Exercise 1: Prepare data for a prediction model

How can I print an exercise to PDF format?

Software requirements

- ArcGIS Online
- ArcGIS Pro 3.2
- ArcGIS Spatial Analyst extension

Introduction

Although climate change is a global issue, it disproportionately affects vulnerable populations. In many parts of Africa, food and income security are tied to the agricultural production of crops like maize. But as climate change causes temperatures to become more extreme and climate-related hazards like drought to become more prevalent, food security for vulnerable populations is diminishing.

ArcGIS Pro includes a Presence-Only Prediction (MaxEnt) tool that applies the Maximum Entropy (MaxEnt) correlative machine learning model to predict the presence of a phenomenon. The machine learning model uses given known-presence locations and values of explanatory variables to make a prediction about a particular phenomenon. This week, you will use this tool to predict future suitable locations for maize cultivation in Africa. Known suitable locations for maize cultivation will serve as your presence points, and baseline and projected climate data will serve as your explanatory variables.

Before running a machine learning model, you must first create a training dataset. This dataset teaches, or trains, the model to perform a specific task—in this case, to predict suitable locations. The training dataset for the presence-only prediction model includes presence points and explanatory variables.

To learn more about using a MaxEnt model to assess relationships between climate and crop suitability, see the Land article, "Evaluation of Maximum Entropy (Maxent) Machine Learning to Assess Relationships between Climate and Corn Suitability."

Scenario

Imagine the following scenario: You are a GIS analyst for a global nonprofit organization working with agricultural ministers in Africa to protect farmers and ensure food security by planning for future climate change impacts. To better understand how climate change will affect where maize can be grown in Africa, you will use the Presence-Only Prediction (MaxEnt) tool to model current and future suitability using baseline variables against the SSP3 7.0 climate scenario for 2050.

In this exercise, you will prepare the training dataset, which includes the presence points and explanatory variables that you need to train the presence-only prediction model.

Note: The exercises in this course include View Result links. Click these links to confirm that your results match what is expected.

Estimated completion time in minutes: 90 minutes

Expand all steps   Collapse all steps

- Step 1: Download the exercise data files

In this step, you will download the exercise data files.

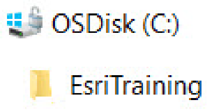
a Open a new web browser tab or window.

b Go to the CLIM Section 4: Maize Prediction Modeling item page.

Note: The complete URL to the exercise data file is https://www.arcgis.com/home/item.html?id=87f01a59ff7c4f4b8a83c5d818a5c64e

c On the right, click Download to download the exercise data ZIP file.

d In File Explorer, extract the ZIP exercise data files to the EsriTraining folder on your local computer.



*Step 1d***: Download the exercise data files.*

Remember that, throughout this course, you will save all your data to this folder. Do not add any spaces or special characters to the folder name.

e After you extract the ZIP file, confirm that the data files are stored in the MaizePredictionModeling folder.

f Close File Explorer.

You have downloaded and extracted the exercise data files.

- Step 2: Explore the explanatory variables

In a presence-only prediction model, the explanatory variables consist of baseline and projected variables that are used to compare change in the outputs of the model. For your model, you will compare 19 baseline bioclimate variables against 19 projected bioclimate variables that have been selected for a future decade based on a future climate scenario. ArcGIS Living Atlas of the World provides the data that you need for the explanatory variables in your prediction model.

In this step, you will go to ArcGIS Living Atlas of the World to explore these bioclimate variables.

a In a web browser, go to the following ArcGIS Living Atlas page: https://livingatlas.arcgis.com/en/browse/?q=worldclim%20by%20@esri_environment#d=2&q=worldclim+by+%40esri_environment.

b Review the list of bioclimate variables.

There are 19 bioclimate variables, some of which include annual mean temperature and precipitation of the driest month. For each bioclimate variable, there is a baseline layer and a projected layer that can be used for comparison. The layers contain multidimensional data, which is represented as data captured at multiple times or at multiple depths or heights or as multiple variables stored in one layer. Each bioclimate projection layer contains multidimensional data for three future climate scenarios.

You will explore the item details for two bioclimate layers.

c From the list of layers, click Bioclimate Projections: (01) Annual Mean Temperature.

d Read the item details for the layer, and then answer the following question.

? Which climate scenario is considered the high end of possibilities for future climate scenarios?

- Answer

SSP3 7.0 is considered the high end of possibilities for future climate scenarios.

e After you finish reading the item details, return to the list of bioclimate data layers.

f From the list of layers, click Bioclimate Baseline 1970-2000.

The Bioclimate Baseline layer provides each of the 19 bioclimate variables averaged for the period of 1970–2000. This layer represent the current, or baseline, conditions that you will use for your model.

g Read the item details, and then answer the following question.

> ? How can you access each of the 19 bioclimate variables stored in this multidimensional layer?
>
> - Answer
>
> You can use a multidimensional filter to access each of the 19 bioclimate variables.

h Close the web browser tabs.

You have explored the bioclimate variables from ArcGIS Living Atlas that you will use to train your prediction model and predict future suitability for maize production in Africa.

To learn more about the bioclimate data from ArcGIS Living Atlas, see ArcGIS Blog: Getting started with new multidimensional climate models.

- Step 3: Open an ArcGIS Pro project

In this step, you will open the ArcGIS Pro project that you downloaded.

a Start ArcGIS Pro.

b If necessary, sign in using your course ArcGIS account username (ending in _CLIM) and password.

c On the Start page, near Recent Projects, click Open Another Project.

Note: If you have configured ArcGIS Pro to start without a project template or with a default project, you will not see the Start page. On the Project tab, click Open, and then click Open Another Project.

d In the Open Project dialog box, browse to the MaizePredictionModeling folder that you saved on your computer.

- Hint

..\EsriTraining\MaizePredictionModeling

e Select MaizePredictionModeling.aprx and click OK.

*Step 3e***: Open an ArcGIS Pro project.*

The map opens and shows only a basemap. The data that you need for this exercise is stored in project folders and a project geodatabase.

