kelpGeoMod User Guide

What is the kelpGeoMod User Guide

This guide outlines the kelpGeoMod Pipeline for effectively understanding the project's data structure and leveraging the tools for expanding your kelp research according to your needs.

What are the kelpGeoMod Data

To access the metadata, descriptions, and understanding of the data, please refer to this link.This project was completed in the R language.

Why was this kelpGeoMod User Guide created?

This user guide pipeline was created to offer a synthesized dataset of kelp and ocean data in the Santa Barbara Channel from various long-term monitoring efforts. This data set is permanently stored and accessible to all. This user guide aims to assist kelp researchers with assimilating their own data observations and integrating their own data with our synthesized data sets. By providing explicit instructions, R Markdowns, README files, and organized data storage, this project enhances data integration efficiency and accessibility, fostering a conducive research environment for significant discoveries.

Who is this user guide designed for?

The kelpGeoMod user guide aims to assist researchers and kelp farmers with a clear pipeline, expanding their understanding of the data in the Santa Barbara Channel (SBC). It provides guidance on how to merge and integrate external data with our synthesized data to address the several interest.

Where are the instructions?

The instructions can be found in the document below. We highly recommend reviewing our technical documentation to understand the content and structure of the kelpGeoMod project data for effective future use.

Where are the project data located?

To access the kelpGeoMod data through our publicly available Google Shared Drive, please refer to this link.

Where is this project code located?

To access the kelpGeoMod code through our publicly available GitHub repository, please refer to this link.

Where is this project area of interest?

The project encompasses data gathered, cleaned, and arranged within the Santa Barbara Channel. The coordinates used to delimit our area of interest are 33.85°- 34.59° N, 118.8°- 120.65°W. If your research extends beyond this area, please refer to the guide and tips section in the provided document.

When was this kelpGeoMod pipeline created?

This pipeline was developed for the kelpGeoMod final Capstone Project, submitted to the Bren School of Environmental Science & Management in **June 2023**.

Who created this kelpGeoMod pipeline?

Erika Egg, Jessica French, Javier Patrón, Elke Windschitl

What are the dates for the data coverage?

The collection period for data pertaining to this project extends from January 1st, 2014, to December 31st, 2023.

Overall Pipeline Outline:

The steps below are designed to provide an overview of the project data and how you can put it to use.

1. **Understand the Current kelpGeoMod Data Structure:** Please visit our GitHub repository and go through the general README to get a grasp of the project structure. Explore more READMEs within the folder structure of our GitHub repository for more specific information pertaining to the files in those folders.
2. **Define Your Goal:** Determine how you can use this project data and pipeline to benefit your kelp question. Set a clear objective for your desired resolution, extent, and position, which will provide a strong base for data resampling and cropping.
3. **Follow Instructions for Data Integration:** Use our guidelines to integrate and utilize our data effectively.

Guide and tips on how to properly merge new data to the synthesized data set:

1. **Define your kelp research question:** Clearly understand your specific research question or objective related to our kelp study.
2. **Identify the data file you want to merge and clean:** Determine the type of data file you are working with for the merge process.

Once you have clearly defined your kelp research question please refer to that step depending on the data file you want to merge and clean.

Comma Separated Values Data (.csv): CSV files are plain text files that use commas as separators to store tabular data. Each line in the file represents a row, and the values within each line are separated by commas. For an example on using CSV files in this project, review the data cleaning section in our repository. To merge and clean CSV data, follow these steps:

1. **Open** the new CSV file
2. **Inspect** metadata, dimensions, and values: Examine the data thoroughly including its metadata, dimensionality, and the actual values it holds.
3. **Convert** the units if needed: Standardize the measurement units across your dataset to ensure reusability with our data
4. **Summarize** the data by the desired time period (e.g., quarterly): Depending on your analysis, you may need to aggregate or resample the data.
5. **Check** the data for any missing or inconsistent values: Always validate the data, check for any missing or bizarre values, including verification of lat and lon columns with the original mask values.
6. **Merge** the new data to your existing dataset: Combine the new data with your existing dataset, aligning by year, quarter, lat, lon, as applicable.
7. **Update** any relevant metadata or documentation to reflect the new data: Keep your metadata and documentation up-to-date, reflecting any new data that has been added.

Raster Data (.tif) : Raster data refers to spatial data represented in a grid-like structure, where each cell or pixel contains a value representing a specific column, which are called attributes. (e.g., sst, kelp area, depth, etc). For an example on using raster files in this project, review the data cleaning section in our repository. To work with raster data files, follow these steps:

1. **Load** the raster data using a suitable package or library (e.g., raster, terra, stars).
2. **Inspect** metadata, dimensions, and values to understand the data.
3. **Convert** units if necessary to match your analysis or visualization needs.
4. **Reproject** coordinate reference system to match other data (e.g., WGS 84)
5. **Mask** the data to focus on a specific area of interest.
6. **Summarize** the data over desired time periods using statistical functions or resampling methods.
7. **Check** for missing or inconsistent values and handle them appropriately. Stack new data onto an existing dataset for your analysis.
8. **Update** metadata and documentation to reflect any changes made.

NetCDF Data (.nc): These are special files often used in scientific and environmental research. They excel at storing different pieces of information all in one place. They're not just for pictures; they can handle many types of attributes and variable data. If you want to work with these files, it's crucial to understand the metadata, which provides essential details about your dataset. As mentioned, NetCDF files store multidimensional data, making them especially useful in the scientific and environmental research fields. Some steps we recommend to have a good approach with NetCDF files are:

- **Review the Metadata:** Make sure to understand the information provided about variables, dimensions, and attributes within the NetCDF file.
- **Identify the Desired Data:** Decide on which variables or data fields you are interested in.

