# MSA\_QGIS manual

(actual name pending…)

Contents

[MSA\_QGIS manual 1](#_Toc130386877)

[Intro 4](#_Toc130386878)

[Prior Knowledge 4](#_Toc130386879)

[Resources 4](#_Toc130386880)

[Installation 4](#_Toc130386881)

[QGIS 4](#_Toc130386882)

[MSA\_QGIS 5](#_Toc130386883)

[Before opening the plugin 5](#_Toc130386884)

[Setting the Coordinate Reference System 5](#_Toc130386885)

[Importing maps 5](#_Toc130386886)

[Input – Spatial and Environmental Input 6](#_Toc130386887)

[Starting Point 6](#_Toc130386888)

[Map Style 6](#_Toc130386889)

[Area of Interest 7](#_Toc130386890)

[Resolution (Simple) 8](#_Toc130386891)

[Resolution (Nested) 8](#_Toc130386892)

[Size of Nested Area 8](#_Toc130386893)

[Available fields/bands: 8](#_Toc130386894)

[Input – Vegetation Input 9](#_Toc130386895)

[Import HUMPOL handbag (optional): 9](#_Toc130386896)

[Taxa 9](#_Toc130386897)

[Communities 9](#_Toc130386898)

[Input – Rules 9](#_Toc130386899)

[Add new rule 10](#_Toc130386900)

[Remove rule 11](#_Toc130386901)

[Edit rule 11](#_Toc130386902)

[Input – Rule Tree 11](#_Toc130386903)

[Insert rule below 12](#_Toc130386904)

[Insert rule in series 12](#_Toc130386905)

[Designate as base group 12](#_Toc130386906)

[Delete branch 12](#_Toc130386907)

[View rule list 12](#_Toc130386908)

[Input – Pollen input 12](#_Toc130386909)

[Sample sites 12](#_Toc130386910)

[Pollen percentage file paths 13](#_Toc130386911)

[Excerpt pollen percentages 13](#_Toc130386912)

[Input – Model Parameters 13](#_Toc130386913)

[Model for distance weighting. 13](#_Toc130386914)

[Fit calculation method 13](#_Toc130386915)

[Desired fit per site 14](#_Toc130386916)

[Desired fit combined 14](#_Toc130386917)

[Data to keep 14](#_Toc130386918)

[Model – HUMPOL mire model 14](#_Toc130386919)

[Model – HUMPOL mixed-basin model [nyi] 14](#_Toc130386920)

[Model – 1/d [nyi] 14](#_Toc130386921)

[Model – 1/d2 [nyi] 14](#_Toc130386922)

[Model – Use Lookup Table [nyi] 15](#_Toc130386923)

[Input – Metadata 15](#_Toc130386924)

[Author(s) and Project name 15](#_Toc130386925)

[Description 15](#_Toc130386926)

[Project changelog 15](#_Toc130386927)

[Notes 15](#_Toc130386928)

[Input – Checklist 15](#_Toc130386929)

[Input – Order of Operations 15](#_Toc130386930)

[Notes on input 16](#_Toc130386931)

[Save 17](#_Toc130386932)

[Load 18](#_Toc130386933)

[Running the plugin/ Output 19](#_Toc130386934)

[Notes on output 22](#_Toc130386935)

[What to do if the MSA run has been/needs to be interrupted. 22](#_Toc130386936)

[Warnings and Errors 23](#_Toc130386937)

[Foreseen Input warnings and errors 23](#_Toc130386938)

[Foreseen Runtime errors 23](#_Toc130386939)

[Unforeseen errors 23](#_Toc130386940)

[Crashes 23](#_Toc130386941)

[Credit 24](#_Toc130386942)

[Cite MSA\_QGIS 24](#_Toc130386943)

[Contact 24](#_Toc130386944)

## Intro

MSA\_QGIS is a plugin for QGIS3 that allows the use of the Multiple Scenario Approach for pollen using the QGIS interface.

The Multiple Scenario Approach is a framework for simulating plausible vegetation maps from pollen percentage data. The approach works by taking maps of the environment, such as waterways, height and/or geology, and applying probabilistic and deterministic rules to assign certain vegetation types. This assignment is done in an iterative way, which means many different vegetation maps are created. From these maps, pollen percentage data is then generated for chosen sites using a Pollen Dispersal and Deposition model. The generated pollen percentages can then be compared to actual pollen percentages, in order to determine which of the simulated vegetation maps are plausible, or can be examined on their own for thought experiments.

The MSA\_QGIS plugin was made with the FAIR (Findable, Accessible, Interoperable, Reusable) principles of open science in mind. Input and output files are in file formats that are intended to be accessible, timeless and readable, so that they may easily be shared and understood. The software and plugin are free and openly available, and the code can be examined and contributed to by everyone (with some moderation, of course). The interface is meant to be user- and beginner friendly, but also flexible. If you use this plugin, please keep the spirit of open science in mind and share your science and data freely if you can.

