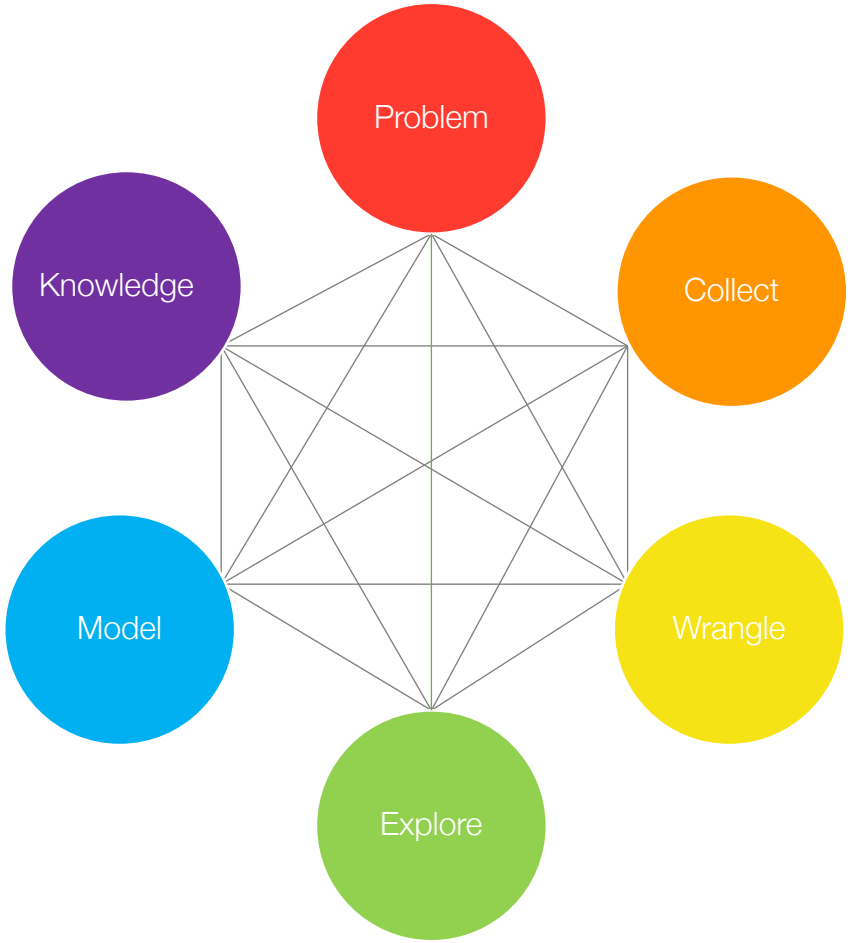
GEOG0114: PRINCIPLES OF SPATIAL ANALYSIS

WEEK 5: SUITABILITY MAPPING (PART 2)

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Contents

1. To understand what are niche models and the concept of niche?
2. To understand the theoretical framework behind niche modelling
3. Knowing the modelling process of niche models with specific focus on MAXENT and LOGIT-based regression models
4. Using the prediction of wildfires in California as a motivating example



QUICK RECAP

Suitability Mapping

Week 4

Knowledge-based
(Mixed methods)

Week 5

Data-driven approach
(Quantitative)

Background: What are Niche Models?

Definition of Niche Modelling [1]

Niche models are a class of methods that use **occurrence data (i.e., an outcome)** in conjunction with **environmental data (i.e., predictor variables)** to make a correlative model of the environmental conditions (or quality and state of the environment) that is tenable for such outcome.

In other words, these are methods that can spatially predict and characterise the relative suitability of areas for where an outcome thrives

The key focus are as follows:

1) To geospatially estimate the potential suitability of an area known for an outcome given a set of environmental data
2) To estimate the geographical limits (or carve out a predicted boundary) for which an outcome can potentially occur (i.e., niches, or suitable area etc.).
3) Useful making forecasts of potential with future scenarios.

Note: In statistical jargon: "Niche modelling" is often referred "Distributional modelling"

Definition of Niche Modelling [2]

Niche models are a class of methods that use **occurrence data (i.e., an outcome)** in conjunction with **environmental data (i.e., predictor variables)** to make a correlative model of the environmental conditions (or quality and state of the environment) that is tenable for such outcome.

At its core, niche models algorithmically identifies associations between environmental variables and known occurrence of a particular outcome – so as to define conditions within which such outcomes are maintained in geographic space (e.g., animal species, disaster events, crimes etc.)

The use of computer algorithms to:

First fit: Mathematical function(s), describing the distribution of an outcome of interest in geographic space for a number for environmental variables

Then: Generate predictive maps of an outcome showing its '**niches**' in a form of probability distribution in geographic space using these mathematical function(s)

What is the meaning of the word “Niche”?

The ‘**Niche**’ in modern science describes an area/region where **ALL** environmental factors influence the fitness of a particular outcome.

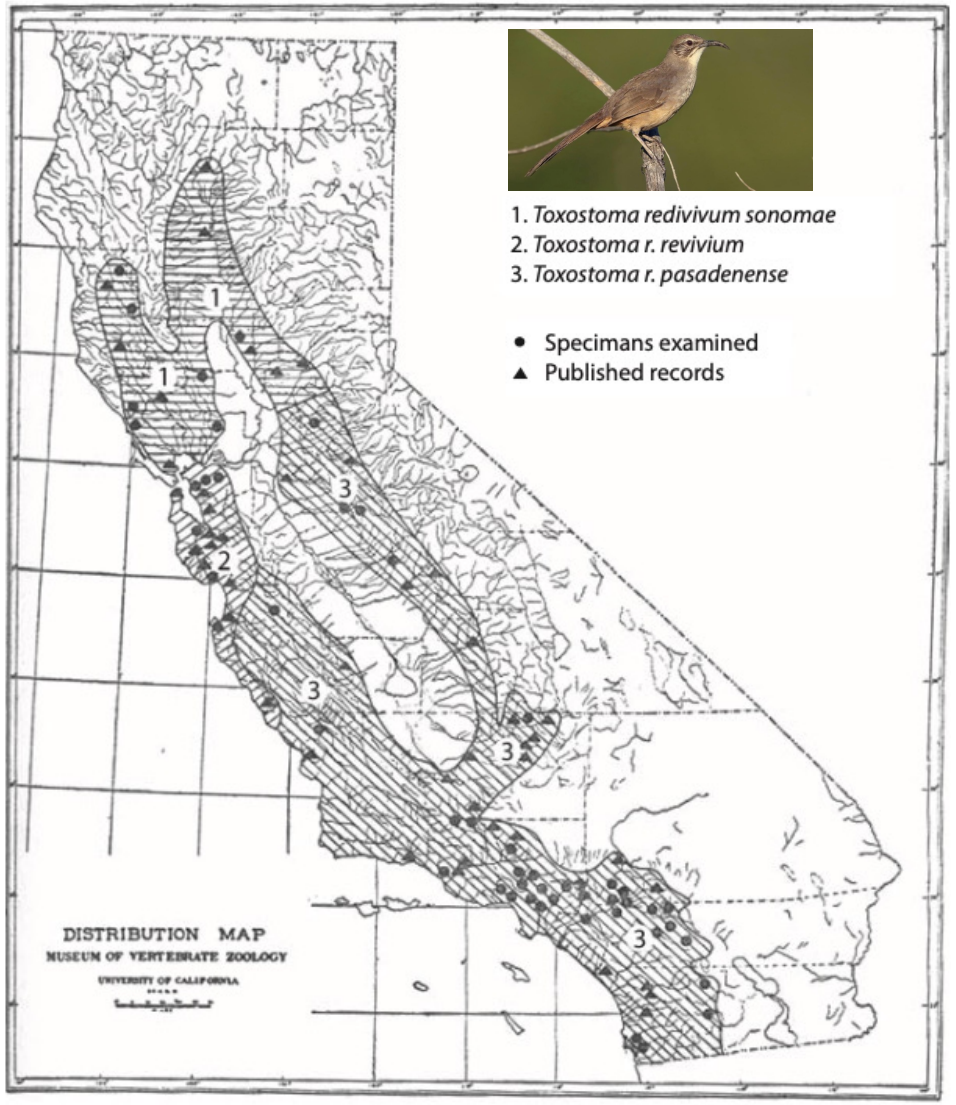
The words ‘**environmental factors**’ has many dimensions in its meaning – this can include physical (e.g., geology, topography, climate etc.), biological, political and/or social environment (e.g., population density, demographic make-up, etc.,)

History of the Niche concept

Joseph Grinnell was the first to offer a comprehensive input to the ideas of niches. From 1904 to 1924, he investigated how variations in environmental conditions were linked to species’ distribution, with special interest on vertebrates. He believed that:

“**the existence […] of a species is vitally bound with environments**” (Grinnell, 1924, p. 226)

