# Mapping Wildlife in the Rockies

Basil Jancso-Szabo Dhvani Doshi Mathieu-Joseph Magri

## I. Problem

Climate change is significantly altering climatic factors such as precipitation patterns, temperatures, and seasonal cycles across the globe [1]. Specifically in Canada, the climatic patterns have evolved with different areas experiencing different effects from climate change. In general, northern Canada has experienced a temperature increase in 2.3°C from 1948 to 2016, which is nearly three times the global average. It has also seen an increase in annual precipitation from both rain and snow at about a 15% increase [2]. Meanwhile, in southern Canada, the changes have been more variable where some regions are facing drier climates while others are experiencing more extreme rainfall events. These changes are further occurring on a seasonal basis where summers have become much drier in certain locations while the winters experience much more precipitation.

These changes in climatic factors inevitably lead to significant transformations in the vegetation and physical ecosystems of these regions. In the north of Canada, increased precipitation and warming has led to an extended growing season, which allows vegetation such as shrubs and trees to take over the tundra environment. This means that the boreal forest is on average expanding northward and replacing the tundra ecosystem which means a loss of habitat for tundra species [3]. The increase in precipitation in the north also leads to changes in freshwater systems throughout the region. In the south of Canada, the vegetation composition is also changing as the forests are facing increased stress due to droughts and reduced soil moisture. Wetlands are also experiencing shifts where reduced precipitation in some areas lead to them drying out whereas others may see expanded coverage due to wetter patterns [2].

All of these impacts place a stress on species as their environments are changing and may not be suitable for them anymore. This is leading to pressures on species to change their ranges to respond to the changing environments or simply leading to a population decline as species cannot respond as quickly as the environment is changing. For example, the moose is experiencing reduced habitat suitability in the south [4]. The World Wide Fund for Nature (WWF) has reported that between 1970 and 2020, the Living Planet Index (LPI), measuring the average size of monitored wildlife populations, has shrunk by 73% [5]. Further, there is a decline in the Biodiversity Intactness index, which measures how much original biodiversity remains within terrestrial communities, while the extinction rate is estimated to be "at least tens to hundreds of times higher than it would be in the absence of human activity" [5]. Alpine species and environments are particularly susceptible to climate change as they are adapted to very specific temperatures and habitat ranges, and thus face increased risks of habitat loss as they cannot shift upwards in elevation indefinitely [6, 7].

## II. Bottleneck

Understanding the distributions of species in the form of species distribution maps (SDMs) or range maps is vital to creating better conservation strategies. In particular, fine spatial and temporal resolution on the range map of species is needed to inform conservation strategies rather than a more national or regional understanding [8]. However, this is difficult due to the lack of quality data available for where these species occur, because the data is usually sparse, incomplete, or biased toward easily accessible areas, and because there is a greater abundance of presence-only data which limits the models that can be used. Moreover, it is generally difficult to create a fine-resolution map for species without including a deeper understanding of local environmental conditions.

Because of this, we focus on presence-only data, which tends to be easier to collect and more widely available than presence-absence data. We further focus on mammals in Banff and Jasper National Parks, as large mammals can serve as umbrella species for other alpine species [9, 10]. Umbrella species require large amount of land to survive, and so by protecting the land needed for these animals, many other species are protected as a positive side effect [9]. Mammals can also

serve as flagship species that can arouse public interest and sympathy to support conservation efforts [9, 10].

Hence to create a fine resolution species distribution model, we seek to develop a model that can predict encounter rates for mammals in Banff and Jasper national park using presence-only data from iNaturalist.

## III. Literature Review

A review of the relevant literature follows, discussing species distribution models, R-Tran, and handling presence-only data.

### A. Species Distribution Models

To create range maps for species, there are many non-machine learning based methods that have been used in the field. Namely, people will collect presence-absence or presence-only data through field surveys or other techniques, which will essentially be observations of the species scattered over some defined region. They will then use this data to create refined range maps from these scatter points across that defined region either using expert knowledge or simple logistic regression [11, 12, 13]. These models will include information about the climate such as temperature, precipitation, and other environmental information. This statistical modelling has been used for things such as habitat suitability models or ecological models [14]. Algorithms such as the Ecological Niche Factor Analysis have been used to model the distribution of species by matching it to certain environmental conditions that may correspond to the species habitat [15]. However, all of these techniques are limited by the assumptions we make about the species and the habitat it prefers which may be oversimplified in the non-machine learning techniques. It is further limited by the data quality, as more data would be needed for simple regression algorithms.

Machine learning techniques are introduced into this field to better capture the very complex relationship between all of the information we can gather on the environment and the species. First, simple machine learning models such as random forest models use presence-absence or presence-only data and environmental factors to create range maps [16]. In these cases again environmental information as well as occurrence data is used to determine the range maps. Random forest models have been seen to outperform other machine learning techniques in cases like creating better range maps for crane species in undersampled areas in Asia [17].

With the increased data availability of satellite imagery, there is a shift to models that incorporate satellite imagery of the environments to provide additional data about the habitats of species. Satellite imagery can be incorporated in two ways, one where a separate machine learning model takes the spatial information from the satellite imagery and has been trained to turn it into abstract feature embeddings. These can then be used as inputs into place into your species distribution model. Separately, you can directly use the satellite imagery in your species distribution model to make range predictions. For example, SatBird uses satellite imagery from Sentinel-2 coupled with presence-absence data to estimate encounter rates for bird species across the United States and parts of Kenya [18].