- Step 4: Add data to the project

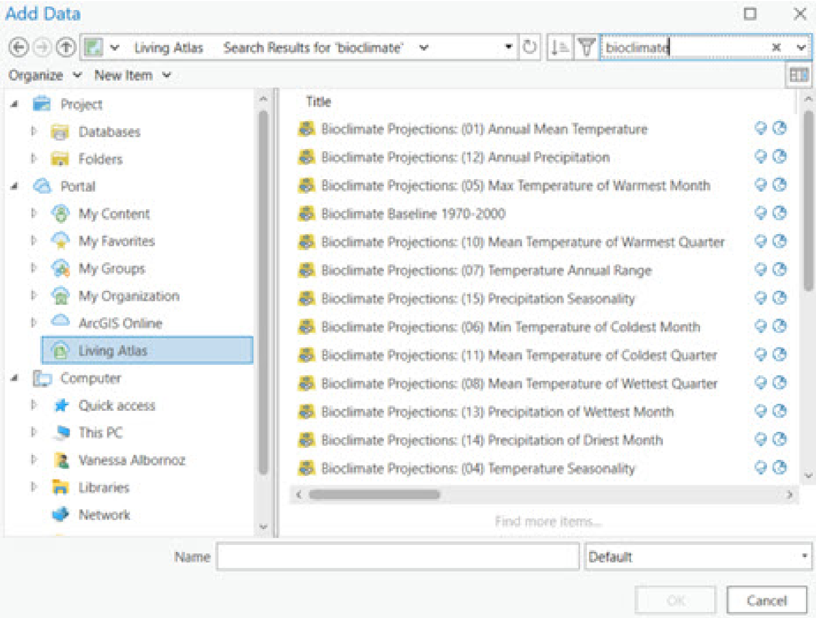
When working on a project in ArcGIS Pro, you can add data from several sources, including ArcGIS Living Atlas.

In this step, you will add the bioclimate data that you explored in ArcGIS Living Atlas.

a On the ribbon, in the Layer group, click the Add Data down arrow and choose Data.

b In the Add Data dialog box, on the left under Portal, click Living Atlas.

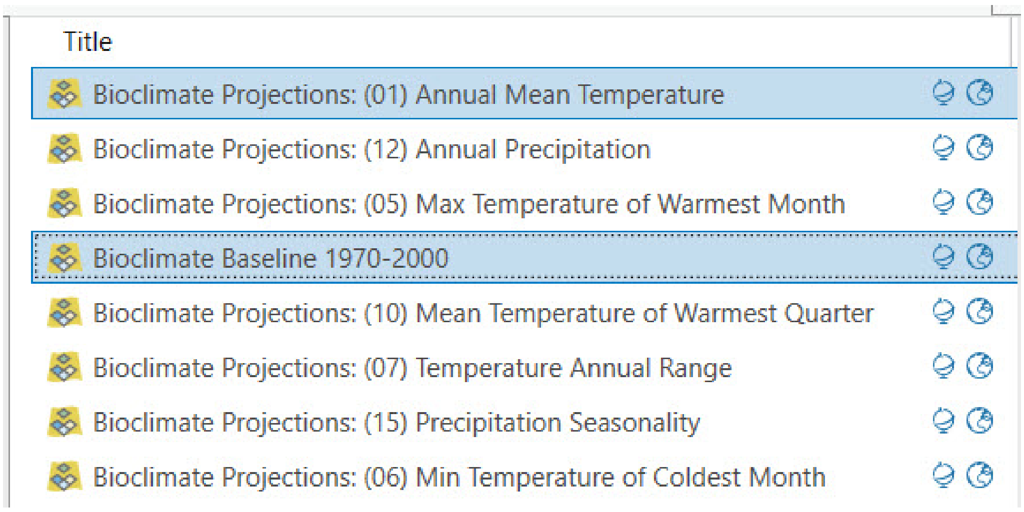
c At the top right of the dialog box, in the Search Living Atlas field, type **Bioclimate**, and then press Enter.



*Step 4c***: Add data to the project.*

All the bioclimate variables from ArcGIS Living Atlas are listed in the Add Data dialog box. You will add two of the bioclimate variables to your project: Bioclimate Projections: (01) Annual Mean Temperature and Bioclimate Baseline 1970-2000.

d In the Add Data dialog box, select Bioclimate Projections: (01) Annual Mean Temperature, press Ctrl on your keyboard, and then select Bioclimate Baseline 1970-2000.



*Step 4d***: Add data to the project.*

e Click OK.

Both layers are added to your map. You will use these layers to prepare explanatory variables for your prediction model.

- Step 5: Subset a multidimensional raster

Subsetting multidimensional data allows you to isolate one dimension of the data and use it as a separate explanatory variable. Subsetting is an important step that prevents you from accidentally averaging values of multiple dimensions for a single variable, which would skew the prediction model.

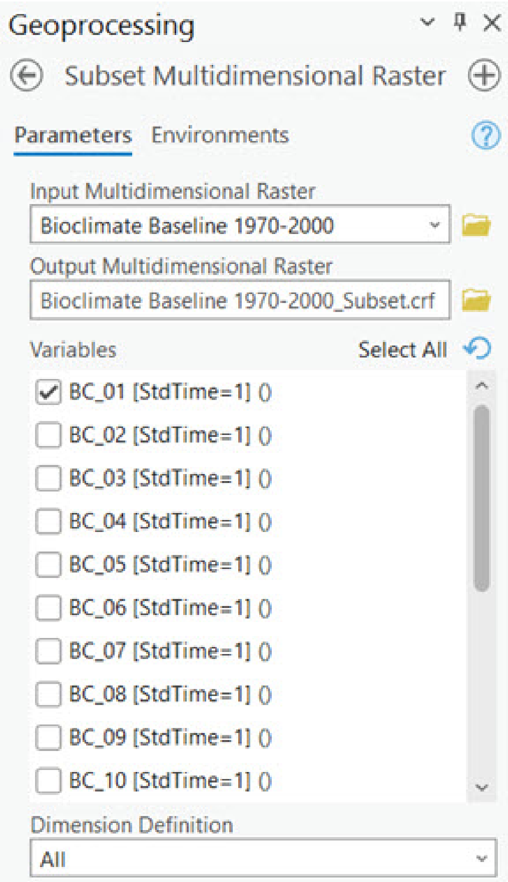
The layers that you added from ArcGIS Living Atlas were multidimensional. Therefore, in this step, you will subset only the layers that you will be using later as explanatory variables in your prediction model.

First, you will subset the multidimensional Bioclimate Baseline 1970-2000 layer to select only the baseline annual mean temperature variable. Then, you will subset the multidimensional Bioclimate Projections: (01) Annual Mean Temperature layer to select only the climate scenario and time period for the future projection of the annual mean temperature variable.

a In the Contents pane, if necessary, select the Bioclimate Baseline 1970-2000 layer.

b On the ribbon, click the Multidimensional tab.

c Click Data Management and choose Subset.



*Step 5c***: Subset a multidimensional raster.*

There are 19 Bioclimate Projection layers, and each of the 19 variables within the Bioclimate Baseline 1970-2000 layer is numbered to match a projection. For example, because you will be using the first projected variable, (01) Annual Mean Temperature, you will subset the first variable from the baseline layer, BC_01, to select the baseline annual mean temperature variable.

You will now set the tool parameters and run the tool to subset the first baseline variable for your prediction model.

d For Output Multidimensional Raster, rename the file to **BioclimateBaseline01.crf**.

e Under Variables, confirm that BC_01 [StdTime=1] () is checked.

f Click Run.

