Biodiversity Data Challenge

Biodiversity Challenge Objectives

The Biodiversity Challenge, referred to as the “Frog Challenge”, is aimed at participants with beginner or intermediate skills in data science and analytics. The goal of the project is to build a machine learning classification model to predict the presence of frog species based on TerraClimate [1] variables extracted from the Microsoft Planetary Computer data catalogue. The TerraClimate dataset provides global monthly climate and water balance variables from 1958 to the present.

Participants will be given locations (latitude and longitude) of frog presence across a portion of southeastern Australia over a period of 2 years from November 2017 to November 2019. This data will be used as the “target variable” in a machine learning model. Data from TerraClimate will be used as the “predictor variable” to train a machine learning model. In the end, this model will be used to predict the presence or non-presence of frogs in specific locations.

The importance of frog biodiversity

The presence of frogs is an important sign of a healthy ecosystem. Scientists believe frogs are an environmental bellwether with declines in their population viewed as early warning signs of environmental damage [2]. Thus, the study of frog habitats is increasingly important to understand the extent and severity of global environmental change. Frog populations have declined significantly since the 1950s and more than one third of species are considered to be threatened with extinction and over 120 are believed to have become extinct since the 1980s.

While in their larval stage, frogs keep waterways clean by feeding on algae. During their adulthood, they reduce the transmission of fatal illnesses by eating large amounts of insects such as mosquitos. Conversely, they are an important part of the animal ecosystem as prey and help other large predators survive. For the aforementioned reasons, scientists and researchers use the presence of frogs as a measure of biodiversity health.

The use of satellite data for detecting frog presence

Researchers and conservation organizations use satellite data to monitor habitats and ecosystems, aiming to identify factors that influence frog presence across different regions. However, like in most ecological studies, a perfect predictive model remains elusive. Advances in remote sensing and machine learning, along with the growing availability of satellite data, continue to push this field forward. By participating in this data challenge, you have the opportunity to develop a model that could contribute to our understanding of amphibian habitats and aid in biodiversity conservation efforts.

One of the valuable satellite data sources for ecological studies is **TerraClimate**. Launched to provide global climate and water balance data, TerraClimate offers monthly climate data at a high spatial resolution of 4 kilometres, dating back to 1958. TerraClimate data includes **14 variables** which are essential for assessing environmental factors affecting frog populations. Because frogs are highly sensitive to climatic conditions and moisture availability, TerraClimate data enables researchers to track habitat suitability with greater precision. There are 6 primary "measured" climate variables in the TerraClimate dataset. These include: maximum temperature (tmax), minimum temperature (tmin), vapor pressure (vap), precipitation accumulation (ppt), downward surface shortwave radiation (srad), and wind-speed (ws). The remaining 8 variables are "derived" parameters and include: actual evapotranspiration (aet), reference evapotranspiration (pet), runoff (q), climate water deficit (def), soil moisture (soil), snow water equivalent (swe), Palmer Drought Severity Index (pdsi), and vapor pressure deficit (vpd).

One unique aspect of TerraClimate data is that it accounts for both climatic variations and hydrological balance. This provides valuable insights into habitat stability over time, which is crucial for studying species dependence on water availability and temperature range. Although TerraClimate data is available on various platforms, like Google Earth Engine, its integration with detailed topographic and climate-adjusted variables on the Microsoft Planetary Computer enhances usability for ecological modelling. In the context of frog habitat studies, TerraClimate data can serve as a foundational layer, helping researchers assess regions of potential frog presence based on climate-driven habitat suitability.