- **Download:** Once you've learned how to download and work with NetCDF files for a specific dataset, continue with that process. Make sure you are correctly reading and collecting the data you need.

For a practical demonstration, refer to the kelpGeoMod project code available on GitHub. This repository contains two Rmd files (sst and the kelp area and biomass script), that show how to read and interpret data from a NetCDF file. Looking at these .Rmd files will help you understand how we accessed the data and what kind of values we extracted.

Shapefile (.shp): Shapefiles are vector data used in geographic information system (GIS) software. It is important to understand that when working with an unknown Shapefile, it can consist of data points, lines, polygons, or multipolygons. For an example on using raster files in this project, review the mask Rmd in our repository. To work with Shapefiles, follow these steps:

1. **Open** the GIS or coding software and import the Shapefile into the environment.
2. **Inspect** the imported shapefile data to ensure its structure contains the required fields in the correct format.
3. **Reproject** the shapefile to align it with the same Coordinate Reference System (CRS) as your existing data, ensuring proper geographic alignment between the datasets.
4. **Clip/Crop** the Shapefile data to include only the specific area of interest, removing any unnecessary data outside of that region.
5. **Check** the shapefile data for any missing or inconsistent values, and address these issues appropriately.
6. **Merge** the new shapefile data with your existing dataset, typically by performing a join based on a shared identifier such as a unique feature ID.
7. **Update** the metadata or documentation related to your dataset to reflect the inclusion of the new Shapefile.

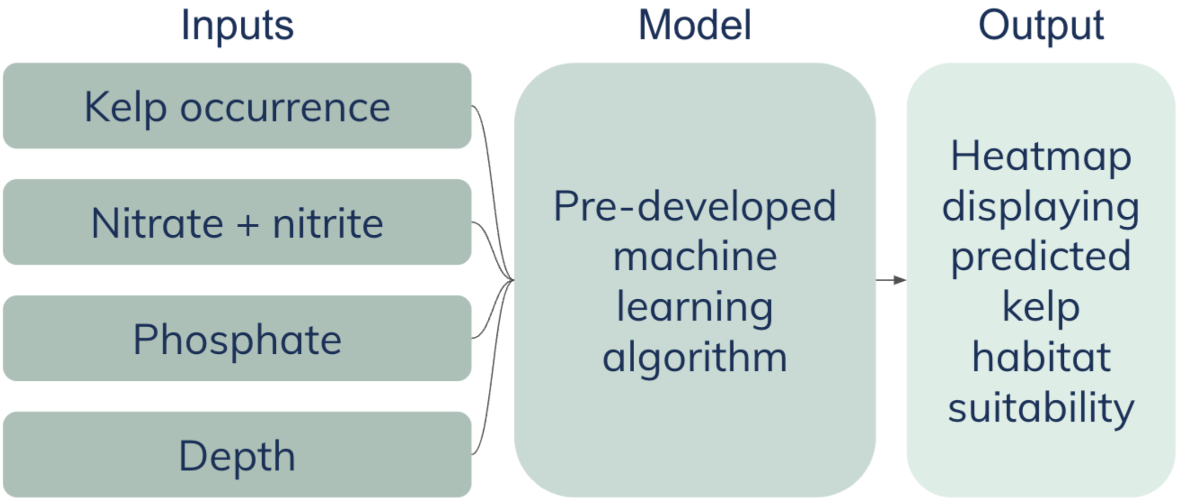
We have focused on the main three file types used in the project, but additional considerations may apply to your specific data. In addition to the provided guide and tips, we highly recommend doing some additional research around your new data so you feel comfortable with your new file structure. To do so, we recommend conducting peer reviews and performing sanity checks.

Once your data has been successfully integrated with the existing project data, we will outline the steps taken by the kelpGeoMod team to determine kelp habitat suitability in the area of interest. For this project, the Maxent model was chosen and the following steps provide a general overview of the Maxent modeling process and how the model utilizes inputs to generate outputs. If you are interested in modeling, please follow the instructions below to proceed.

Feeding the Data to Maxent

Define your model

Inputs: Kelp presence with geographic information, depth, nitrate + nitrite spatial distribution, & phosphate spatial distribution (all averaged per quarter).
Output: Spatial distribution of predicted probability of habitat suitability.


Inputs
Model
Output
Kelp occurrence
Nitrate + nitrite
Phosphate
Depth
Pre-developed machine learning algorithm
Heatmap displaying predicted kelp habitat suitability

Wallace Workflow:

1. Ensure that all data is within the same area of interest and has identical specifications for position, extent, CRS, and origin.
2. Utilize script 02-prep-kelp-presence.Rmd to divide the **observed species data** into a CSV dataset formatted for Wallace. This script aggregates kelp averages kelp area over all years by quarter. Format your column names as:
    a. "scientific_name" (in binomial nomenclature with the genus capitalized).
    b. "longitude"
    c. "latitude"
3. Utilize script 03-arrange-quarterly-data.rmd to divide the environmental factors into raster layers by quarter and reorganize into folders for each quarter containing all necessary data..
    a. Name each layer with a standardized naming convention. Eg. “sst”, “depth”, etc.

4. Run Wallace from scratch using the script run_maxent.Rmd or replicate our maxent runs with the following scripts…
    a. Will add these later
    b. See the Wallace vignette for instruction on how to run maxent through Wallace.