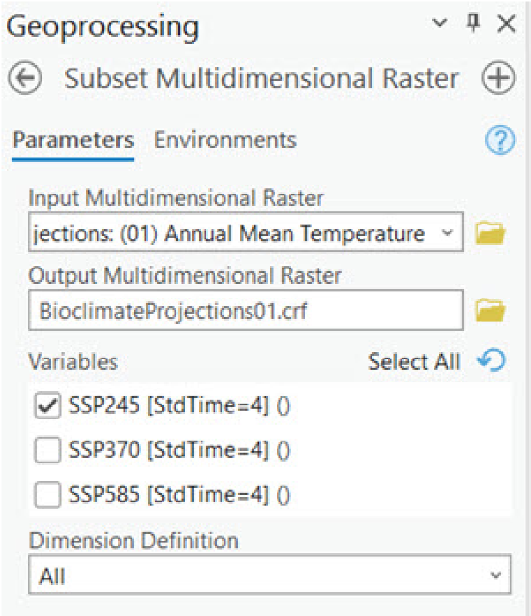
The multidimensional raster is added to your map.

By subsetting the first variable from the Bioclimate Baseline 1970-2000 layer, you now have 1 of the 19 variables needed to create a model to identify current suitable locations for maize production in Africa. For the sake of time, the other 18 baseline variables have already been subset for you using the same workflow.

You will now subset the Bioclimate Projections: (01) Annual Mean Temperature layer to select the climate scenario and time period for your prediction model.

g In the Geoprocessing pane, for Input Multidimensional Raster, click the down arrow and choose Bioclimate Projections: (01) Annual Mean Temperature.

h For Output Multidimensional Raster, rename the file to **BioclimateProjections01.crf**.



*Step 5h***: Subset a multidimensional raster.*

The bioclimate projection layer's variables include emissions scenarios. For your prediction model, you will use the SSP370 climate scenario and specify it for 2050.

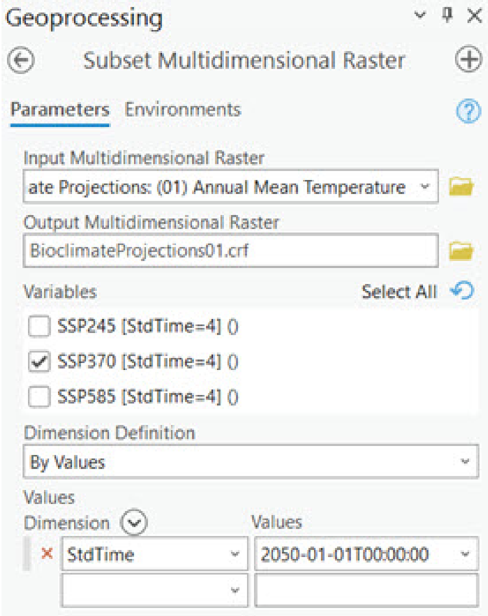
i Under Variables, uncheck the box for SSP245 [StdTime=4] ().

j Check the box for SSP370 [StdTime=4] () to select it as the emissions scenario.

k For Dimension Definition, click the down arrow and choose By Values.

l In the Values section, for the empty field under Dimension, click the down arrow and choose StdTime.

m For the field under Values, click the down arrow and choose 2050-01-01T00:00:00.



*Step 5m***: Subset a multidimensional raster.*

You have set the parameters to subset the annual mean temperature projection layer to select the SSP370 climate scenario for 2050.

n Click Run.

The multidimensional raster is added to your map.

By subsetting the Bioclimate Projections: (01) Annual Mean Temperature layer, you now have 1 of the 19 projection variables needed to create a model to predict future suitable locations for maize production in Africa. For the sake of time, the other 18 bioclimate projection variables have been subset for you using the same workflow.

Because the other explanatory variables have already been subset, you can remove the bioclimate layers that are currently in the Contents pane. Removing layers that you do not need keeps your project organized. It is important to note that removing a layer from the Contents pane does not delete the layer from your project–the bioclimate variables that you subset are stored in a project folder.

o In the Contents pane, remove the four bioclimate layers: Bioclimate Baseline 1970-2000 layer, Bioclimate Projections: (01) Annual Mean Temperature layer, BioclimateBaseline01.crf, and BioclimateProjections01.crf.

- Hint

  In the Contents pane, click the first bioclimate layer and then press the Ctrl key to select the other three bioclimate layers. Right-click the selection and choose Remove.

p Save your project.

You have learned how to subset multidimensional raster data for your baseline and projected bioclimate variables. The baseline bioclimate variables will be used as explanatory variables for your training dataset, and the bioclimate projection variables will be used when you are ready to run your prediction model. The remaining variables needed for your prediction model were already subset for you using the workflow that you learned in this step.

- Step 6: Generate points

In addition to explanatory variables, your prediction model will require presence points to identify locations of known presence. To create the presence points, you must create random points within a study area before combining multiple data sources that identify presence. Using multiple data sources to identify presence covers any gaps or failures that may exist in one dataset alone.

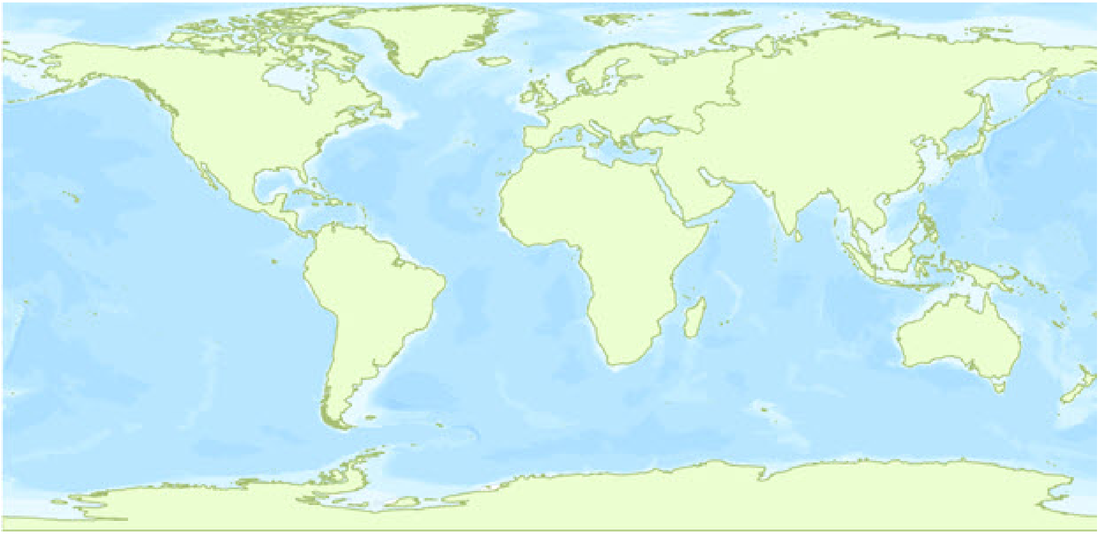
In this step, you will begin creating the training dataset of random points across your study area, Africa.