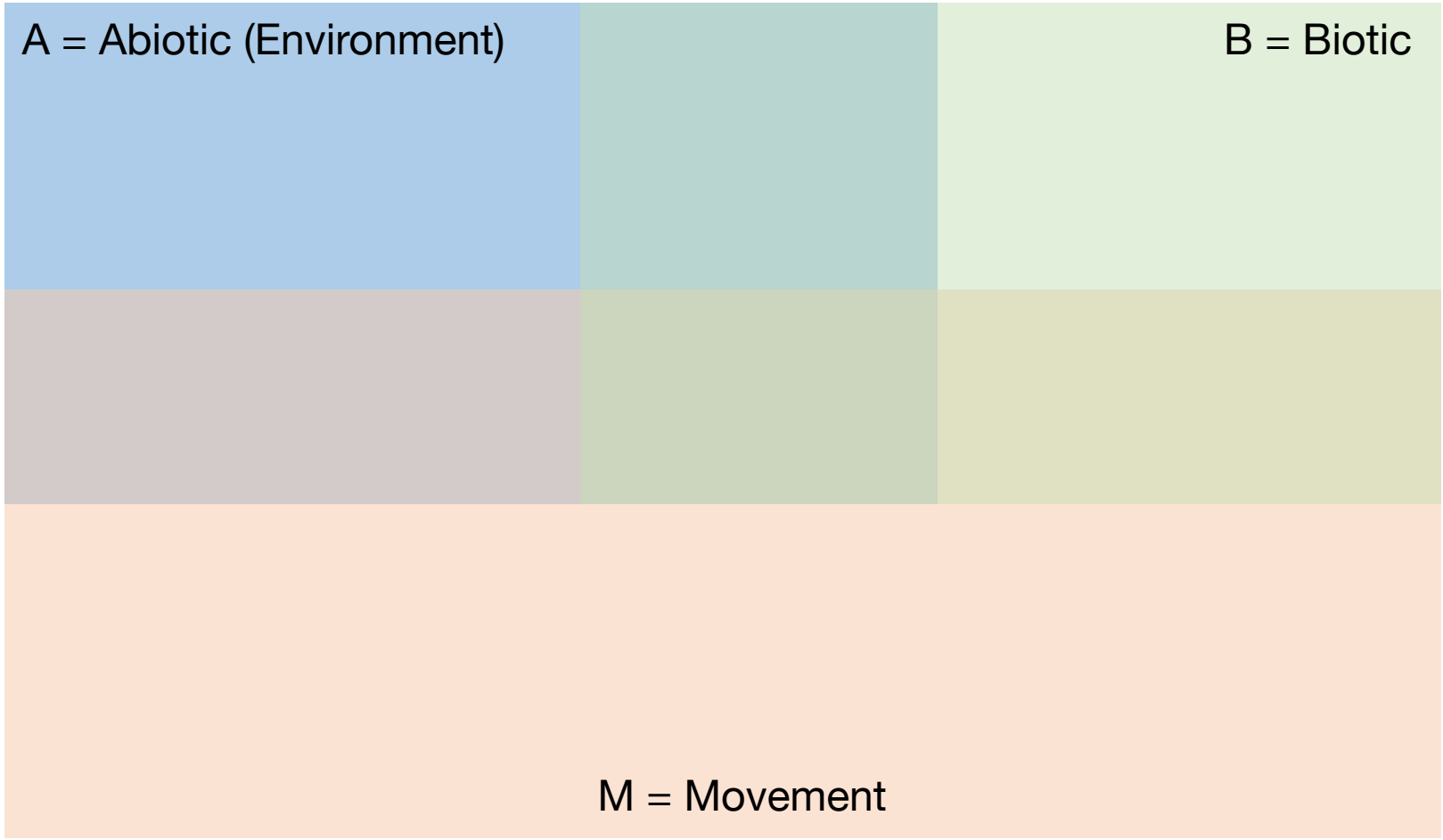
He defined the **niche** as all factors determining the species’ existence in a given location (i.e., geography) and its interplay with abiotic (i.e., physical environment [e.g., temperature, rainfall etc.,]) and biotic (e.g., food supply, nest sites etc.,) factors as well as movement within such space.



Distribution of 3 different species of the California Thrasher in California. Dots are occurrences of the species, or triangles are published records of sightings. Shade are is the distributional limits of such species. **Source: Joseph Grinnell (1917)**

B-A-M Diagram: Theoretical Framework behind Niche Modelling

G = Geographical space



G = represent the reference of study area

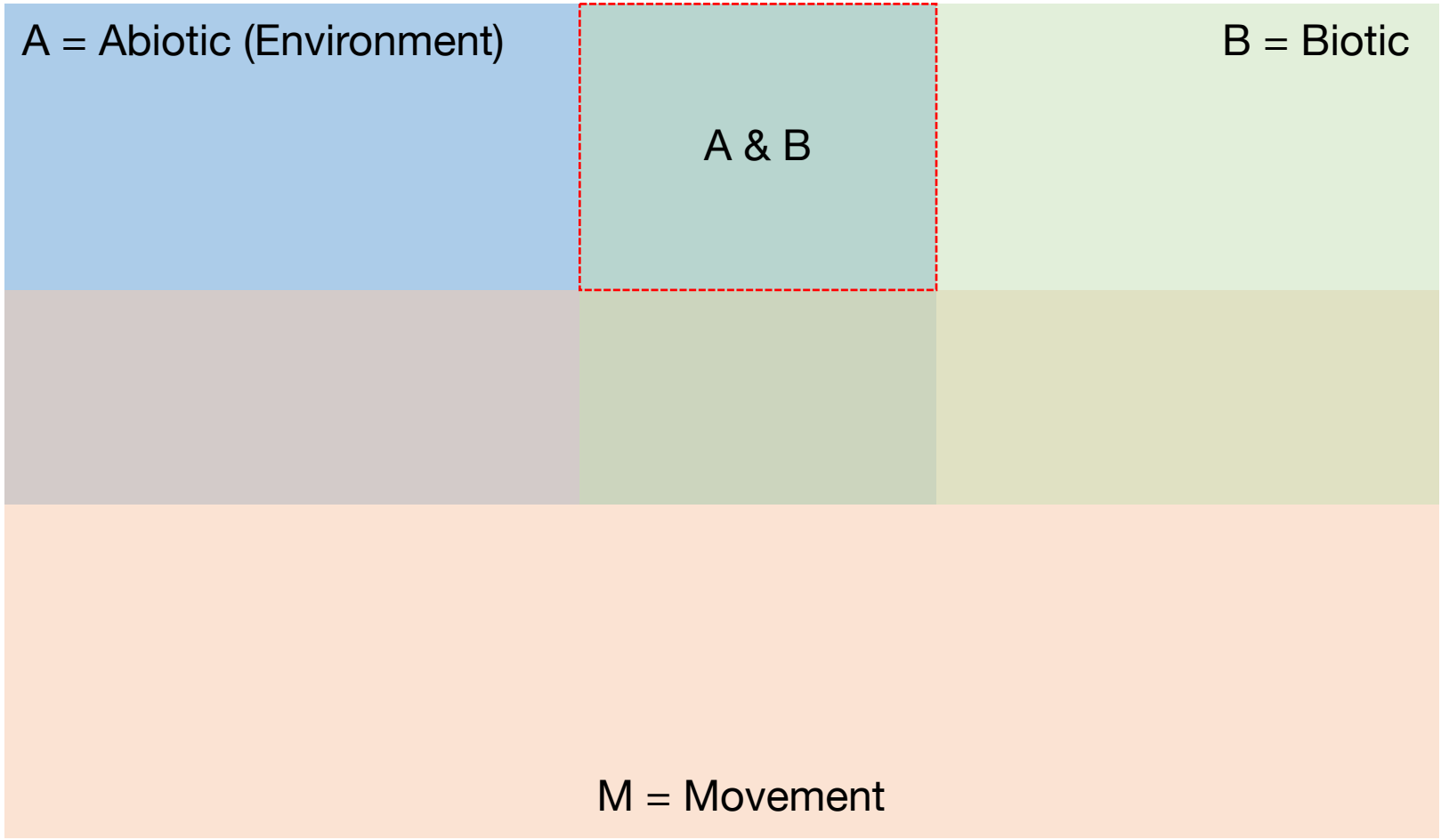
B = represent the variation in biologic environmental conditions (resource suitable for species)

A = represent the variation in physical environmental conditions

M = Movement potential within study area

B-A-M Diagram: Theoretical Framework behind Niche Modelling

G = Geographical space



G = represent the reference of study area

B = represent the variation in biologic environmental conditions (resource suitable for species)