Regardless, these methods still perform poorly in regimes where little occurrence data is available for the target species. As a result, researchers now consider the joint modelling of multiple species in order to determine the range of a target species. The idea is that species should share similar ranges or will have some sort of dynamic with surrounding species that have potentially more data available. This information can be used to determine the range of the target species. This has been shown with the SatButterfly dataset [19], where the encounter rates of butterflies were determined using information about the occurrence of birds, satellite imagery, and presence-absence data.

### B. R-Tran

As explained in Section III-A, machine learning models are now integrating a variety of factors, whether environmental or species-related, in combination with satellite imagery to create accurate range maps of a target species. One of these models that has been previously used for this task is R-Tran [19]. In [19], R-Tran is used to create species range maps for both birds and butterflies using the SatBird [18] and SatButterfly [19] datasets. More broadly, R-Tran is capable of integrating a variety of abiotic environmental factors, satellite imagery, and related species data within a shared habitat to make species distribution predictions.

R-Tran, short for regression transformer, is a model developed for predicting species distribution patterns. R-Tran is capable of leveraging partial observational data from related species to make SDMs predictions of other species occupying the same environment and habitat [19]. The model takes in a variety of input data to make predictions such as Sentinel-2 satellite images as well as WorldClim climate-related data and SoilGrids soil-related data. WorldClim is a publicly available weather and climate database. It includes data

describing an environment's precipitation volume, solar radiation numbers, wind speed, water vapor pressure, and maximum, minimum, and mean temperatures [20]. This data can be used for environmental mapping. On the other hand, SoilGrids is a machine learning based system that handles and develops soil property maps. The maps include information on the soil's coarse fragments, its bulk density, its organic carbon, its clay, silt, and sand percentages, its pH level, and its cation exchange capacity [21]. SoilGrids maps combined with both Satellite images and WorldClim maps can be used to accurately model a given habitat and environment which allows us to learn the specifics of a species' habitat.

To accomplish the previously described species distribution task, R-Tran will take in as input WorldClim's bioclimatic rasters, SoilGrids' pedologic rasters, Sentinel-2's RGB satellite images as well as a set of target embeddings and state embeddings which both represent different types of species' observational data [19]. Target embeddings model all possible species classes that may be present while state embeddings represent the known or unknown state of a specific species in a given environment [19]. The images and rasters are fed to a ResNet-18 convolutional neural network (CNN) model, and then the CNN outputs a variety of features $I$ from the previously mentioned satellite and abiotic data [19]. Features $I$, Target embeddings $T$, and state embeddings $S$ are then finally concatenated all together as one. Once this is done, the concatenation of all embeddings and extracted features is used as input for a transformer encoder model [19]. A transformer encoder is used as it allows the relationships and relatedness between target classes and features to be modeled accordingly. Once the process of modelling target classes and features is finished, the model can make species checklist predictions [19].

In [19], the target and state embeddings are derived from eBird and eButterfly complete checklists. These checklists have citizen scientists record all species that they see or hear in a hotspot (a popular site) during their observation. This can be used to derive presence-absence data for these hotspots. In [19], this presence-absence data was calculated as follows:

1) Continental United States data for many different species was extracted between 2010 and 2023.
2) Each location's encounter rate $h$ was then calculated as follows: $\boldsymbol{y_h} = (y^h_{s_1}, \ldots, y^h_{s_n})$, where $y^h_{s_i}$ is the number of checklists where species $s_i$ was observed divided by the total number of checklists completed in $h$, for $i = 1, \ldots, n$.
3) Data related to a particular species was then combined over the 13 year period between 2010 and 2023. These are the finalized target encounter rates.

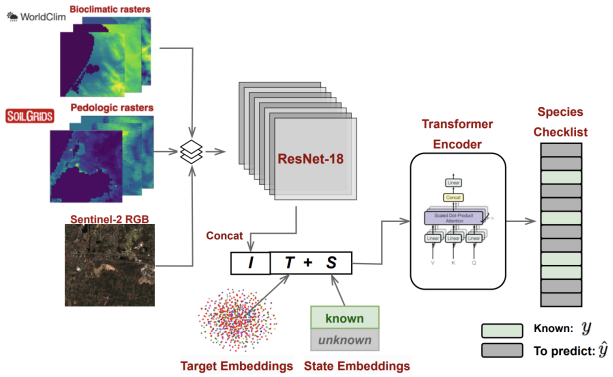



Fig. 1: R-Tran, proposed in [19], is a generalized model for creating SDMs where the input data is a variety of environmental characteristics, satellite images, and information about subset species in a given location. $T$ represents embeddings of all possible species classes and $S$ represents whether or not the state of a species is recognized. $T$ and $S$ embeddings are combined with the features generated from the environmental and satellite imagery data ($I$). Finally, this is fed to a transformer where predictions are then made.

During the initial training stage of the model, the authors of [19] provide all encounter rates barring a few that are randomly masked out and set to a value of "unknown". The percentage of total encounter rates that end up being randomly masked is chosen between $0.25n$ and $n$, $n$ being the total number of targets [19]. Furthermore, it is important to note that R-Tran can be used with or without partial species information as the model can still make predictions simply based on satellite and environmental data [19]. R-Tran's model architecture can be viewed in Figure 1.