To generate points in only your study area, you must first create a query that identifies the boundary of the African continent. You will add a layer from ArcGIS Living Atlas to set this boundary.

a On the ribbon, click the Add Data icon and, if necessary, choose Living Atlas.

b Search for **World Continents**.

c Add the World Continents layer from ArcGIS Living Atlas to your map.



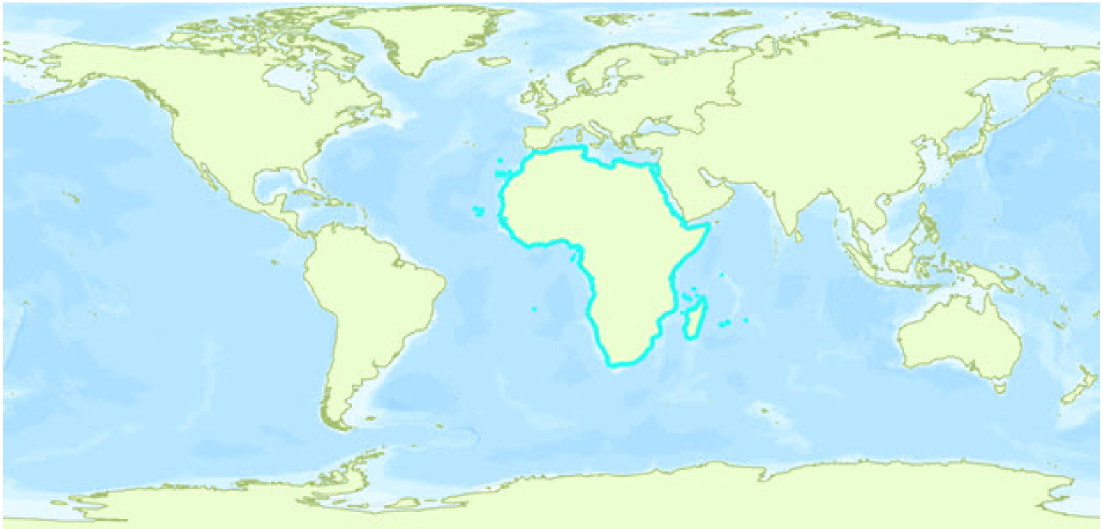
*Step 6c***: Generate points.*

The World Continents layer contains the boundaries for all continents, so you will create a query to select Africa.

d On the ribbon, click Select By Attributes.

e Create the query `Where continent is equal to Africa`.

f Click Apply, and then click OK to close the tool.



*Step 6f***: Generate points.*

With Africa now selected in the layer, you can generate random points in your study area. This will result in a point feature class, which is a layer that contains only points.

g Find and open the **Create Random Points** (Data Management) tool.

- Hint

    In the Geoprocessing pane, click the Back button. In the search field, type **Create Random Points** and press Enter. Click the Create Random Points (Data Management Tools) tool to open it.

You will specify several parameters for the tool, including the name of the point feature class and the geodatabase where you will store the point feature class. You will also set the constraining feature class to your study area of Africa and specify the number of points to be created.

h For Output location, leave the default geodatabase.

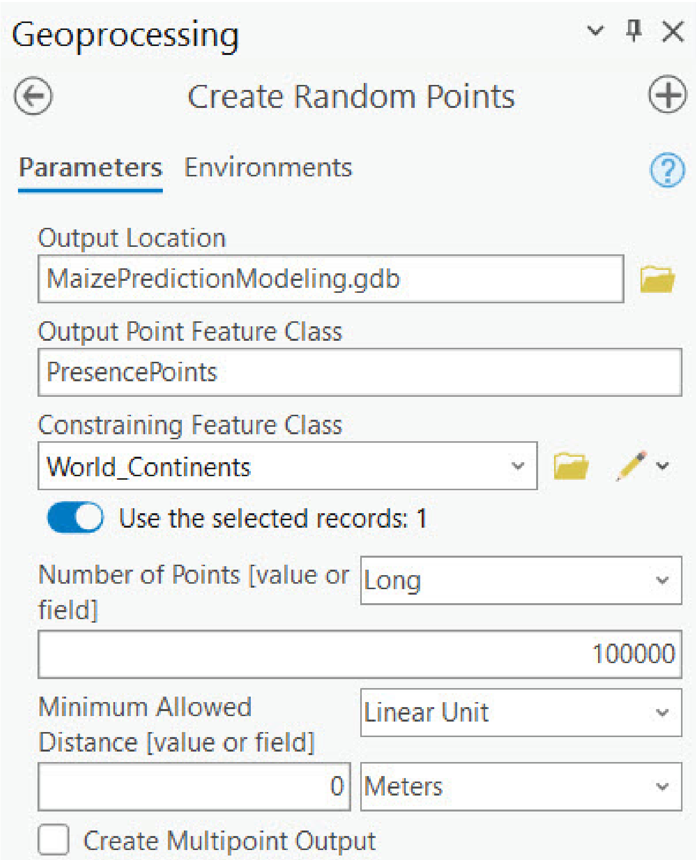
i For Output Point Feature Class, type **PresencePoints**.

j For Constraining Feature Class, click the down arrow and choose World_Continents.

Because you have Africa selected in the World Continents layer, the tool will use the selection to set the constraining boundary for the points to be created.

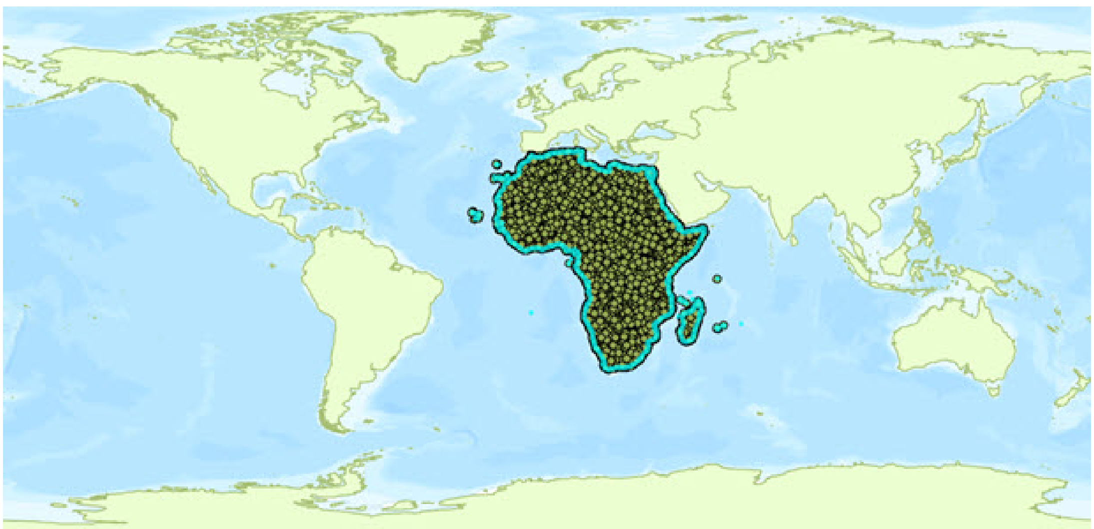
k Under Number of Points [Value Or Field], delete 100 and type **100000**.