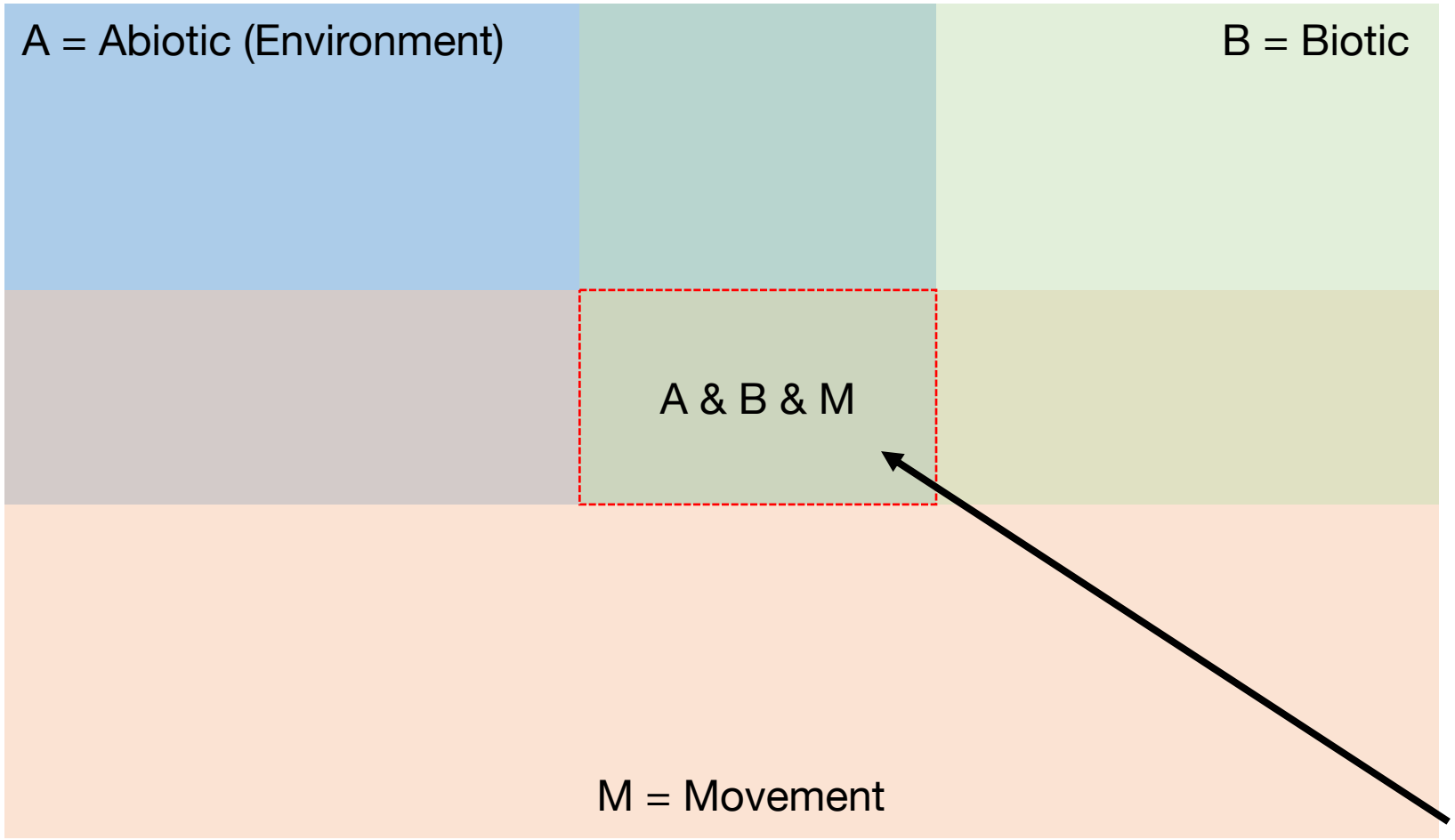
A = represent the variation in physical environmental conditions

M = Movement potential within study area

A & B = This is the interplay between A & B, where they intersect – these are the regions considered to be favourable for species in terms of the conditions from A and B

B-A-M Diagram: Theoretical Framework behind Niche Modelling

G = Geographical space



G = represent the reference of study area

B = represent the variation in biologic environmental conditions (resource suitable for species)

A = represent the variation in physical environmental conditions

M = Movement potential within study area

A & B & M = This is the interplay between A & B & M, where they intersect – these 3 regions considered to be most optimal for species to live-in in terms of the conditions from A and B and M