### C. Presence-only data

Many methods for range mapping require presence-absence data, where species are labeled as either present or absent for an observation [22]. This includes R-Tran [19]. However, many large datasets are presence-only, where only the sightings of species are recorded. Presence-only data tends to be much easier to collect through citizen-science projects such as iNaturalist.

Models typically only work with presence-only or presence-absence data [22]. One of the most popular approaches for presence-only species distribution models is Maxent, which estimates the probability of observing a

species as a function of environmental factors [23, 22]. The other popular method for handling presence-only data is to generate pseudo-absences, artificial absences that can be based on environmental or geographic factors, as well as using data from other species [24], however, these methods have not consistently demonstrated improved performance over random sampling [25].

An alternative approach is to use target-species background points, where occurrences of other species can be used as pseudo-absences [25]. This has been shown to sometimes improve performance as the sampling of these other species may have similar biases, allowing for more precise differentiation between presence and absence [25].

Recently, [25] proposed to use target-group pseudo-absences alongside random pseudo-absences using the full weighted loss function

$$\begin{aligned} L_{full}\left(y,\hat{y}\right) = -\frac{1}{S}\sum_{s=1}^{S}[\mathbb{1}_{[y_s=1]}\lambda_1 w_s \log(\hat{y}_s) + \\ \mathbb{1}_{[y_s=0]}\lambda_2 \frac{1}{1-\frac{1}{w_s}} \log(1-\hat{y}_s) + \\ (1-\lambda_2)\log(1-\hat{y'}_s)], \end{aligned} \tag{1}$$

where $y_s$ is 1 if species $s$ is observed and 0 otherwise, $\hat{y}_s$ denotes the predicted suitability score for species $s$ between 0 and 1, $\hat{y}'_s$ is the model's prediction for species $s$ at a random location, $S$ is the number of species considered, and $\mathbb{1}_{[\cdot]}$ is the indicator function. This builds off of the full assumed negative loss proposed from [26], with the addition of the weighting terms $w_s$, $\lambda_1$, and $\lambda_2$ to address class imbalances between species, and prioritize different pseudo-absences differently [25].

The species weight $w_s$ accounts for the large class imbalance between different species and is defined for each species $s$ as

$$w_s = \frac{n}{n_{p(s)}}, \tag{2}$$

where $n_{p(s)}$ is the number of presence records for species $s$ and $n$ is the total number of presence locations in the training set [25]. The weighting $\lambda_1$ allows for more or less priority to be placed on correctly predicting presences, while $\lambda_2$ adjust the importance of random pseudo-absences and target-group pseudo-absences [25]. It is suggested that these weightings should be tuned to datasets, but as a baseline, it is suggested that $\lambda_1 = 1$ so that presences and absences are weighed similarly, and $\lambda_2 = 0.8$ as in practice target-group pseudo-absences tend to provide better representations of the species distribution [25].

## IV. Proposed Dataset

The proposed dataset combines satellite and environmental data with presence-only data from iNaturalist, and is, to our knowledge, a novel dataset for estimating encounter rates from iNaturalist data.

### A. Rasters for Satellite and Environmental Data

For the proposed methodology, we will need to include environmental data as inputs to predict the encounter rates of the species. We will include three types of environmental data: climate, soil, and satellite data. The climate data will be taken from WorldClim 1.4 at the 30 arcsec resolution [20]. This corresponds to a spatial resolution of 1km. This high resolution is required as we will be looking at a very small area, just two national parks, for our inferencing. We will average the data between the four seasons as we do not want to incorporate any temporal variation in our inferencing. There will be 19 bioclimatic factors that are taken which include things such as the annual mean temperature and annual precipitation. The data taken for the bioclimatic factors will be 1 value per hotspot for each of the 19 factors. As a result, we will only take the value corresponding to the specific longitude and latitude of the center of the hotspot cell that we define. An example of the bioclimatic data can be seen in Figure 4, where the annual mean temperature is plotted for the Jasper and Banff region from the 2.5 arcmin dataset.

For the soil data, data is taken from SoilGrids for 8 various soil properties [21]. The properties include factors such as bulk density, organic carbon density, and the soil's pH level. The resolution of the data is 250m and only the mean value is used at a depth of 0 to 5cm. Again, we only take the value corresponding to the specific longitude and latitude of the center of the hotspot cell that we define. An example of the soil grid pH value in the Banff and Jasper area can be seen in Figure 5.

Finally, for the satellite imagery, we will use Sentinel-2 images from the bands B02, B03, B04 and B08, which correspond to the RGB and NIR images. Once we have developed the cells for the hotspot, we will use the center location of the grids to extract the Sentinel-2 images using the planetary computer package. We will extract an image with a 1km by 1km bound corresponding to the hotspot grid. The images themselves have a 10m resolution. We will only choose images that are from the year 2020, as this is around the time that matches

our presence-only data. Moreover, we will choose the images that have less than 10% cloud cover. We will further choose an image from the summer as this will be most representative of what the landscape looks like without features getting washed out by snow cover. An example of a satellite image from Sentinel-2 in RGB is shown in Figure 6 of the Jasper and Banff area.