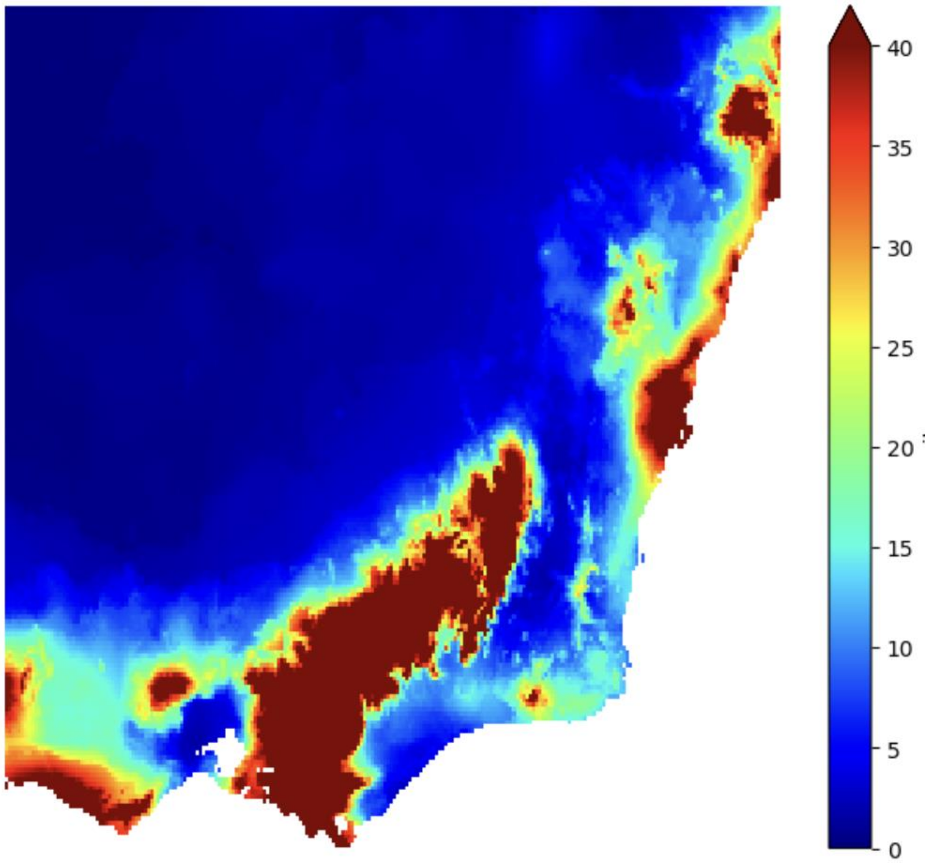

Figure 1. *Soil moisture is a known proxy for frog presence. This figure shows the median soil moisture variability across southeastern Australia for the period of November 2017 through November 2019. Such data could be important for frog presence modeling*.

Frog Dataset and Analysis Region

The data for this challenge originates from the Global Biodiversity Information Facility (GBIF), an international organization that provides open access to biodiversity data contributed by governments and institutions worldwide [3]. GBIF's mission is to support research and conservation by making comprehensive biodiversity data available for analysis, facilitating a deeper understanding of species distribution and ecological trends globally.

The dataset represents expert-validated occurrence records of calling frogs across Australia collected via the national citizen science project FrogID [4]. FrogID relies on participants recording calling frogs using smartphone technology, after which point the frogs are identified by expert validators, resulting in a database of georeferenced frog species records. The total dataset (see Figure 2) includes 771,542 records of 218 species which is 86% of the known frog species in Australia.