This is what we want to estimate in our models


Zoology


Food security

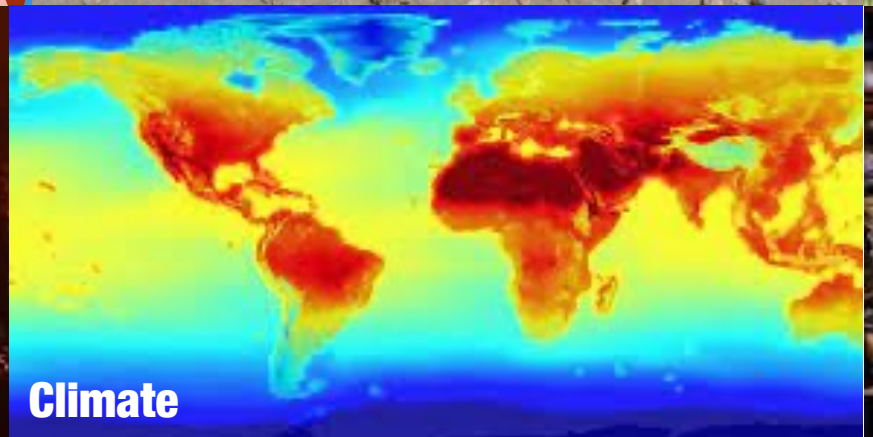

Environmental Criminology


Environmental & Spatial Epidemiology
• Vector-borne disease


Palaeontogy and Archology


Landscape ecology


Natural Disaster Science


Climate


Humanitarian crisis


Zoology


Food security

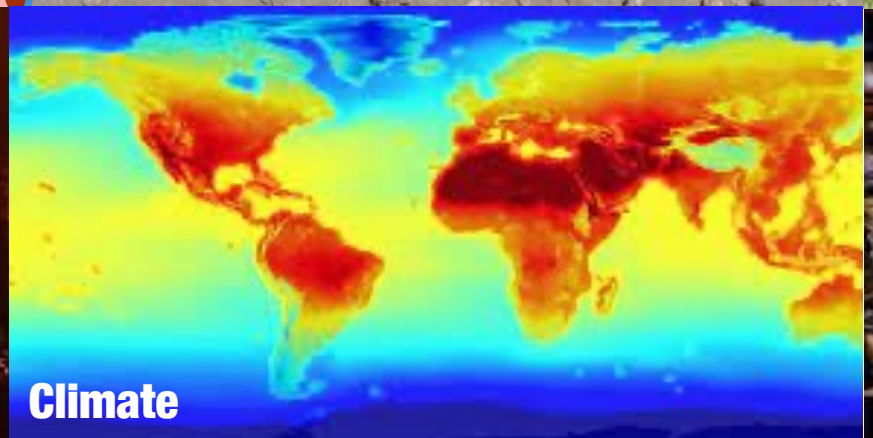

Environmental Criminology


Environmental & Spatial Epidemiology
• Vector-borne disease


Palaeontogy and Archology


Landscape ecology


Natural Disaster Science

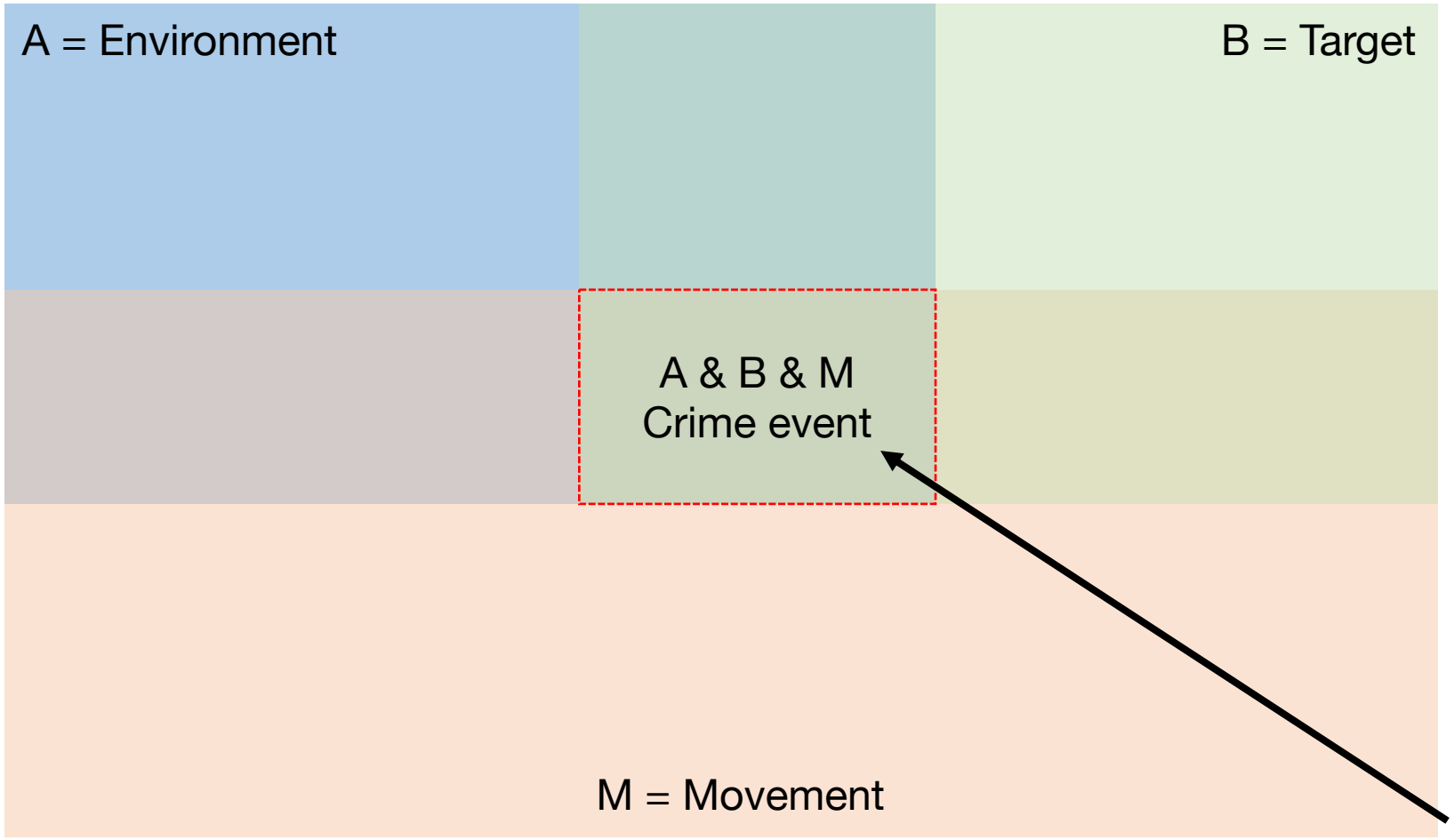

Climate


Humanitarian crisis

B-A-M Diagram: Crime Pattern Theory

Example of application to quantitative criminology

G = Activity space for potential offender



This is what we want to estimate in our models

G = represent the reference of study area

B = Target refers to the victim (person, house, shop etc.)

A = represent the variation in physical environmental conditions – environmental factors that attract a potential offender (security, social disorganisation etc.)

M = Movement potential within study area (street network; accessibility etc.)

A & B & M = This is the interplay between A & B & M, where they intersect – these 3 regions considered to be most optimal for crimes to occur in terms of when conditions from A and B and M converge

What is the methodology behind niche models?

Modelling process [1]

Things to consider:

1. Definition of the research question

2. Data identification and preparation (i.e., environmental layers, presence only vs presence & absence data (from surveys), or random generation of background data (to act as “pseudo absences”)

3. Selection of a modelling algorithm

4. Statistical inference and testing predictive performance (i.e., examine of variable contribution, response curves for each predictor variable and Area Under the Curve (AUCs))

5. Interpretation of model outputs (i.e., predicted probabilities and suitability areas/niches)

Modelling process [2]

Definition of the research question

The research question, in whatever context or domain of science it is in. You must be able to link whatever scientific theory (or plausible mechanism) that explains the outcome of interest with the BAM diagram.

Ask yourself the following questions:

- What is the outcome of interest?
- What are the environmental (i.e., abiotic and biotic) factors that influence the occurrence of an outcome?
- Is there a scientific explanation or plausible mechanism?

This is how you justify the use of niche models to addressing research question.

Modelling process [2]

Data identification and preparation

- **Environmental variables**: these are a series of **gridded raster datasets**

- **Outcome of interest**: This is **spatial point data only**; it should represent the presence of an outcome. There are three types of formats – which they can be presented.

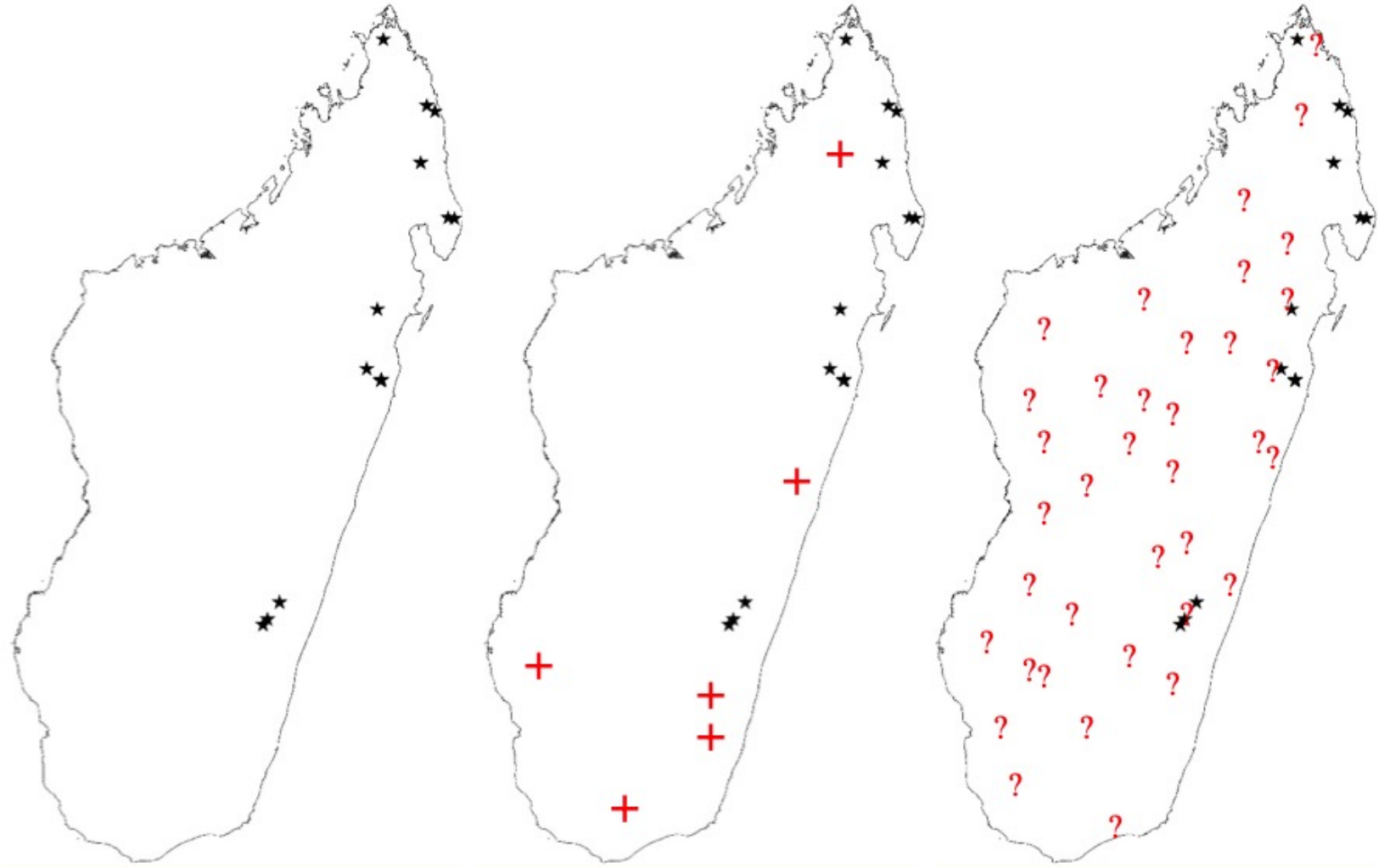
i. **Presence-only**: This refers to point sample locations with information on known presence of a particular outcome. But at the same time, does not contain point sample location of known absences of that outcome.

ii. **Presence and absence**: This refers to point sample locations with information on known presence and absence of a particular outcome.

iii. **Presence and pseudo-absence**: This refers to point sample locations with information on known presence of a particular outcome only. But the absence points were generated at random spatially to be used as controls.

Example: Malaria prevalence survey in Madagascar

Research question: Where will Malaria transmission occur?



Presence only: Surveys of communities with at least one person infected with Malaria

Presence & absence: Surveys of communities where we have at least one infected person with Malaria at a location; a community with people disease free.