### B. Encounter Rates

To create species encounter rates, presence-only data from iNaturalist will be used, alongside target-group pseudo-absences by creating checklists when any species is observed. The following pipeline will be followed for creating these checklists:

1) Occurrences will be filtered to mammals in the time range of 2010-2023 in the geographic region of the study.
2) The study area will be divided into hexagonal H3 cells with resolution 7, corresponding to approximately 5.16 $\mathrm{km}^2$ [27].
3) All areas with a total number of occurrences less than $n_{min}$, an empirically determined threshold, will be ignored. Each remaining area will be enumerated.
4) For each remaining area $h$,
   - Create a checklist for each month in which observations were made. Record what species were seen or not seen in that month in $h$.
   - Calculate $\boldsymbol{y_h} = (y^h_{s_1}, \ldots, y^h_{s_n})$, where $y^h_{s_i}$ is the number of checklists where species $s_i$ was observed divided by the total number of checklists completed in $h$, for $i = 1, \ldots, n$.

### C. Contribution

This preprocessing pipeline provides a dataset that can be used for new predictive tasks that may provide better species distribution models. However, this estimate for encounter rates is imperfect and has several limitations, discussed in Sections VII-A and IX.

## V. Proposed Methodology

We propose to use R-Tran [19] to predict encounter rates, using a loss function $L_{weighted}$ inspired by $L_{full}$ [25].

### A. Loss Function

The weighted loss is defined as

$$L_{weighted}(y, \hat{y}) = -\frac{1}{N_h}\sum_{h=1}^{N_h}\sum_{s=1}^{S}[\lambda w_s y_h^s \log(\hat{y}_s^h) + \frac{1}{1-\frac{1}{w_s}}(1-y_h^s)\log(1-\hat{y}_s^s)], \tag{3}$$

where $y_s^h$ is the actual encounter rate and $\hat{y}_s^h$ is the predicted encounter rate for species $s$ in location $h$ between 0 and 1. $N_h$ is the number of locations considered, $S$ is the total number of species studied, and $w_s$ is defined for each species $s$ as

$$w_s = \frac{N_h}{N_h^s}, \tag{4}$$

where $N_h^s$ is the total number of locations where $s$ has a nonzero encounter rate. This is done to mitigate the effect of the class imbalance between species.

This loss is similar to 1, but with $\lambda_2 = 1$, and adjusted for our case where we will be predicting encounter rates in specific locations.

Different definitions of $w_s$ should be explored, and $\lambda$ will need to be fine-tuned for this task. The performance will be compared to the model trained on cross-entropy loss, as in the original experiments with R-Tran [19]. More research should be conducted to find other potentially better-suited loss functions.

### B. Contribution

Using R-Tran in combination with the previously described loss function may improve performance for species that have less data, and allows for the tuning of R-Tran to be more or less sensitive to low encounter rates. This can help generalize R-Tran to be used on presence-only data. Beyond this, applying R-Tran to data from iNaturalist is novel and may provide better performance than existing models for tasks like generating range maps due to the inclusion of satellite imagery, abiotic environmental factors and joint distributions of related species within a shared environment and habitat.

Some pitfalls and potential limitations may arise by using the methodology described in this section. The major limitations are discussed in Section IX.

## VI. Preliminary Work

An initial small sample of the occurrences dataset was created for the region of Banff and Jasper National Parks. This used only data directly surrounding the parks to demonstrate the preprocessing steps involved in generating the encounter rates and limits the species studies to

only mammals that have occurrence data in this region from iNaturalist. Figure 2 shows the filtering of the data to determine the spots with enough occurrence data, while Figure 3 shows a sample set of occurrence data for *Ursus americanus*

The preprocessed encounter rate data can be found at https://github.com/BasilJ-S/RangeMapEstimation.

## VII. Next Steps

The next steps that should be taken to progress this project are described in this section.

### A. Dataset

The current dataset includes all seasons indiscriminately, potentially missing differences in species distributions in different seasons. A next step would be to divide the dataset between summer and winter, as in [18].

However, this is likely to run into challenges as the initial dataset is significantly smaller than would be needed for accurate predictions. The initial filtering reduced the number of valid points to 57. As such we will discuss several avenues that can be explored to gather more data, both with iNaturalist data and using other data sources.

Using iNaturalist data, the study area can be increased to include all of the rocky mountains or all of North America, as this should still provide useful data for the region of Banff and Jasper. The species within the dataset could also be expanded beyond mammals so that more information can become available from co-occurrences.

Outside of iNaturalist, there are several other data collection methods that could be incorporated. Parks Canada has many camera traps in national parks, which could be used as additional data for cells [28]. Further, if the dataset were increased to include all of North America integrating this dataset into the SatBird and SatButterfly datasets should be explored, as co-occurrences may provide an increase in performance to models [19, 18].

### B. Benchmarks

Appropriate benchmarks need to be selected, both to evaluate R-Tran and to assess the quality of the dataset.

To assess R-Tran, several baselines were suggested in [18], including predicting the average of encounter rates over the training set and using Gradient Boosted Regression Trees on the encounter rates and data extracted from the environmental rasters. The baselines from [19] can also be used, which are ResNet18 [29] and Feedbackprop [30].

To assess the dataset, these baseline models and R-Tran can be tasked with creating range maps for different species, by finding all locations where the encounter rate is significantly greater than 0. This can then be compared to current "ground-truth" expert-made range maps from [31]. These range maps can indicate whether the dataset is providing valuable information.

Further research should be conducted to determine more adequate baselines.

### C. Future Uses

Beyond the application of this model to Banff and Jasper National Park, future work would attempt to apply similar models to low-data regions, as this model should be better suited for these regions due to the wide accessibility of satellite data. This could provide a significant benefit to the conservation efforts of these regions.