## Prior Knowledge

The MSA\_QGIS plugin is strictly a tool and could be used without prior knowledge, but this is strongly discouraged for anything other than learning. In no circumstances should you apply MSA\_QGIS professionally unless you:

* Understand the assumptions and limitations of the Pollen Dispersal and Deposition model you are applying
* Understand the assumptions and limitations of the Multiple Scenario Approach.
* Understand the basics of Coordinate Reference Systems and map projections.

Please consult the resources below if you are unsure.

### Resources

* [A gentle Introduction to GIS](https://docs.qgis.org/3.22/en/docs/gentle_gis_introduction/index.html)
* [Bunting & Middleton, 2005, Modelling pollen dispersal and deposition using HUMPOL software, including simulating windroses and irregular lakes.](https://doi.org/10.1016/j.revpalbo.2004.12.009)
* [Bunting & Middleton, 2009, Equifinality and uncertainty in the interpretation of pollen data: The Multiple Scenario Approach to reconstruction of past vegetation mosaics.](https://doi.org/10.1177/0959683609105304)
* [paper(s) for Prentice-Sugita-Sutton]
* [paper(s) for RPPE]
* [Paper(s) for 1/d]
* [Paper(s) for 1/d^2]
* [paper(s) for RSAP]
* [Documentation for QGIS3.22](https://docs.qgis.org/3.22/en/docs/)

## Installation

### QGIS

As MSA\_QGIS is a plugin for QGIS, installing QGIS is required. QGIS is a free and open source GIS (Geographical Information System) that is updated regularly.

QGIS can be downloaded from their website: <https://qgis.org>

#### version

MSA\_QGIS was produced using QGIS3.22.7, as such earlier versions of QGIS are discouraged. Later version should be fine, although function cannot be guaranteed unless they have been tested. The type of download (OSGeo4W or regular install) does not matter.

MSA\_QGIS has currently been tested on the following versions:

* QGIS 3.22.7-Białowieża for Windows 10

### MSA\_QGIS

The plugin itself can be installed in 2 ways, from the QGIS plugin repository, or installing directly from ZIP.

#### From repository

[nyi]

#### From ZIP

A Zip folder containing QGIS can be downloaded from [GitHub link]. Once downloaded, the zip folder should **NOT** be unzipped. Follow these steps:

* Go to “Plugins” in the Menu bar of QGIS.
* Open “manage and install plugins”
* Click “Install from ZIP”
* Click the ellipsis and find the zip folder containing MSA\_QGIS in your file explorer (or alternatively, paste the URL to the zip folder in the input box).
* Click install plugin.

new small icon should have appeared in the toolbar. If it has not, make sure the plugin toolbar is turned on by going to “view” 🡪 “toolbars” 🡪 and check “plugin toolbar”.

## Before opening the plugin

### Setting the Coordinate Reference System

The coordinate reference system (CRS) should be set to the **projected** coordinate reference system that is most accurate for the area of interest. MSA\_QGIS will not produce accurate results when using geographical coordinate reference systems, as the maths being done in the background will not work with degrees.

### Importing maps

Any maps you would like to use as input for the creation of the simulated vegetation maps need to be imported prior to opening the plugin. Raster and vector polygon maps of any file type can be used. To be able to use lines and points you should create a vector polygon layer using a buffer. Keep in mind the [resolution](#_Resolution_(Simple)) of your final project, since the buffer needs to be wide enough to overlap with the points of the final map, or else it will not influence your data at all. Polygon layers may not have polygons that overlap.

MSA\_QGIS currently does not have the capability to place vegetation communities at given coordinates. If you do want to place vegetation communities at given coordinates, you need to create a new map layer and draw (or generate) the area that you would like to fill prior to running the plugin, using the (advanced) editing tools of QGIS.

## Input – Spatial and Environmental Input

### Starting Point

The starting point can be changed if you have previously run the plugin and already have a point-sampled map or basemap, or if you want to use your own, custom made map. There are three options:

#### Create a map from scratch

Use this option if you do not yet have any point vegetation map and you want to generate one using MSA-QGIS. This will likely be the case if you have not run the plugin before, or are starting a fresh project.

#### Load Point sampled map

This will load a map layer with points to which the data from your input maps have already been copied. Consequently, it will skip assigning an area of interest, resolution and input maps. See [running the plugin].