Presence and pseudo-absence: We only surveys of communities with at least one person infected with Malaria. No absence data, and hence we generate them randomly to act as proxies for absences (background)

Modelling process [3]

Selection of a modelling algorithm

Niche Models	Data type
Climatic envelope Gower Metric	Presence-only
Maximum entropy (MAXENT)	Presence and pseudo-absence (background)
Regression-based models: e.g., Generalised linear model (GLM) and Generalised additive model (GAM)(non-linear)	Presence and absence Presence and pseudo-absence (background)
Machine learning models: Artificial Neural Network (ANN); Classification and regression-trees for GLM, GAM and ANNs	Presence and absence

There are a tonne of other models out there which I have not listed because I just don't know what they do…

Modelling process [3]

Selection of a modelling algorithm

Niche Models	Data type
Climatic envelope Gower Metric	Presence-only
Maximum entropy (MAXENT)	Presence and pseudo-absence (background)
Regression-based models: e.g., Generalised linear model (GLM) and Generalised additive model (GAM)(non-linear)	Presence and absence Presence and pseudo-absence (background)
Machine learning models: Artificial Neural Network (ANN); Classification and regression-trees for GLM, GAM and ANNs	Presence and absence

There are a tonne of other models out there which I have not listed because I just don't know what they do…

Modelling process [3]

Selection of a modelling algorithm

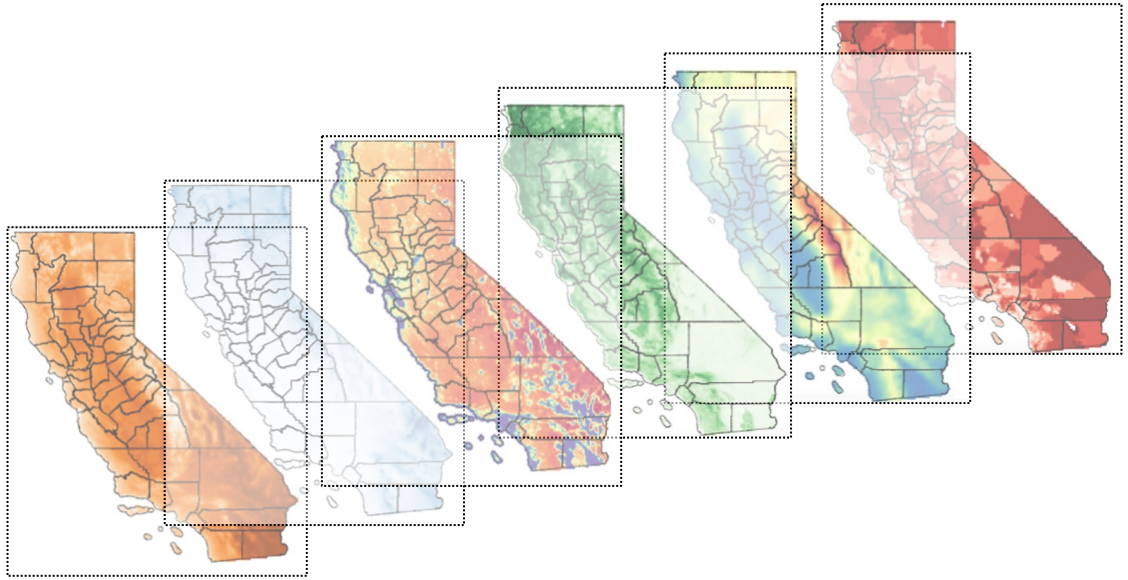
Niche Models	Data type
Maximum entropy (MAXENT)	Presence and pseudo-absence (background)
Regression-based models: e.g., Generalised linear model (GLM) and Generalised additive model (GAM)(non-linear)	Presence and absence

Occurrence point data

Location	Longitude	Latitude	Outcome
Site 1	x_1	y_1	1
Site 2	x_2	y_2	1
Site 3	x_3	y_3	0
Site 4	x_4	y_4	0
⋮	⋮	⋮	⋮
Site n	x_n	y_n	1

Raster-to-point extraction

Environmental data (multi-band raster)



Modelling process [3]

Selection of a modelling algorithm

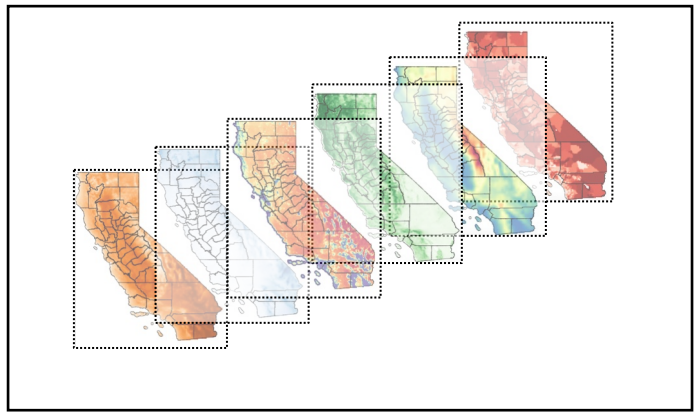
Niche Models	Data type
Maximum entropy (MAXENT)	Presence and pseudo-absence (background)
Regression-based models: e.g., Generalised linear model (GLM) and Generalised additive model (GAM)(non-linear)	Presence and absence

Linked occurrence point data with environmental data

Location	Longitude	Latitude	Outcome
Site 1	x_1	y_1	1
Site 2	x_2	y_2	1
Site 3	x_3	y_3	0
Site 4	x_4	y_4	0
⋮	⋮	⋮	⋮
Site n	x_n	y_n	1

Temp	NDVI	Dryness	⋯
30	128	457	⋯
27	291	239	⋯
34	302	124	⋯
29	305	54	⋯
⋮	⋮	⋮	⋮
25	281	345	⋯

Extracted environmental values from Rasters



Modelling process [4]

Statistical inference and testing predictive performance

Linked occurrence point data with environmental data

Location	Longitude	Latitude	Outcome
Site 1	x_1	y_1	1
Site 2	x_2	y_2	1
Site 3	x_3	y_3	0
Site 4	x_4	y_4	0
⋮	⋮	⋮	⋮
Site n	x_n	y_n	1

Temp	NDVI	Dryness	⋯
30	128	457	⋯
27	291	239	⋯
34	302	124	⋯
29	305	54	⋯
⋮	⋮	⋮	⋮
25	281	345	⋯

We use the linked point data and split this according for **k-fold cross-validation**:

- Testing (20%) | Training (80%) (5 equal portion (1 selected for testing))
- Testing (25%) | Training (75%) (4 equal portion (1 selected for testing))

Training data: Is the large portion of records we use for the actual prediction – this creates our trained model

Testing data: Is the small portion of records we use to assess the predictive performance of our trained model.

Evaluation of testing data, we are interested in two outputs which measures for model accuracy:

- **Area Under a Curve (threshold for good model: 0.5)**
- **Optimized estimate or thresholds for where predictions are a true positive and true negative [max TPR+TNR]**

If the overall AUC > 0.5, proceed to use the trained model on multi-band raster generate probability estimates across region >> Use optimized estimate threshold from [max TPR+TNR] to map the extent for suitability for the outcome.

Example: Predicting the extent of wildfires in California

Taylor, L. (2022): The social side of fires: assessing the inclusion of human social factors in fire prediction models