## VIII. Pathways to Impact

The proposed model should provide high-resolution species distribution maps for mammal species for the current climate. These maps will highlight regions that need increased protection in Banff and Jasper National Parks.

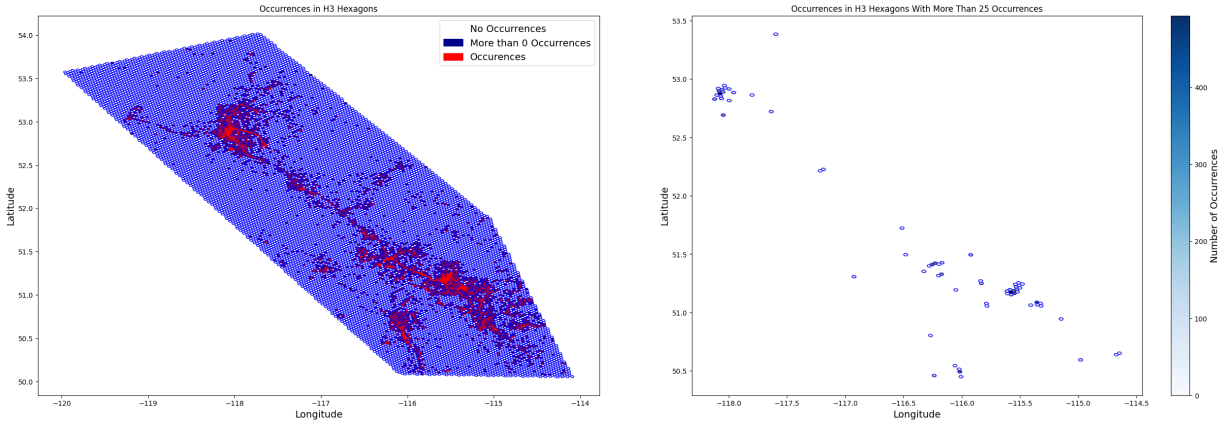
### A. Stakeholders

The main stakeholders who could be influenced by these species distribution maps are:

- Parks Canada, which has initiatives to preserve the biodiversity of Jasper and Banff National Parks, including efforts to improve the connectivity of the parks for land-based species [32].
- Conservation organizations including the National Conservancy of Canada.
- Indigenous peoples who have lived on the land and are partners of Parks Canada [32].
- Environmental science and machine learning researchers.
- Provincial and federal governments that can use this data for policy decisions affecting land use.
- Industry, which may want to use the area around these parks for resource extraction or development.

### B. Stakeholder-Specific Considerations

In this section, we analyze how environmental, financial, and social factors might influence stakeholders.




(a) Map dividing the study area into H3 hexagons, showing all occurrences for mammals and noting any cells that have more than 0 occurrences.

(b) Cells filtered to only include cells with more than 25 occurences.

Fig. 2: Initial preprocessing step, where the study area is initially divided into H3 cells and the occurrences are binned into those cells. We then take only the cells with more than 25 occurrences, to limit areas that are likely to provide false occurrence rates of 0.

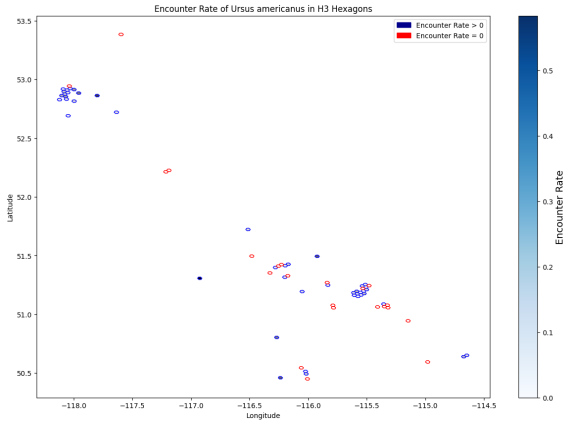



Fig. 3: Visualization of the estimated encounter rates for *Ursus Americanus*, commonly the American black bear. This demonstrates potential issues in the dataset, as there are several locations with an encounter rate of 0 that are directly next to areas with a relatively high encounter rate.

*1) Environmental Considerations:* Parks Canada has the greatest potential for environmental impact in this project, as species distribution maps could directly help their conservation efforts by indicating the regions most important for different species. Parks Canada also communicates with several other mentioned stakeholders, so this could lead to wide dissemination of the models.

*2) Financial Considerations:* There are several financial constraints to consider in driving impact with these models.

Firstly, the spending of Parks Canada is planned to shrink to 62% of its amount from 2024-2025 by the fiscal year of 2026-2027 [32]. This may limit the ability of Parks Canada to lead conservation efforts in parks and increase the wariness of unneeded spending. As such, the model should be benchmarked against existing models to ensure their usefulness before involving these stakeholders.

Secondly, this model would ideally lead to the protection of land that is needed for the survival of the studied species in the region. This could impact industries seeking to use the land for other purposes. This further motivates the need for the reliability of results so that they can be used in a regulatory context by provincial and federal governments.

These financial factors also show that providing results to other environmental agencies such as the National Conservancy of Canada may help to bolster conservation given the decreased resources of Parks Canada and the potential pressure on regulators from industry.