#### Load basemap

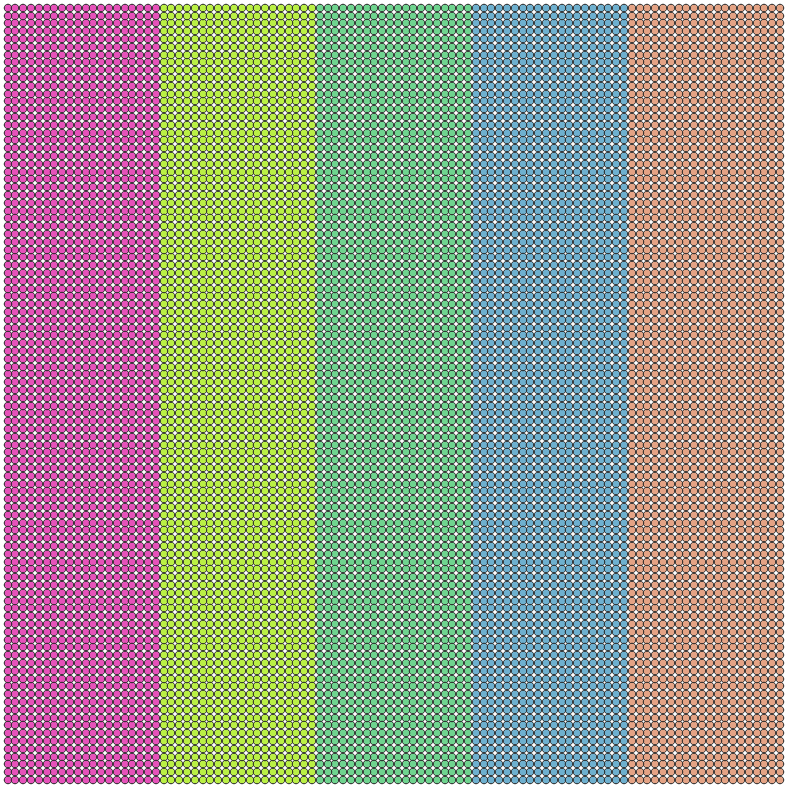
This will load a map layer with points to which the data from your input maps have already been copied, and the [[base group](#_Designate_as_base)] rules applied. Consequently, it will skip assigning an area of interest, resolution and input maps. See [[running the plugin](#_Running_the_plugin/)].

### Map Style

#### Simple

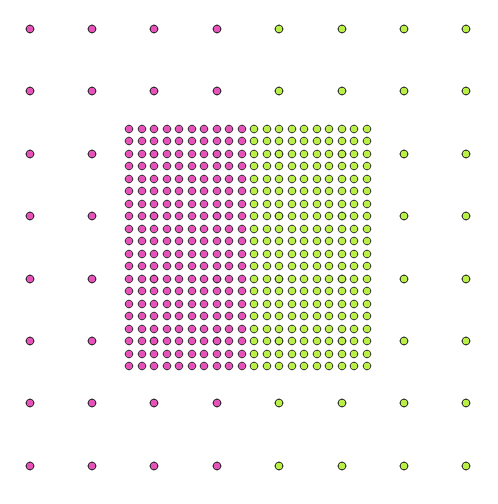
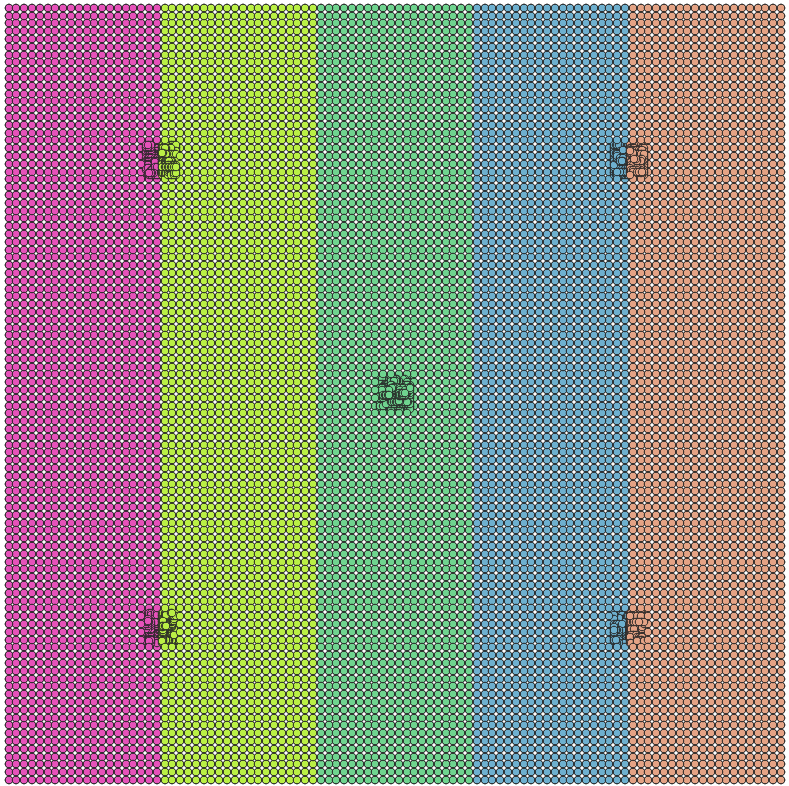
If a map is created, a “simple” map means that the map that will be created will be a single rectangle of equally spaced points, that all have one size. This is the most suitable when testing, doing thought experiments, excluding the background component, or when the overall resolution of your map is very high.