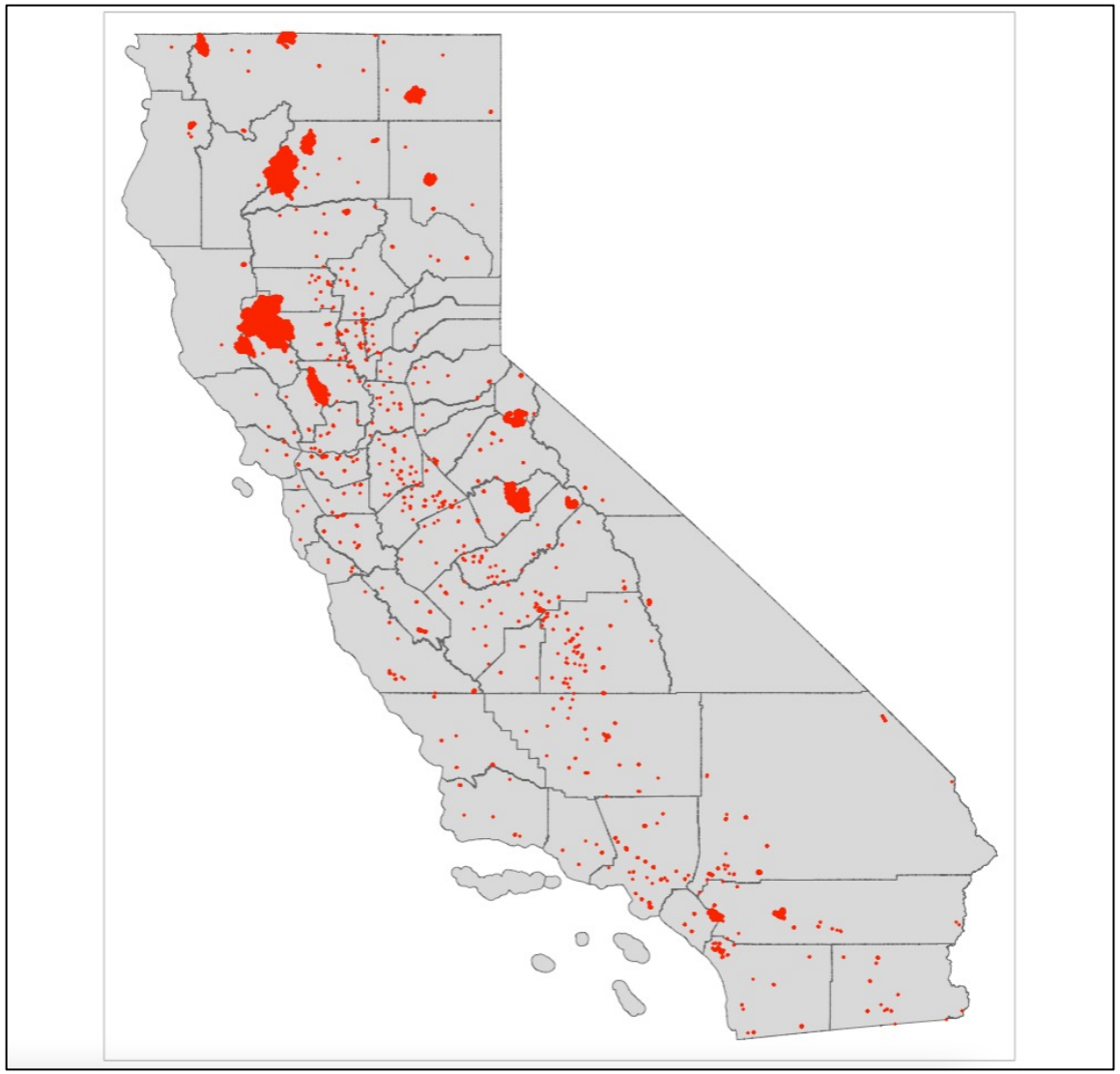


Figure shows the study area of California (counties) and points are occurrence of wildfires during the summer period of 2018. Presence-only points.

The objectives are to determine the occurrence of wildfires, as well as infer the extent (or zones) for such environmental hazard in California given a set of predictor variables (i.e., climate, vegetation, anthropogenic and socioeconomic risk factors which are raster)

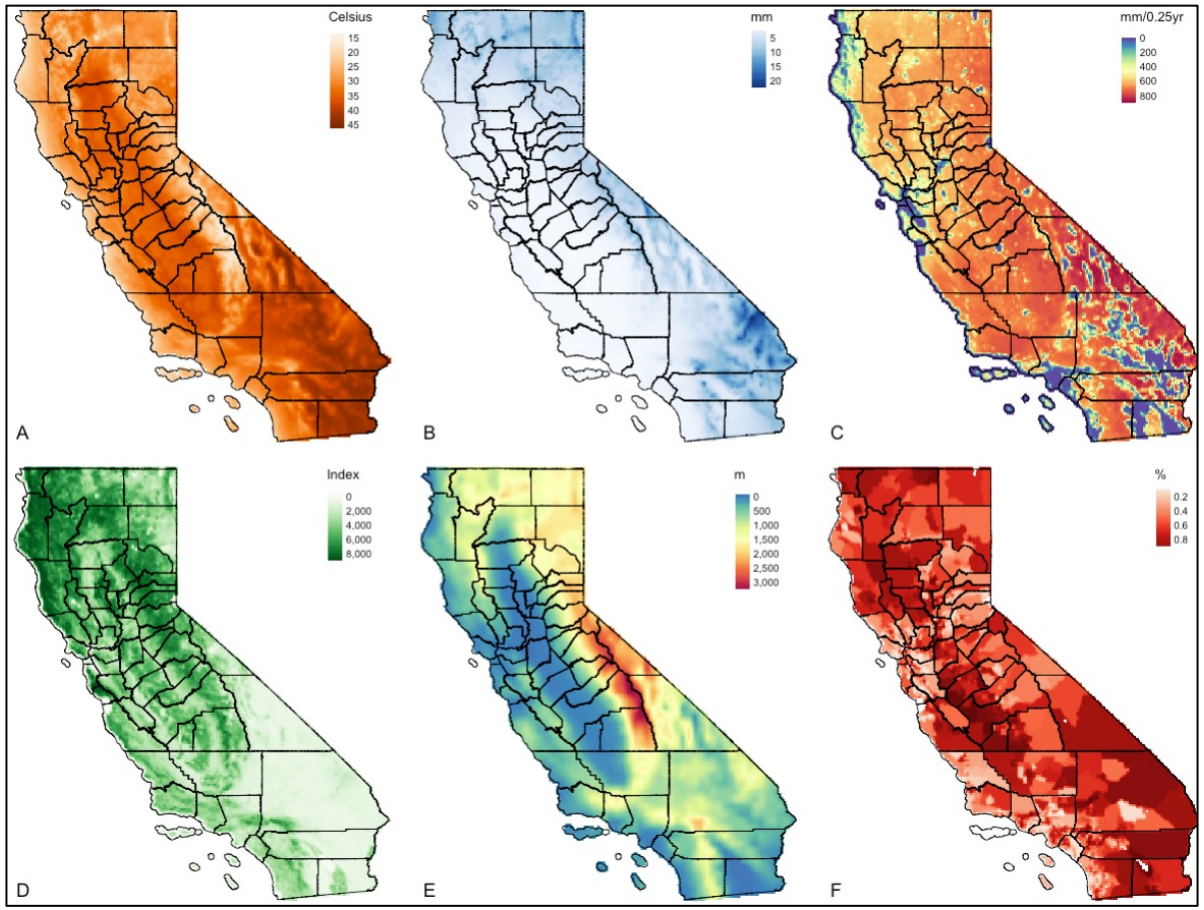
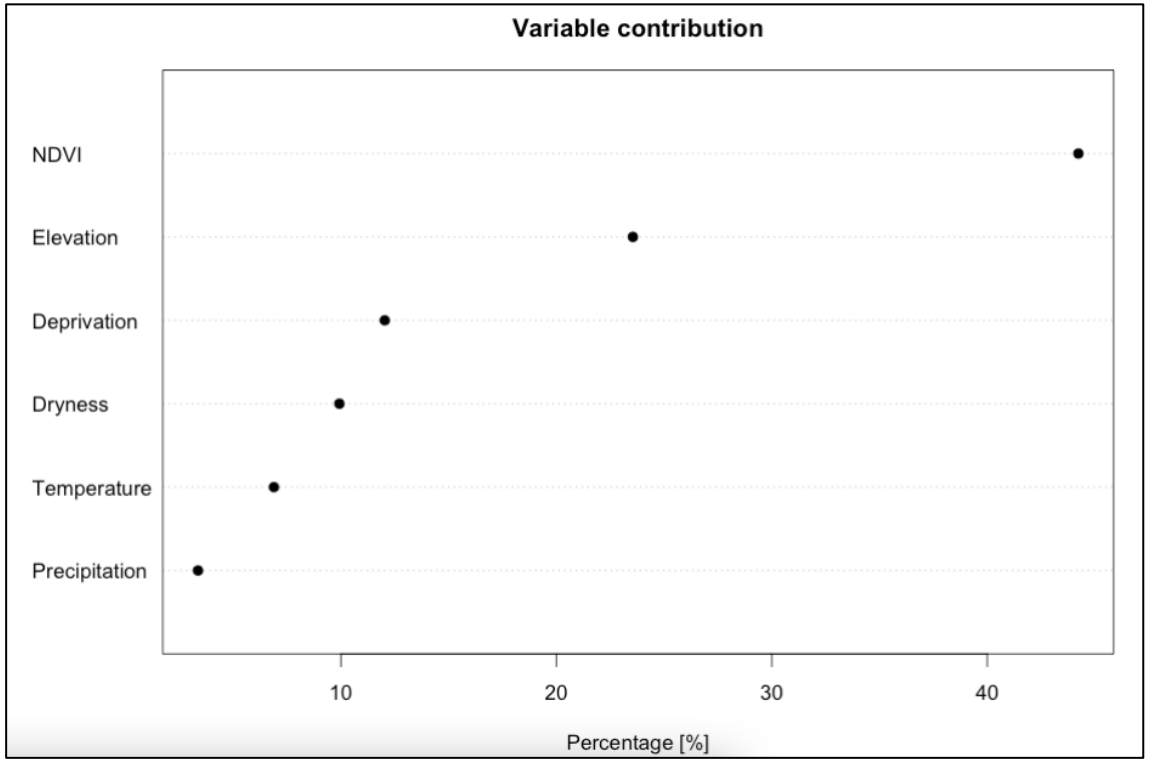


Figure panel shows A: Temperature (degree Celsius); B: Precipitation (mm); C: Dryness (Evapotranspiration) (mm/0.25 year); D: Vegetation (NDVI); E: Elevation (meters [m]); & F: Socioeconomic vulnerability index (%)

Using MAXENT

Result 1: Variable Contribution & Response curves



Interpretation: Here, we can see the following contribution estimates: NDVI (44.2321%); Elevation (23.5530%); Deprivation (12.0339%); Dryness (9.9266); Temperature (6.8892%); and Precipitation (3.3653%). The contribution estimates should sum up to 100%. From this plot, we can see that the model is most sensitive to variation in NDVI, followed with additional contributions from land surface elevation, and from increased levels of socioeconomic deprivation (reporting top three).

