

Climate risk assessment pipeline

Filtering occurrence data and extracting climate conditions -> Data collation -> multivariate Climate Risk Score calculation -> Δ mCRS calculation -> Analyses

1. Filtering occurrence data

These steps are contained in the `1_CoordinateFiltering` and `1_Thinning_Collating_Points` R scripts. Occurrence data were obtained from [OBIS](#) and [GBIF](#). Filtering occurrence points involves loading raw occurrence points and filtering them to only include the relevant genetic populations studied (North-East Pacific, North-West Pacific, South Pacific and South Africa). Using the 'sp' package the coordinates are converted to a `SpatialPoints` object and the points that occur over land (Natural Earth 10m Projection) are filtered out. Due to the grid size of the CMIP6 data some coordinates occur too close to land to extract data from the raster file. These occurrences need to be moved to the closest valid data extraction point on the coastline. This is done by extending these points to the closest point within the valid raster data region using the 'sf' package and a line extension function. Coordinates are then thinned to reduce the clustering that may occur when multiple species data collection types (such as surveys and fisheries) are collated into the dataset. The 'spThin' package `thin()` function was used with a 100km thinning distance to match the spatial resolution of environmental data. This resulted in a final dataset of 180 occurrence points that can then be used for data extraction.

2. Data Extraction

The data extraction step, in the form of `2_DataManipulation` R scripts and the `env.extract.bulk.ecoregion` function from `env_extract_ecoregionSummary.R` is then used to extract climate values from the processed .tiff files at the coordinates previously specified. This extracts data from the .tiff files using the 'raster' package `extract()` function. This step also has time inputs to specify the time range to be extracted (1850-2014 for historical and 2015-2100 for ssp). Extraction is looped to extract nitrate, pH, salinity and temperature data for a time variant/model and combine them into a `time.model.stack` summary dataframe for later use.

3. Data collation

This step involves combining the summary dataframes for each model into a multi-model ensemble for later use. Each model summary dataframe is loaded into R and the depth levels are categorised into bins to allow for consistent averaging between models. Depth bins of 0-10m, 10-50m, 50-100m and 100-150m are used. The population and region identifiers are also joined onto the data. This results in a large dataframe containing all historical conditions for the coordinates, for all models (This is converted to a data table to allow for more efficient manipulation using the 'data.table' package). As the models may contain different numbers of values in each depth bin, climate variable values across a depth bin were averaged. The average for each variable across the different models was then taken to account for variation between model sources. Date was then added to combine Year and Month values, and the datatable was converted into a wide format for downstream CRS analyses.

Exploratory analyses of the historical conditions (`HistoricalConditions_Exploration.R`) found that the New Zealand region included a much larger number of fish observation

coordinates than Australia. These observations were lower in min, mean and max temperatures experienced and influenced the conditions included in the East population. As a result, the New Zealand observations were separated from the Australian observations to avoid impacting the CRS calculations.

4. Climate Risk Score calculation

Climate risk score calculation involves transforming historical and projection data using principal component analysis to analyse the distances of a coordinate to optimum and hazardous conditions. These processes are performed on one population, depth level and ssp scenario at a time. First two principal components, that explain >71% of the variation in the conditions are calculated for the historical and ssp environmental conditions. `chull()` is used to create a convex hull around all of the historical condition PC values. This hull is then converted to a `SpatialPolygon` object for later analysis. The centroid points are the mean of the two PC values for the reference period (1850-2000). To calculate the distance from the centroid to the hull boundary via an occurrence point a line is created from the centroid to the occurrence point. This line is then extended linearly to a point that lies outside of the convex hull (using information from <https://gis.stackexchange.com/questions/154479/how-to-find-the-distance-from-a-point-to-a-polygon-via-another-point-in-r>). The intersection point can then be found and is called an edge point. This dataset with occurrence points, corresponding edge points and the centroid can then be used to calculate CRS scores. DM is the distance from the occurrence point to the edge of the hull, DC is the distance from the occurrence point to the centroid and DE is the distance from the centroid to the edge point via the occurrence point. These are calculated using the `dist()` function using `method = "euclidean"`. $mCRS$ values are then calculated as the $-\log(DE/DC)$. (DM values are only calculated to validate that $DM + DC = DE$). The result of this process is a results dataset containing $mCRS$ values for each Date and ID. $mCRS$ scores show a date/ID's risk relative to the entire population's niche at that depth level and are referred to as regional risk. Risk score calculation was also performed at the species level for comparisons with population level calculations (found in scripts containing 'Spl', while population level scripts contain 'ensemble').