*3) Social Considerations:* The methods described in this report can further strengthen the ability of citizen science projects like iNaturalist to have an impact. This in turn can encourage more people to participate in citizen science. Increasing engagement in these projects increases the quantity of data that can be gathered, and may also lead to further engagement in issues surrounding conservation.

### C. Path

The model should first be documented and made open source to enable future work. The proposed dataset can be published online to enable future modelling. However, considering all of the stakeholders, the pathway to impact that will be explored is through Parks Canada due to the direct potential for impact.

Preliminary results should be shared with Parks Canada to gather feedback and determine what models would be most helpful for their conservation efforts. From this feedback, species distribution models can be generated for specific species and benchmarked for tasks from Parks Canada, and the model can be tuned to provide the best results for these tasks. These results can then be provided to Parks Canada with the model parameters and training information to enable further modelling.

## IX. Limitations and General Considerations

There are some limitations and general considerations for our proposed methodology. As stated in Section V, we are generating pseudo-absence data to mimic the performance of real absence data. Although [25] shows promising results in creating realistic pseudo-absence data, this is still an assumption we are making and a limitation of our methodology. Ideally, we would be using true absence data, but that is not available.

Further analysis should be conducted of the dataset to determine if the estimated encounter rates are valuable. By considering each checklist across a month, the current method attempts to approximate focused checklists by assuming that multiple partial checklists can constitute a full checklist. However, due to the relatively small number of observations in each H3 cell as discussed in Section VII-A and inherent bias in what species are recorded, this may not be true. Adjustments such as increasing the region size, increasing the time range, or recording the encounter rate as simply the number of occurrences of a species divided by the total occurrences in a cell should all be explored.

Furthermore, we make two important assumptions in our work that are somewhat contradictory. We assume that species distributions are related, but, through our target-group pseudo-absences, we also assume that the presence of one species implies the absence of another. As a simplified example of how this could be contradictory, one could assume that wolves will tend to be in the same areas as one of their prey, deer. This would mean that if we see a wolf, we would expect that there are also deer nearby. However, if deer are not seen in that period in the area, this would be recorded as an absence of deer by our current preprocessing pipeline. This may have the effect of limiting the performance gained from multi-species modelling.

Moreover, creating species distribution maps to model species occurrences is a topic that is of relevance to a wide range of stakeholders and communities as we saw in Section VIII. Depending on the stakeholder or community, the usage and needed output of the model may change. Therefore, it is important to keep these potential differences in mind and to set up dialogues with these different stakeholders, actors, and communities as early as possible.

Finally, beyond the scope of the work proposed, it would be interesting to integrate a time component or value within the R-Tran model. Environments and habitats change over time, whether that be over years or months due to the changing seasons. Moreover, seasons can also heavily affect the behaviour of certain species which can change their occurrence patterns and maps. Additionally, some of these habitat changes or animal behaviour changes have been accelerated due to the effects of climate change and it would be important to keep track of these accelerated changes through some sort of time metric. Therefore, by having an added time component within the model, we would be able to keep track of this sort of data which would allow us to make more accurate predictions by incorporating and keeping in mind the passage of time and its various effects on animals and their habitats.

## References


[1] Camille Parmesan and Gary Yohe. "A globally coherent fingerprint of climate change impacts across natural systems". In: *Nature* 421.6918 (Jan. 2003). Publisher: Nature Publishing Group, pp. 37–42. ISSN: 1476-4687. DOI: 10.1038/nature01286. URL: https://www.nature.com/articles/nature01286 (visited on 12/10/2024).

[2] E Bush and D S Lemmen. *Canada's changing climate report*. en. 2019. DOI: 10.4095/314614. URL: https://ostrnrcan-dostrncan.canada.ca/handle/1845/143725 (visited on 12/04/2024).

[3] Logan T. Berner et al. "Summer warming explains widespread but not uniform greening in the Arctic tundra biome". en. In: *Nature Communications* 11.1 (Sept. 2020). Publisher: Nature Publishing Group, p. 4621. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18479-5. URL: https://www.nature.com/articles/s41467-020-18479-5 (visited on 12/04/2024).

[4] Dennis L. Murray et al. "Continental divide: Predicting climate-mediated fragmentation and biodiversity loss in the boreal forest". en. In: *PLOS ONE* 12.5 (May 2017). Publisher: Public Library of Science, e0176706. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0176706. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176706 (visited on 12/04/2024).

[5] WWF. *Living Planet Report 2024 - A System in Peril*. Gland, Switzerland: WWF, 2024.

[6] ROBIN ENGLER et al. "21st century climate change threatens mountain flora unequally across Europe". In: *Global Change Biology* 17.7 (2011), pp. 2330–2341. DOI: https://doi.org/10.1111/j.1365-2486.2010.02393.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2010.02393.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2010.02393.x.

[7] Yohann Chauvier-Mendes et al. "Transnational conservation to anticipate future plant shifts in Europe". In: *Nature Ecology & Evolution* 8.3 (Mar. 2024), pp. 454–466. ISSN: 2397-334X. DOI: 10.1038/s41559-023-02287-3. URL: https://www.nature.com/articles/s41559-023-02287-3 (visited on 12/10/2024).

[8] Laura J. Pollock et al. "Protecting Biodiversity (in All Its Complexity): New Models and Methods". In: *Trends in Ecology & Evolution* 35.12 (2020), pp. 1119–1128. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2020.08.015. URL: https://www.sciencedirect.com/science/article/pii/S0169534720302305.