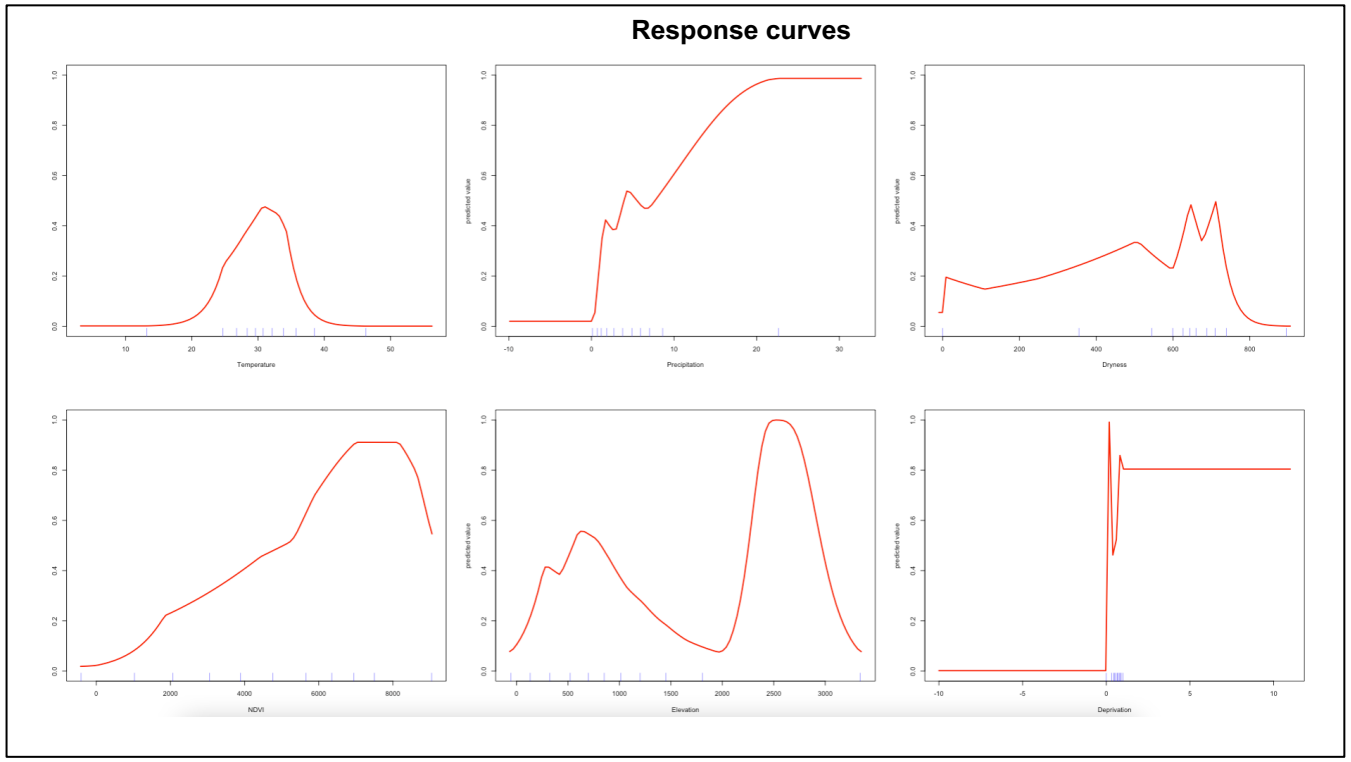
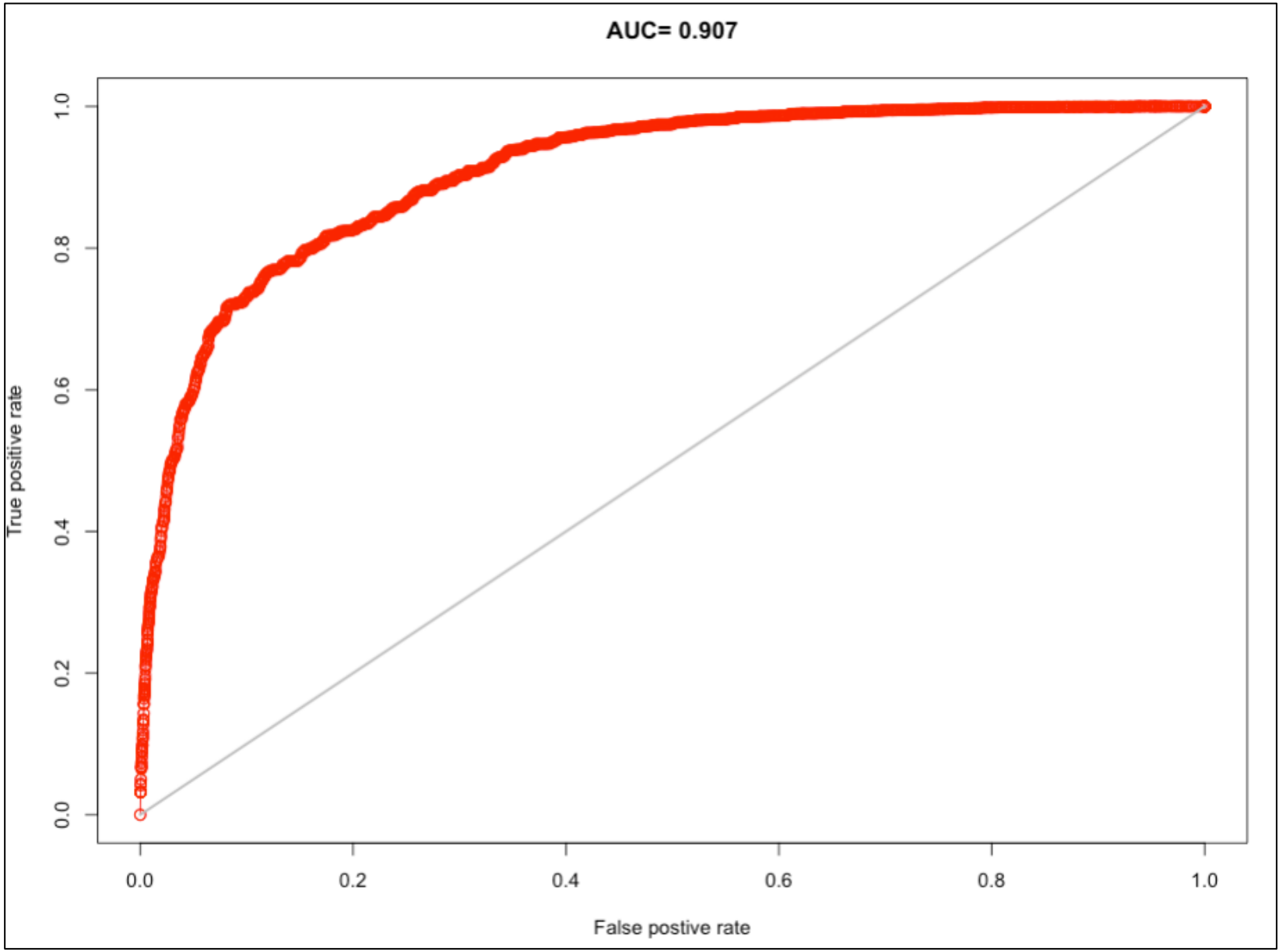


Figure panel shows from top left: Temperature (degree Celsius); Precipitation (mm); Dryness (Evapotranspiration) (mm/0.25 year); Vegetation (NDVI); Elevation (meters [m]); & Socioeconomic vulnerability index (%)

Interpretation: In the response plots, we are looking at how the probability of fire occurrence (Y-axes, from zero to one) varies with each the environmental predictors (X-axes). From these plots, we can see that the MAXENT models can include complex environmental responses including plateau, linear, and nonlinear shapes, and some which are utterly unclear. For example, if we look at mean temperature during the summer, we can see that the probability for fire occurrence peaks around 0.60 when temperatures are around 30 degrees Celsius. We can also see that the probability of such outcome increases with more and more vegetation during the summer period. Probability in terms of fires in relation to deprivation is a flat line. For precipitation, dryness and elevation - the patterns are unclear.