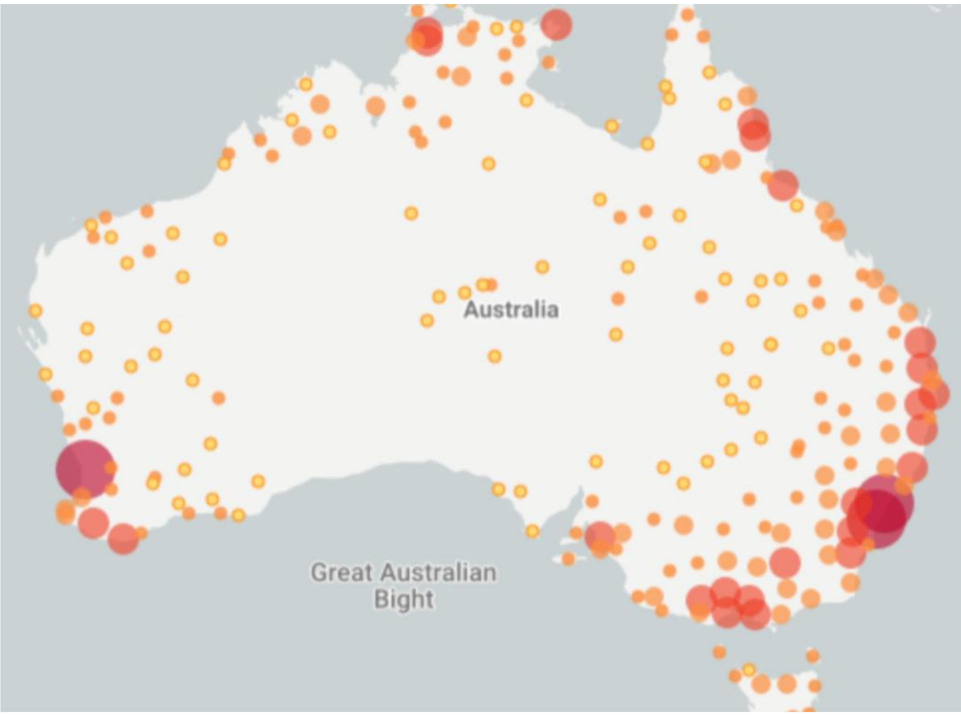

***Figure 2**. The FrogID dataset includes 771,542 records across Australia. This data challenge will focus on a subset of the data from the southeastern portion of Australia. This sub-region has excellent diversity and large frog populations.*

The specific dataset for this challenge (see Figure 3) focuses on frog presence across southeastern Australia, spanning a two-year period from November 2017 to November 2019. A complementary frog non-presence dataset was developed using random equally spaced locations across the region. It should be noted that the frog non-presence dataset could be adjusted by challenge participants to improve results. This data is not actually "measured" in the field but reflects random locations that are not coincident with measured frog presence locations.

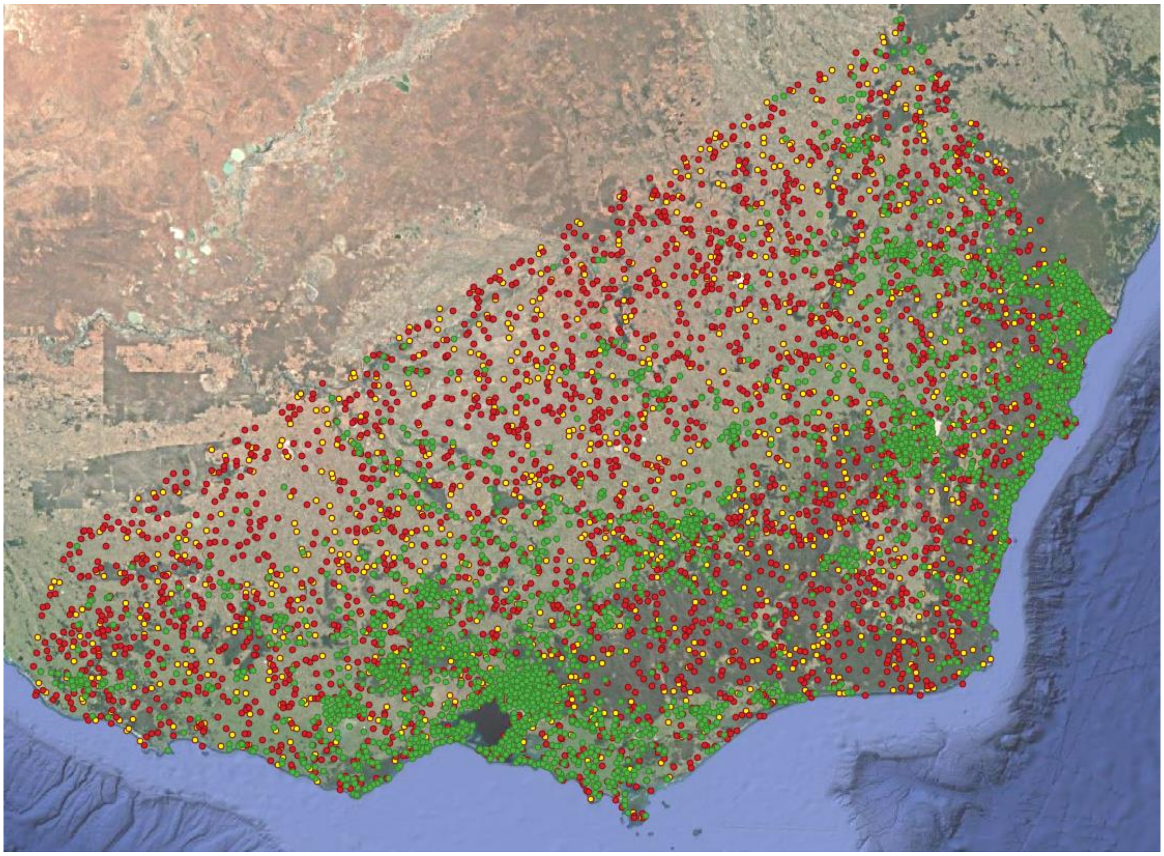

***Figure 3**. The dataset for this challenge is focused on a region in southeastern Australia. There are 3792 frog presence locations (GREEN), 2520 frog non-presence locations (RED) and 2000 validation locations (YELLOW) to test model predictions.*