[9] Robin Steenweg et al. "Testing umbrella species and food-web properties of large carnivores in the Rocky Mountains". In: *Biological Conservation* 278 (Feb. 1, 2023), p. 109888. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2022.109888. URL: https://www.sciencedirect.com/science/article/pii/S0006320722004414 (visited on 11/02/2024).

[10] Daniel Simberloff. "Flagships, umbrellas, and keystones: Is single-species management passé in the landscape era?" In: *Biological Conservation*. Conservation Biology and Biodiversity Strategies 83.3 (Mar. 1, 1998), pp. 247–257. ISSN: 0006-3207. DOI: 10.1016/S0006-3207(97)00081-5. URL: https://www.sciencedirect.com/science/article/pii/S0006320797000815 (visited on 11/02/2024).

[11] Antoine Guisan, Thomas C Edwards, and Trevor Hastie. "Generalized linear and generalized additive models in studies of species distributions: setting the scene". In: *Ecological Modelling* 157.2 (Nov. 2002), pp. 89–100. ISSN: 0304-3800. DOI: 10.1016/S0304-3800(02)00204-1. URL: https://www.sciencedirect.com/science/article/pii/S0304380002002041 (visited on 12/04/2024).

[12] Antoine Guisan et al. "Making better biogeographical predictions of species' distributions". en. In: *Journal of Applied Ecology* 43.3 (2006), pp. 386–392. ISSN: 1365-2664. DOI: 10.1111/j.1365-2664.2006.01164.x. URL: https://onlinelibrary.wiley.com/doi/abs/%5Callowbreak10.1111%5Callowbreak/j.1365-2664.%5Callowbreak2006.01164.x (visited on 12/04/2024).

[13] Antoine Guisan and Wilfried Thuiller. "Predicting species distribution: offering more than simple habitat models". en. In: *Ecology Letters* 8.9 (2005), pp. 993–1009. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2005.00792.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2005.00792.x (visited on 12/04/2024).

[14] Veronika Braunisch et al. "Living on the edge—Modelling habitat suitability for species at the edge of their fundamental niche". en. In: *Ecological Modelling* 214.2-4 (June 2008), pp. 153–167. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2008.02.001. URL: https://linkinghub.elsevier.

com/retrieve/pii/S0304380008000598 (visited on 12/04/2024).

[15] A. H. Hirzel et al. "ECOLOGICAL-NICHE FACTOR ANALYSIS: HOW TO COMPUTE HABITAT-SUITABILITY MAPS WITHOUT ABSENCE DATA?" In: *Ecology* 83.7 (2002), pp. 2027–2036. DOI: https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/0012-9658%282002%29083%5B2027%3AENFAHT%5D2.0.CO%3B2. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658%282002%29083%5B2027%3AENFAHT%5D2.0.CO%3B2.

[16] Anantha M. Prasad, Louis R. Iverson, and Andy Liaw. "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction". en. In: *Ecosystems* 9.2 (Mar. 2006), pp. 181–199. ISSN: 1435-0629. DOI: 10.1007/s10021-005-0054-1. URL: https://doi.org/10.1007/s10021-005-0054-1 (visited on 12/04/2024).

[17] Chunrong Mi et al. "Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence". In: *PeerJ* 5 (Jan. 2017), e2849. ISSN: 2167-8359. DOI: 10.7717/peerj.2849. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5237372/ (visited on 12/04/2024).

[18] Mélisande Teng et al. "Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data". In: *Advances in Neural Information Processing Systems* 36 (2024).

[19] Hager Radi Abdelwahed, Mélisande Teng, and David Rolnick. "Predicting Species Occurrence Patterns from Partial Observations". In: (2024). Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.2403.18028. URL: https://arxiv.org/abs/2403.18028 (visited on 12/04/2024).

[20] Stephen E. Fick and Robert J. Hijmans. "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". en. In: *International Journal of Climatology* 37.12 (2017), pp. 4302–4315. ISSN: 1097-0088. DOI: 10.1002/joc.5086. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086 (visited on 12/10/2024).

[21] Niels H. Batjes, Eloi Ribeiro, and Ad van Oostrum. "Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019)". English. In: *Earth System Science Data* 12.1 (Feb. 2020). Publisher: Copernicus GmbH, pp. 299–320. ISSN: 1866-3508. DOI: 10.5194/essd-12-299-2020. URL: https://essd.copernicus.org/articles/12/299/2020/ (visited on 12/10/2024).

[22] Sara Beery et al. "Species Distribution Modeling for Machine Learning Practitioners: A Review". In: *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. COMPASS '21: ACM SIGCAS Conference on Computing and Sustainable Societies. Virtual Event Australia: ACM, June 28, 2021, pp. 329–348. ISBN: 978-1-4503-8453-7. DOI: 10.1145/3460112.3471966. URL: https://dl.acm.org/doi/10.1145/3460112.3471966 (visited on 11/03/2024).

[23] Stephen J Phillips, Miroslav Dudík, and Robert E. Schapire. *Maxent software for modeling species niches and distributions*. URL: http://biodiversityinformatics.amnh.org/open_source/maxent/.

[24] Joséphine Broussin, Maud Mouchet, and Eric Goberville. "Generating pseudo-absences in the ecological space improves the biological relevance of response curves in species distribution models". In: *Ecological Modelling* 498 (Dec. 1, 2024), p. 110865. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2024.110865. URL: https://www.sciencedirect.com/science/article/pii/S0304380024002539 (visited on 12/10/2024).