The number of points that you choose to create depends on factors such as the size of your study area and the distance that you want between points.



*Step 6k***: Generate points.*

l Click Run.



*Step 6l***: Generate points.*

You have created a point feature class of 100,000 random points within your study area of Africa.

m In the Contents pane, turn off the PresencePoints and World_Continents layers.

n Save your project.

- Step 7: Reclassify data

After the random points are created for the prediction model's training dataset, the data sources specifying areas of known presence need to be incorporated into the training dataset. However, the data sources may require preparation if any additional values are needed to identify the presence of maize. Using the Reclassify tool, you can change values to specify only those that are needed for the prediction model.

In this step, you will reclassify the values for two raster datasets to specify the presence of maize.

a In the Contents pane, right-click PresencePoints and choose Zoom To Layer.

The first raster is the GFSAD.tif file, which is a global dataset that was created by NASA (National Aeronautics and Space Administration) using satellite imagery to identify the dominant crop in an area.

b Find and open the **Reclassify** (Spatial Analyst Tools) tool.

c For Input Raster, click the down arrow and choose GFSAD.tif.

The layer uses values to identify the dominant crop types. You are interested in maize, which is represented by values of 4, 6, and 7 in the dataset. You will reclassify the 4, 6, and 7 values as "1" to indicate that maize is present and then reclassify all other values as "0" to indicate that maize is not present.

d Confirm that Reclass Field is set to Value.

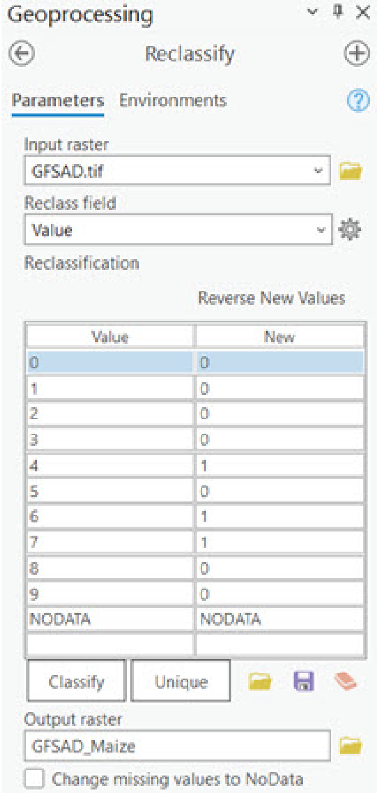
e Under Reclassification, click Unique.

f In the table, for the value 0, double-click the 1 in the New column and type **0**.

g Reclassify the remaining values using the following table as a guide.

Value	New
1	**0**
2	**0**
3	**0**
4	**1**
5	**0**
6	**1**
7	**1**
8	**0**
9	**0**

h For Output Raster, type **GFSAD_Maize**.



*Step 7h***: Reclassify data.*

i Click Run.

The GFSAD_Maize layer is added to the Contents pane and shows only two values in the layer: 0 for where maize is not grown and 1 for where maize is grown.

You need to reclassify the second dataset to identify where maize is grown. However, this dataset is hosted on ArcGIS Online, so you must first add it to your project.

j In the Contents pane, turn off the GFSAD_Maize layer.

k Open the Add Data dialog box.

- Hint

On the ribbon, click the Map tab and, in the Layer group, click Add Data.

l Under Portal, click ArcGIS Online.

m Search for **Africa Crop Maize**.

n Select the Africa Crop Maize - Harvested Area layer and click OK.

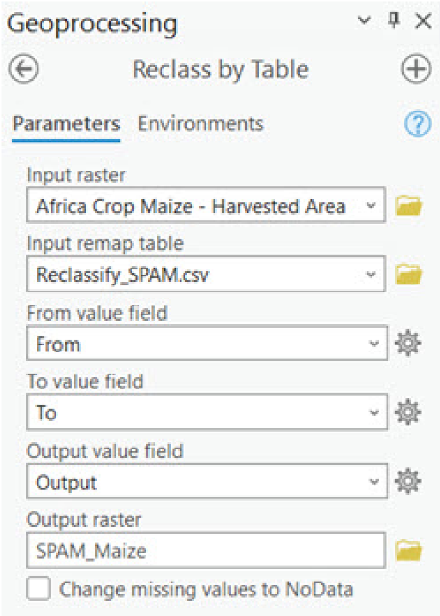
The Africa Crop Maize - Harvested Area layer shows the average annual area of maize harvested in Africa from 1999 to 2001. This dataset was created by spatially disaggregating national and subnational harvest data using a modeled crop presence workflow called the Spatial Production Allocation Model (SPAM). Using the SPAM workflow, the yield of maize harvested across Africa is provided in units of hectares per cell. The dataset uses continuous values, rather than unique values, to identify areas where maize is grown. For this dataset, values of 180 or less represent areas of higher amounts of hectares of harvested maize.

In accordance with the SPAM model, you will reclassify values that are less than or equal to 180 as "1" to indicate areas where more maize is harvested. You will also reclassify any values greater than 180 as "0" to indicate areas where either no maize or less maize is harvested. Rather than typing in these values, you will utilize a CSV file, Reclassify_SPAM, that has been stored in your data.

o In the Geoprocessing pane, open the **Reclass By Table** (Spatial Analyst Tools) tool.

p Set or confirm the following parameters:

- Input Raster: Africa Crop Maize - Harvested Area
- Input Remap Table: Reclassify_SPAM.csv
- From Value Field: From
- To Value Field: To
- Output Value Field: Output
- Output Raster: **SPAM_Maize**



*Step 7p***: Reclassify data.*

q Click Run.

The SPAM_Maize layer is added to the map and shows the reclassified values: 1 for areas with higher amounts of maize harvested and 0 for areas with no or lesser amounts of maize harvested.

r Turn off the SPAM_Maize and Africa Crop Maize - Harvested Area layers.

s Save your project.

