation of Evolutionary Distinctness in Birds." *Current Biology* 24 (9): 919–30.

Knight, Gregory P., and Nasseh Tabrizi. 2016. "Using n-Grams to Identify Time Periods of Cultural Influence." *J. Comput. Cult. Herit.* 9 (3): 15:1–19. https://doi.org/10.1145/2940332.

London, Zoological Society of. 2008. "Edge of Existence Programme." www.edgeofexistence.org.

Metrick, Andrew, and Martin Weitzman. 1996. "Patterns of Behavior in Endangered Species Preservation." *Land Economics* 72 (1): 1–16. https://EconPapers.repec.org/RePEc:uwp:landec:v:72:y:1996:i:1:p:1-16.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph Porter Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.

Moore, Frances C, Arianna Stokes, Marc N Conte, and Xiaoli Dong. 2022. "Noah's Ark in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation Costs in the United States." *Journal of the Association of Environmental and Resource Economists* 9 (5): 981–1015.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://here.r-lib.org/.

NatureServe. 2024. "Unlocking the Power of Science to Guide Biodiversity Conservation." https://www.natureserve.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Redding, David W, Klaas Hartmann, Akihiro Mimoto, Dragan Bokal, Maud Devos, and Arne Ø Mooers. 2008. "Evolutionarily Distinctive Species Often Capture More Phylogenetic Diversity Than Expected." *Journal of Theoretical Biology* 251 (4): 606–15. https://doi.org/10.1016/j.jtbi.2007.12.006.

Safi, Kamran, Katrina Armour-Marshall, Jonathan E. M. Baillie, and Nick J. B. Isaac. 2013. "Global Patterns of Evolutionary Distinct and Globally Endangered Amphibians and Mammals." *PLoS ONE* 8 (5): 4–12.

Schoch, Conrad L, Stacy Ciufo, Michael Domrachev, Carol L Hotton, Sujatha Kannan, Rita Khovanskaya, Detlef Leipe, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database: The Journal of Biological Databases and Curation* 2020. https://doi.org/10.1093/database/baaa062.

Smith, Alan R., Kathleen M. Pryer, Eric Schuettpelz, Petra Korall, Harald Schneider, and Paul G.

Wolf. 2006. "A Classification for Extant Ferns." *Taxon* 55 (3): 705–31. https://doi.org/10.2307/25065646.

USFWS. 2020. *Federal and State Endangered Species Expenditures Fiscal Year 2020*. Washington, DC: U.S. Department of the Interior, Fish and Wildlife Service.

Weitzman, Martin L. 1992. "On Diversity*." *Quarterly Journal of Economics* 107 (2): 363–405.

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

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

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