[25] Robin Zbinden et al. *On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning*. Jan. 3, 2024. DOI: 10.48550/arXiv.2401.02989. arXiv: 2401.02989. URL: http://arxiv.org/abs/2401.02989 (visited on 12/05/2024).

[26] Elijah Cole et al. "Spatial Implicit Neural Representations for Global-Scale Species Mapping". In: *Proceedings of the 40th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 3, 2023, pp. 6320–6342. URL: https://proceedings.mlr.press/v202/cole23a.html (visited on 12/05/2024).

[27] Uber Open Source. *H3*. Version 4.x. original-date: 2017-12-21T01:38:35Z. Dec. 8, 2024. URL: https://github.com/uber/h3 (visited on 12/08/2024).

[28] Parks Canada. *Mammals - Jasper - Open Government Portal*. 2018. URL: https://open.canada.ca/data/en/dataset/932b4e4d-9c09-4864-8646-24db93fc1b36/resource/946766c4-c5bf-415d-826e-20ba9a27ef0a (visited on 11/04/2024).

[29] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385[cs]. URL: http://arxiv.org/abs/1512.03385 (visited on 12/10/2024).

[30] Tianlu Wang, Kota Yamaguchi, and Vicente Ordonez. "Feedback-Prop: Convolutional Neural Network Inference Under Partial Evidence". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 898–907. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Feedback-Prop_Convolutional_Neural_CVPR_2018_paper.html (visited on 12/10/2024).

[31] *The IUCN Red List of Threatened Species*. IUCN Red List of Threatened Species. URL: https://www.iucnredlist.org/en (visited on 12/10/2024).

[32] Parks Canada. *Parks Canada Departmental plan for the fiscal year 2024 to 2025*. Last Modified: 2024-11-01. Government of Canada, Feb. 13, 2024. URL: https://parks.canada.ca/agence-agency/bib-lib/plans/dp/coupdoeil-2024-2025-glance/plan-ministeriel-departmental-plan (visited on 12/10/2024).

## APPENDIX

Figures 4, 5 and 6 show sample environmental rasters for the Banff and Jasper regions.

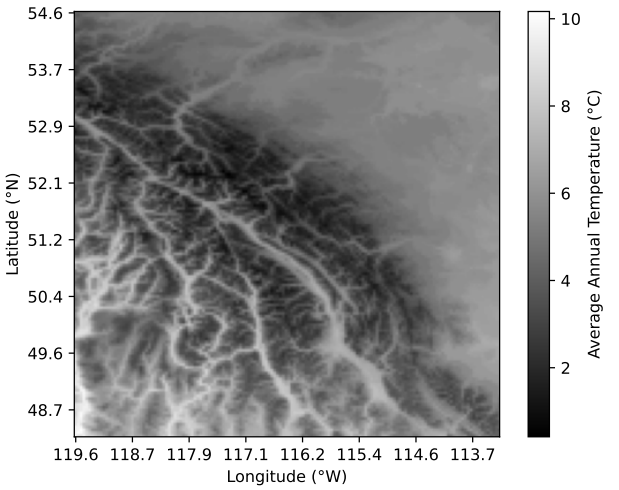



Fig. 4: This shows the annual mean temperature of the Banff and Jasper region from WorldClim.

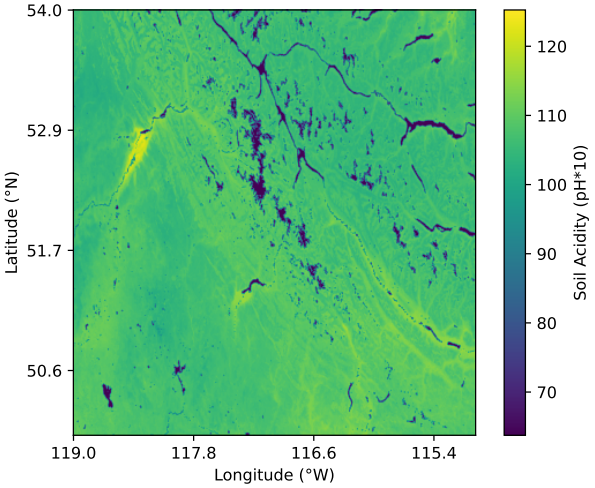



Fig. 5: This shows the soil acidity levels of the Banff and Jasper region from SoilGrids.

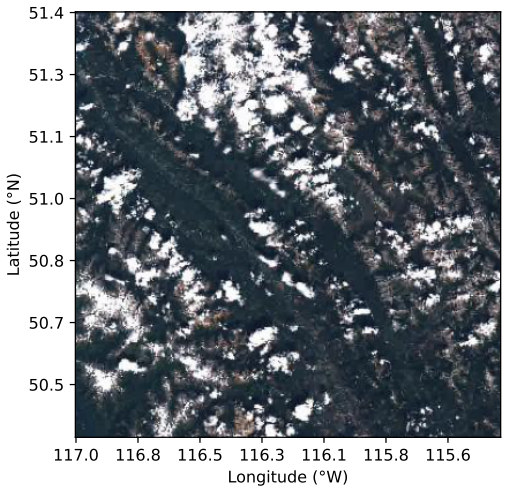



Fig. 6: This shows the RGB data from the Sentinel-2 satellite of the Banff and Jasper region.