References

[1] **TerraClimate** – Microsoft Planetary Computer – LINK HERE

[2] Wikipedia – Frogs – LINK HERE

[3] GBIF - **Global Biodiversity Information Facility** – LINK HERE

[4] Rowley JJL, Callaghan CT (2020) **The FrogID dataset: expert-validated occurrence records of Australia's frogs collected by citizen scientists**. ZooKeys 912: 139–151. https://doi.org/10.3897/zookeys.912.38253

[5] Campbell, K.S., Baltensperger, A.P. & Kerby, J.L. **Random Frogs: using future climate and land-use scenarios to predict amphibian distribution change in the Upper Missouri River Basin**. *Landsc Ecol* 39, 61 (2024). https://doi.org/10.1007/s10980-024-01841-z

Participant Package Contents

1. Participant Overview (**Biodiversity_Challenge_Overview.docx**) – this document
2. Benchmark Notebook (**Biodiversity_Challenge_Benchmark.ipynb**) – a Python notebook that provides a starting point for participants
3. TerraClimate Sample Notebook (**TerraClimate.ipynb**) – a Python notebook that outputs a GeoTIFF file with specific climate variables. This output is used by the Benchmark Notebook.
4. Requirements (**requirements.txt**) – a text file listing Python package dependencies for the benchmark notebook execution
5. Training Dataset (**Training_Data.csv**) – a CSV file containing **3792 frog presence** (Occurrence Status =1) and **2520 frog non-presence** (Occurrence Status =0) locations (latitude and longitude) for model training
6. Validation Dataset (**Validation_Template.csv**) – a CSV file containing **2000 locations** (latitude and longitude) for participant model validation testing

==

Instructions for Participants/Students

a) Understand the problem statement and dataset

The objective of this data challenge is to predict the presence, or non-presence, of frog species at given locations in southeastern Australia. This exercise will use frog presence data over a period of 2 years from Nov-2017 to Nov-2019. Please refer to the file "*Training_Data.csv*" to find training data consisting of 3792 frog presence locations and 2520 frog non-presence locations.

b) Set up the development environment

Building a machine learning model requires a Python notebook development environment and a Python computing environment with the appropriate libraries. Please refer to the file "*requirements.txt*", which contains a list of necessary Python libraries required to run the benchmark notebook. You can use a local computing environment (e.g., PC, Mac) or any cloud environment (e.g., Azure Cloud, Google Cloud). For reference, the sample notebooks have been successfully tested on a local laptop computer with 4 cores and 32GB memory, so a paid cloud environment is not required.

c) Build the model

Once your development environment is ready, you need to create a Python notebook and start building the model. Participants may want to install “JupyterLab” to manage notebook development on a local computer. Please refer to the file “*Biodiversity_Challenge_Benchmark.ipynb*”, where we have demonstrated a basic machine learning workflow designed to help students gain practical experience. The focus of this example is to predict the presence of frog species using two features from the TerraClimate dataset as predictor variables. This notebook is simplified and intended as a learning tool for students to experiment with remote sensing data and machine learning techniques. The model should run very quickly (~minutes) on any local computer. This notebook should be considered as a starting point only and students should extend this model to build a more robust model by running multiple experiments.

d) Make the prediction on the validation dataset

After building the model, students need to validate the model performance on a dataset with different locations. To validate the model, students need to use their model to make predictions about the presence or non-presence of frogs for a set of 2000 test coordinates (latitude and longitude) we have provided in the "*Validation_Template.csv*" file and share the predicted results with the instructor/professor. For reference, the benchmark notebook will produce a starting F1-score of 0.6949. Participants should attempt to maximize this score up to an ideal value of 1.0.