5. $\Delta mCRS$ calculation

$\Delta mCRS$ values are calculated in the `5_DeltaCRS` scripts. The first step is calculating the reference period (1850-2000) mean $mCRS$ values for each depth level for an occurrence location ID. Then that mean is subtracted from each date's $mCRS$ value for a particular location. This produces a $\Delta mCRS$ score which represents a date/ID's risk relative to historical levels of climate risk experienced at that ID location. This is referred to as local risk. Location specific $mCRS$ and $\Delta mCRS$ results are saved as a csv for later use. Scripts containing 'Spl' refer to species level calculations, while 'ensemble' scripts refer to population level calculations.

6. Statistical analyses and figures

Statistical analyses and figure construction are contained in the `Analyses` R script. Analyses are performed separately for each SSP scenario. For analyses $mCRS$ and $\Delta mCRS$ values were averaged annually. For regional risk emergence analyses the year when $mCRS$ values became positive are calculated for each location ID at each depth layer (novel year). These are then analysed in a linear regression that incorporates Population and Depth as factor explanatory variables against novel year ($\text{Novel Year} \sim \text{Depth} * \text{PopID}$).

The differences between depth and population levels is then analysed using emmeans contrasts and bonferroni adjustment method. For rate of change in regional risk analyses annually averaged mCRS values are incorporated into a linear regression with Year, Depth and Population as explanatory variables ($mCRS \sim Year * Depth * PopID$). Temporal autocorrelation is incorporated into this model using the corA1(form = ~Year|PopID/Depth/Extract.ID) and gls() functions. This incorporates a separate temporal correlation structure for each ID location within each depth and population. emtrends() is then used to examine the trends in mCRS across Year at each population and depth level. For local risk magnitude analyses $\Delta mCRS$ values were used at the time when a warming level was reached for the population/depth level. Warming levels were calculated as the modelled (gam model of ΔT and Year) temperature change relative to the mean reference (1850-2000) temperature for a certain population/depth level. Each population and depth reaches the warming level at a different time, thus a warming level is only used if it is reached by 2100 in all population and depth levels for a scenario. 0.75°C is the warming level used for SSP1-2.6 1°C is the warming level used for SSP2-4.5 and 1.75°C is the warming level used for SSP5-8.5. This method standardises $\Delta mCRS$ values for each population and depth level at a similar magnitude of warming across these levels. Annual $\Delta mCRS$ ID values were then analysed in a linear regression against longitude, Distance from the equator (EqDist) , depth and population ($\Delta mCRS \sim Long + EqDist * Depth * PopID$), incorporating spatial autocorrelation with corExp(). Longitudinal values were adjusted by a random value (from 0.000000000001° to 0.000000002448°) to allow for spatial autocorrelation to be applied across depth layers, as this process cannot include duplicate coordinate values. Differences across depths and population levels were analysed using emmeans contrasts and bonferroni adjustment method. Latitudinal effects were analysed using emtrends to look for trends in $\Delta mCRS$ across distance from the equator for each population and depth level. Figures were created using sf, ggplot2 and ggeffects packages. Analyses are organised into sections including: A) Loading $\Delta mCRS$ results B) Loading environmental climate data C) Calculating population and depth specific warming levels and years when levels are reached D) Regional risk emergence calculation E) Statistical Analyses F) Figures