You have reclassified the values for two raster datasets to identify where maize is grown. Using multiple datasets to confirm the presence of maize covers any gaps or failures in one dataset alone.

- Step 8: Extract values from raster datasets

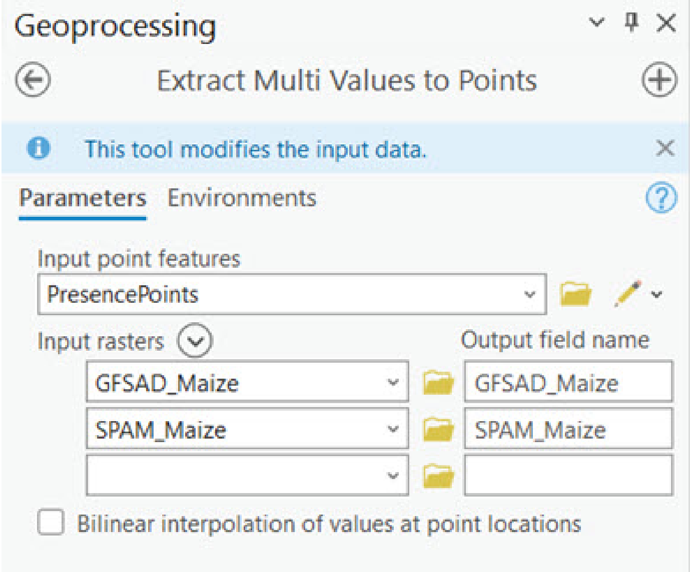
The presence-only prediction model only requires input data that represents known-presence locations. With the values from the data sources reclassified to specify presence locations for maize, you can now extract and add those values to the point feature class layer of the study area, PresencePoints. These values can then be recorded in the layer's attribute table for developing presence points.

In this step, you will extract the values from the GFSAD_Maize and SPAM_Maize raster datasets to the PresencePoints layer to identify areas where maize is present in both raster datasets.

a In the Geoprocessing pane, open the **Extract Multi Values To Points** (Spatial Analyst Tools) tool.

b For Input Point Features, click the down arrow and choose PresencePoints.

c For Input Rasters, click the Add Many button, check GFSAD_Maize and SPAM_Maize, and click Add.



*Step 8c***: Extract values from raster datasets.*

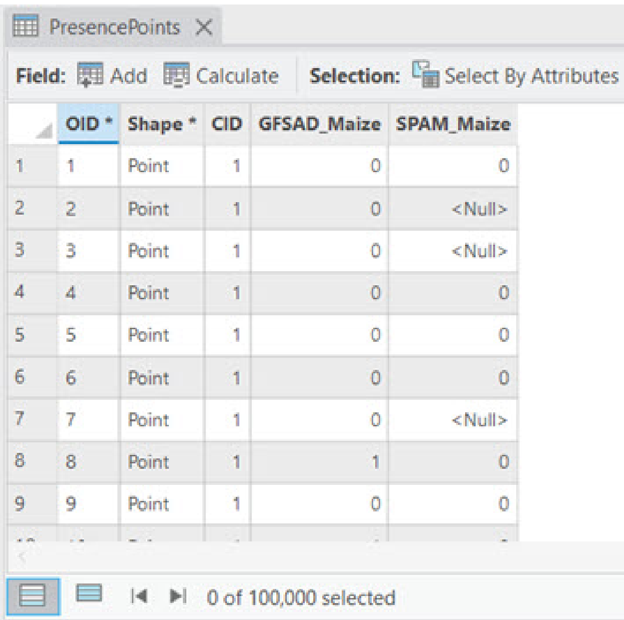
d Click Run.

You have used the Extract Multi Values To Points tool to extract the values from the GFSAD_Maize and SPAM_Maize raster layers to the PresencePoints layer.

e In the Contents pane, open the PresencePoints attribute table.

- Hint

Right-click PresencePoints and choose Attribute Table.



	OID *	Shape *	CID	GFSAD_Maize	SPAM_Maize
1	1	Point	1	0	0
2	2	Point	1	0	<Null>
3	3	Point	1	0	<Null>
4	4	Point	1	0	0
5	5	Point	1	0	0
6	6	Point	1	0	0
7	7	Point	1	0	<Null>
8	8	Point	1	1	0
9	9	Point	1	0	0

*Step 8e***: Extract values from raster datasets.*

The GFSAD_Maize and SPAM_Maize fields have been added to the PresencePoints attribute table.

You will notice that the GFSAD layer is a TIF file and the SPAM layer is an image service. Both formats are common GIS raster data formats; however, image services provide more functionality for raster analyses. This difference in data format explains why, in the attribute table, the SPAM_Maize field displays null values to represent areas of no data and zeroes to represent areas where maize is not present, whereas the GFSAD_Maize field displays zeroes to represent areas that have either no data or no maize present.

f Turn on the PresencePoints layer.

g Save your project.

You have extracted the values from two raster datasets, GFSAD_Maize and SPAM_Maize, to the PresencePoints layer to identify areas areas where maize is present in both raster datasets.

- Step 9: Identify presence with attributes

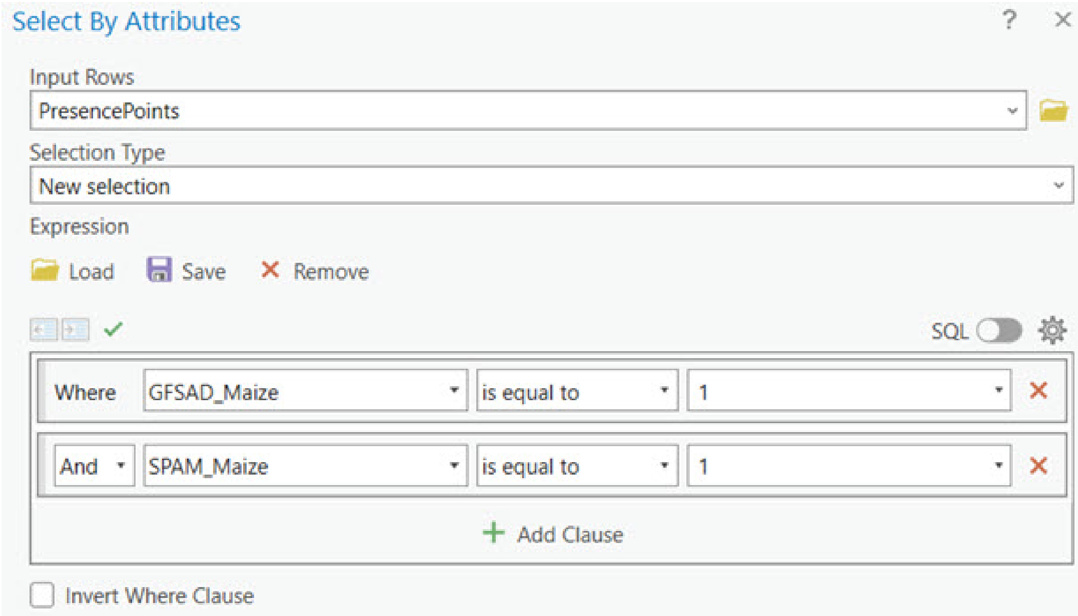
After the values that specify presence locations from two data sources are added, the common presence locations in both data sources can be identified.