Using MAXENT

Result 2: Model validation

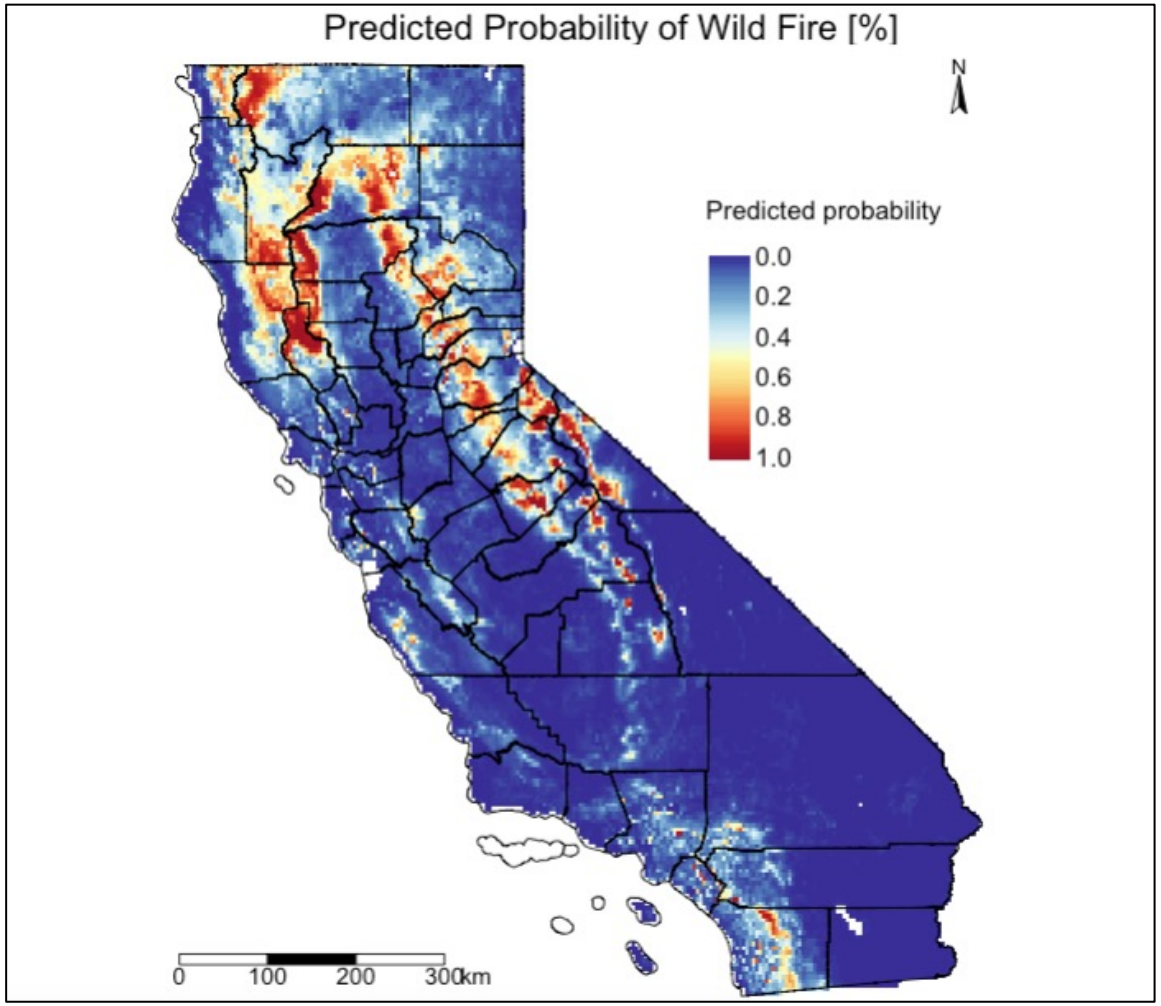


Interpretation: On the receiver operator curve (ROC), the 1:1 line give an AUC of 0.5. From our curve and the AUC, it is clear that our model appears to do substantially better than random guessing (high AUC value = **0.907 [90.7%]** > 0.5).

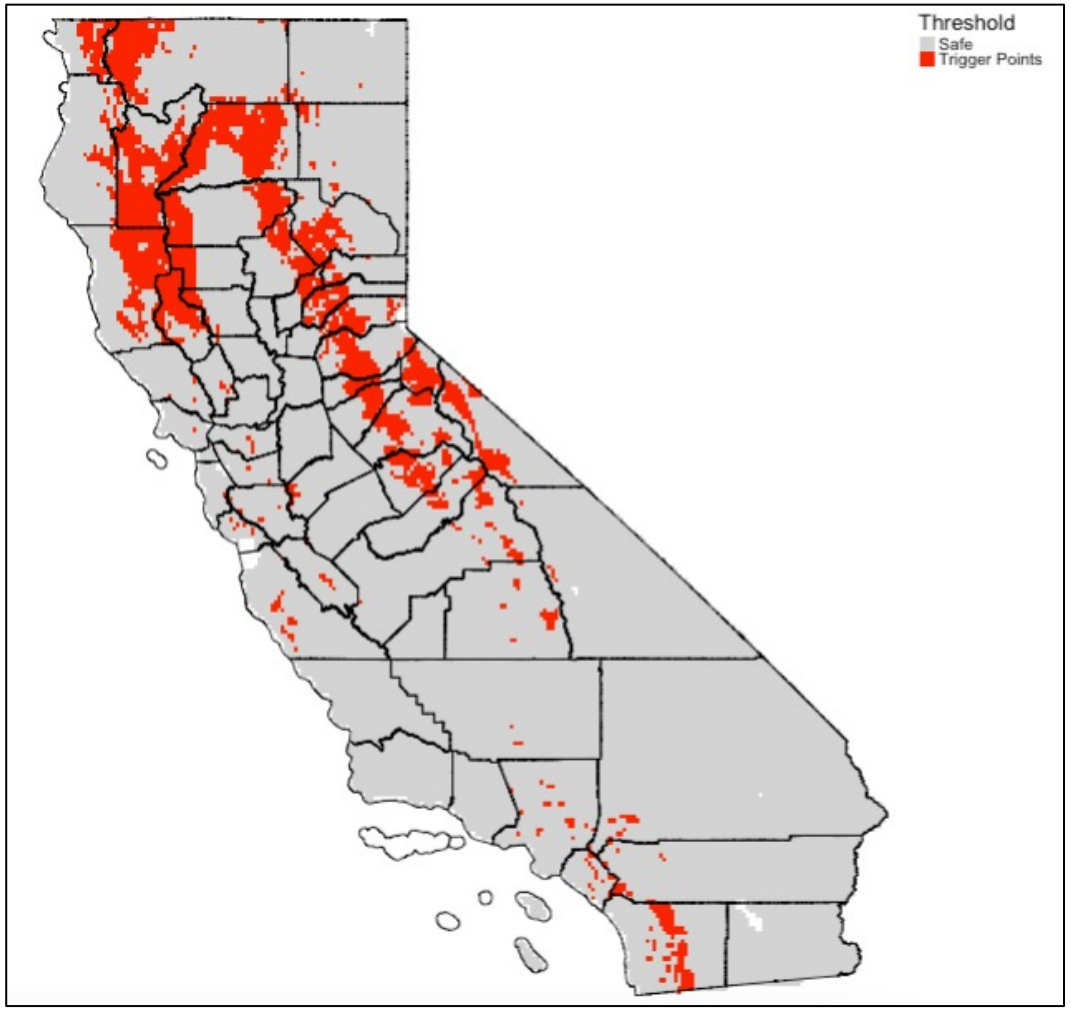
The optimal probability threshold at which our model maximizes the True Positive Rate and the True Negative Rate is 0.4054474 (40.55%). Hence, we will use **predicted probability > 0.4054** to delineate areas of suitability (or trigger points) for wildfires.

Using MAXENT

Result 3: Predicted probability of fire occurrence & areas for environmental suitability for fire hazards (i.e., trigger points)



We mapped predicted probability of fires using the trained model after making sure its valid. The multi-band raster is fed to the trained model to make full scale predictions.



This is based on the optimized estimate obtained after model validation i.e., maximizes the True Positive Rate and the True Negative Rate is 0.4054474 (40.55%). Here, we mapped **predicted probability > 0.4054** as a reclassified raster.

Any questions?