In this step, you will use the attribute table to select where maize is present in both raster datasets to create a new field in the attribute table. Using more than one dataset to identify locations that are suitable for maize helps minimize errors in your data.

a In the PresencePoints attribute table, click Select By Attributes.

You will write an expression to identify points where maize is present in both datasets, as indicated by the value 1.

b Create a clause that states `Where GFSAD_Maize Is Equal To 1 And SPAM_Maize Is Equal To 1`



*Step 9b***: Identify presence with attributes.*

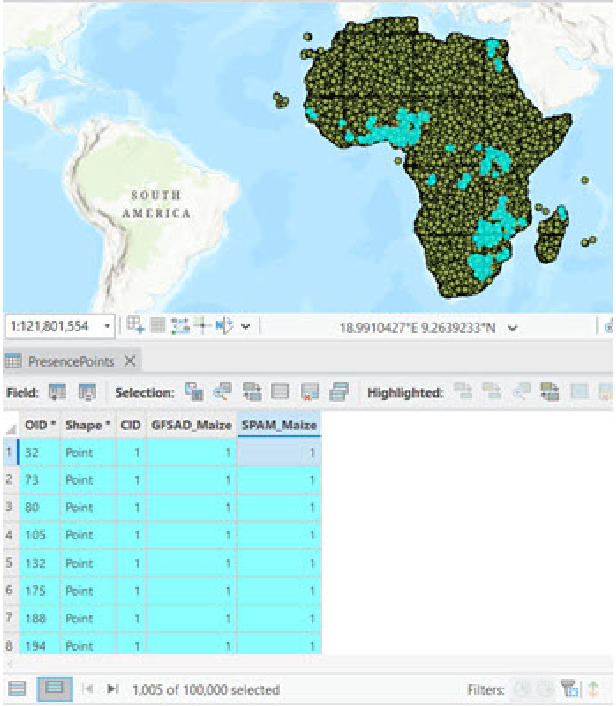
c Click Apply, and then click OK.

d At the bottom of the attribute table, click the Show Selected Records button .

e Zoom to the PresencePoints layers.

- Hint

In the Contents pane, right-click PresencePoints and choose Zoom To Layer.



*Step 9e***: Identify presence with attributes.*

Because random points are generated, the number of selected points may vary slightly.

The map and attribute table update to highlight the selected points. The selected points represent the points where maize is present in both the GFSAD layer and the SPAM layer.

To save the selection, you will calculate a new field.

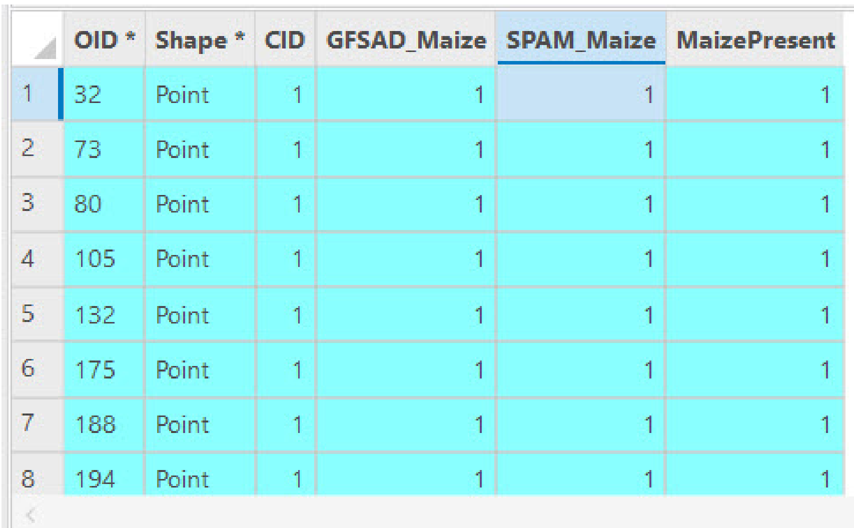
f In the attribute table, click the Calculate Field button .

g For Field Name, type **MaizePresent**.

h For Field Type, choose Short (16-bit Integer).

i Under MaizePresent=, type **1**

j Click Apply, and then click OK to close the Calculate Field tool.



	OID *	Shape *	CID	GFSAD_Maize	SPAM_Maize	MaizePresent
1	32	Point	1	1	1	1
2	73	Point	1	1	1	1
3	80	Point	1	1	1	1
4	105	Point	1	1	1	1
5	132	Point	1	1	1	1
6	175	Point	1	1	1	1
7	188	Point	1	1	1	1
8	194	Point	1	1	1	1

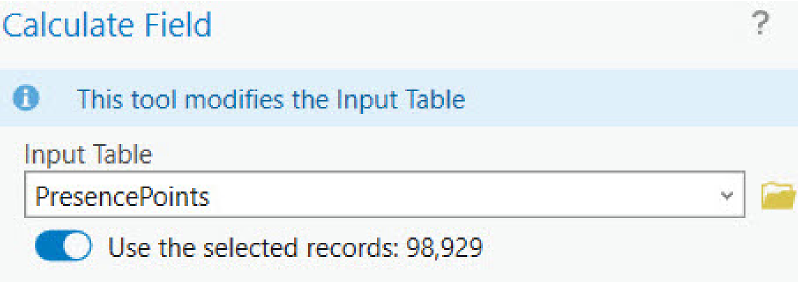
*Step 9j***: Identify presence with attributes.*

A new field called MaizePresent is created. This field combines the points identifying where maize is present in both the GFSAD layer and the SPAM layer.

You will now switch the selection in the attribute table to calculate the selected fields that should be zero, to reflect the points in both layers where maize is not present.

k In the attribute table, click the Switch Selection button .

l Open the Calculate Field tool.



*Step 9l***: Identify presence with attributes.*
Because random points are generated, you may see a slightly different number.

The input selection refreshes to show the number of selected points that will be set to zero.

m Under MaizePresent =, delete 1 and type **0**

n Click Apply, and then click OK.

You have created a new field in the PresencePoints attribute table that identifies where maize is present in both the GFSAD and SPAM raster datasets.

o Click the Clear Selection button and close the attribute table.

p Save your project.

- Step 10: Extract baseline bioclimate data

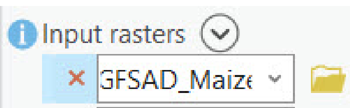
The final step in creating a training dataset for the presence-only prediction model is to add the explanatory variables.

In this step, you will add the baseline bioclimate variables for each presence point to complete the training dataset for your prediction model. To do so, you will use the Extract Multi Values To Points tool to add the baseline bioclimate variables to your PresencePoints layer.

You previously ran the tool, so it is still open in the Geoprocessing pane. You simply need to remove the GFSAD_Maize and SPAM_Maize layers before running the tool again with the new parameters.

a In the Geoprocessing pane, for Input Point Features, ensure that PresencePoints is selected.

b Under Input Rasters, point to GFSAD_Maize and click the red x to remove the layer, as shown in the following graphic.



c Use the same process to remove the SPAM_Maize layer.

You will now add the baseline bioclimate variables to the tool.

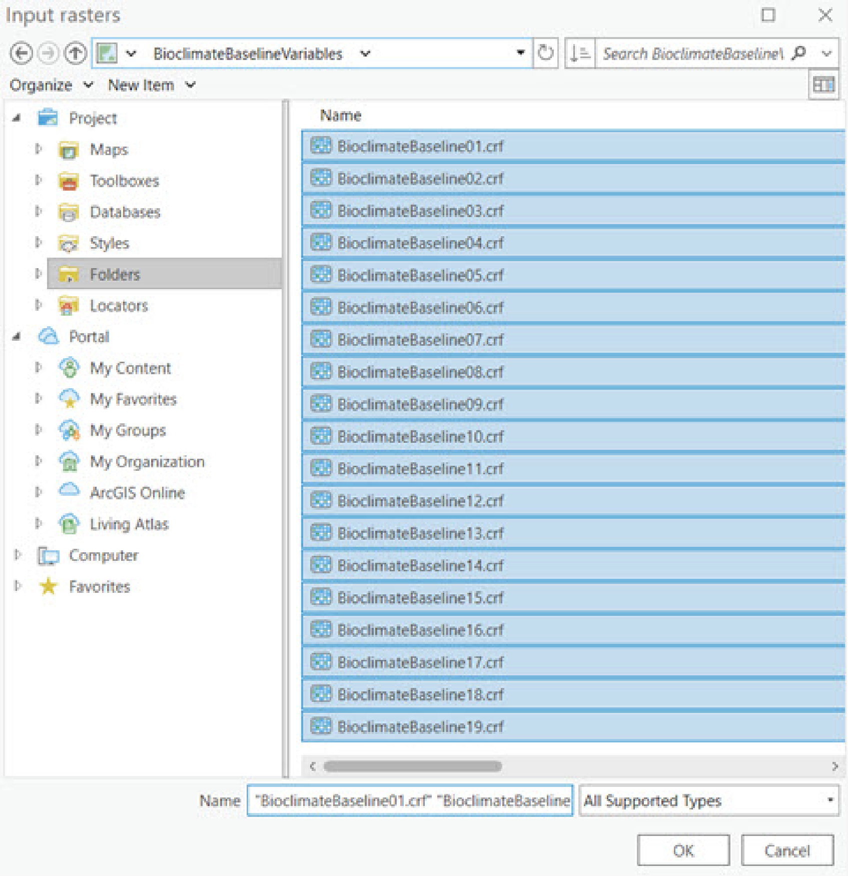
d For Input Rasters, click the Browse button .

e On the left, under Project, click Folders, and then double-click the MaizePredictionModeling folder.

f Double-click the BioclimateBaselineVariables folder.

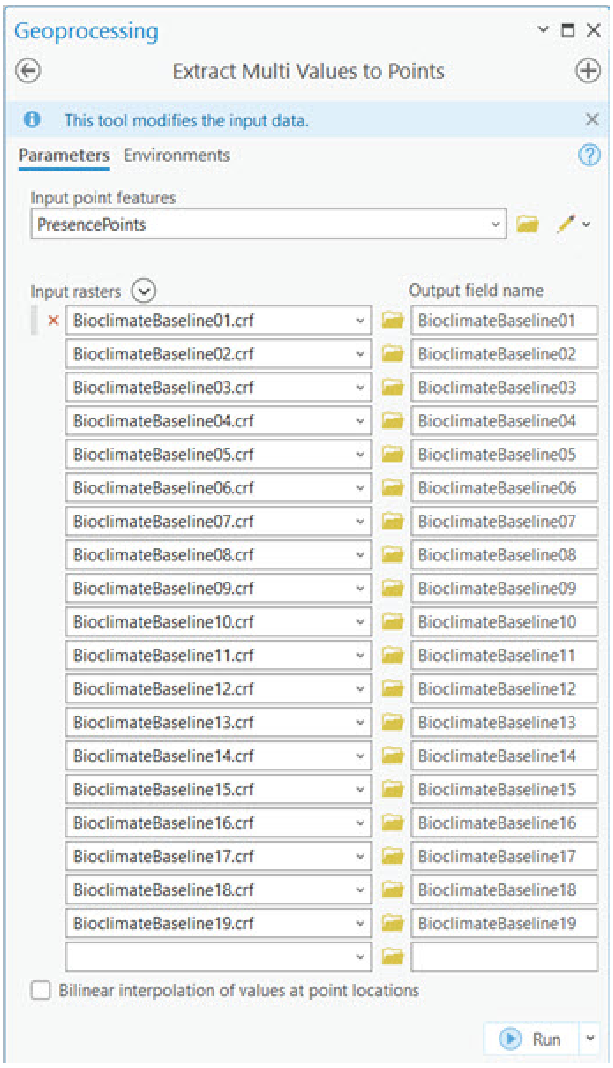
g Click BioclimateBaseline01.crf.

h Press the Shift key and click BioclimateBaseline19.crf to select all the files.



*Step 10h***: Extract baseline bioclimate data.*

i Click OK to add the selected baseline bioclimate variables to Input Rasters.



*Step 10i***: Extract baseline bioclimate data.*

j Click Run.

k Open the PresencePoints attribute table.

The 19 baseline bioclimate variables are added to the PresencePoints attribute table. Each baseline bioclimate variable has a value that is associated to a point. These values are based on different calculation methods for each variable. The amount of a value that is present in the location of the point is the value that is added to the attribute table. For example, some points may have a value of zero as a baseline bioclimate variable; a value of zero means that the baseline bioclimate variable has no effect on the location of that point.

For more information on the calculation methods for the bioclimate variables, see the USGS publication "Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States."

l Close the attribute table.

m Save your project.

In this exercise, you created a training point dataset for your prediction model by identifying areas of known maize presence in Africa based on two raster datasets and then combining 19 baseline bioclimate variables to incorporate explanatory variables.

n If you are continuing to the next exercise, leave ArcGIS Pro open; otherwise, exit ArcGIS Pro.