If a map is imported, it means that the map imported consists of equally spaced points that are representative of equally spaced grid sizes (although the grid nor the overall shape of the map have to be square).



#### Nested

A “nested” map means that the map created will contain one larger, lower resolution square of equally spaced points, and within that square, will contain smaller squares of higher resolution centred around the sampling sites. The outer, lower resolution parts of the map can be considered the “background” component and the smaller, higher resolution parts of the map can be considered the “local” component, for the purposes of simulating the pollen loading and percentages. This is most suitable when high resolution is required, but computing resources or time are limited. See [[notes on determining the size of the nested area](#_Notes_on_input)]



#### Nested/Variable

This option is only available when not creating a map from scratch. See [[Starting Point](#_Starting_Point)].

The imported map is either a nested map as described above, or has points that are in some other capacity not all representative of the same area. A map imported this way should have a “resolution” column that contains the (relative) size of the area the various points represent.

### Area of Interest

This option is only available when creating a map from scratch. See [[Starting Point](#_Starting_Point)].

The area of interest marks the North, East, South and West outer bounds of the square area you want to model the vegetation for. There are three options:

**Calculate from layer**: This will copy the outer extents of a chosen map layer to be the area of interest. The layer must have been loaded into QGIS. Recommended method.

**Map canvas extent**: This will copy the outer extents of your current screen view in QGIS to be the area of interest.

**Draw on canvas**: This will direct you back to the QGIS interface where you can freehand draw a square on the map to be your area of interest.

All input that uses some kind of distance or size on the map use the map units associated with the Coordinate Reference System chosen. (e.g., British National Grid EPSG:27700 uses meters, Arkansas North (ftUS) ESPG:3485 uses feet).

### Resolution (Simple)

This option is only available when creating a map from scratch. See [[Starting Point](#_Starting_Point)].

This is the distance between points on the vector point grid. In case of the [[simple](#_Simple)] map style, this is for the entire grid. In case of a [[nested](#_Nested)] map this is the resolution of the outer (background) grid, and it must be a larger number than Resolution (nested) and Size of Nested Area needs to be dividable by it.

All input that uses some kind of distance or size on the map use the map units associated with the Coordinate Reference System chosen.

### Resolution (Nested)

This option is only available when creating a nested map from scratch. See [[Starting Point](#_Starting_Point)] and [[Map Style](#_Map_Style)].

When creating a nested map, this is the distance between points of the nested parts of the vector point grid. It must be a smaller number than Resolution (simple), and Size of Nested Area must be dividable by it.

All input that uses some kind of distance or size on the map use the map units associated with the Coordinate Reference System chosen.

### Size of Nested Area

This option is only available when creating a nested map from scratch. See [[Starting Point](#_Starting_Point)] and [[Map Style](#_Map_Style)].

This is the length of the sides of the square of the nested area(s) around sampling sites. See [[notes on determining the size of the nested area](#_Notes_on_input)]. The number must be a multiple of both Resolution (Simple) and Resolution (nested).

All input that uses some kind of distance or size on the map use the map units associated with the Coordinate Reference System chosen.

### Available fields/bands:

This option is only available when creating a map from scratch. See [[Starting Point](#_Starting_Point)].

**Vector**: The left two tables show the vector polygon map layers and their fields that are loaded into QGIS and set to visible. These are the environmental variables that will constrain the rules of the MSA. The upper table are all those that are currently available in QGIS. The lower are those that you have selected from the upper. Select fields by clicking, or clicking and dragging in the upper table. Deselect by doing the same again. All fields that are in the lower table will be available later on when [[creating rules](#_Add_new_rule)].

**Raster:** The right two tables show the raster map layers and their available colour bands. Selecting otherwise works the same as with Vector layers.

## Input – Vegetation Input

### Import HUMPOL handbag (optional):

This option exists for backwards compatibility. If you have previously done a project using the HUMPOL suite and would like to recreate it in MSA\_QGIS, this will load the information of the taxa, vegetation communities, and sampling points [nyi] from the HUMPOL handbag into MSA\_QGIS. Please note that only the information from the handbag is **not** enough to run the model.

### Taxa

The table contains data for the taxa that will be used in the modelling. [It is currently configured for only the PSS model!]. Single click in the table to select a single taxon, or click and drag to select multiple. Click “remove” underneath the table to delete selected taxa. Click “add new” underneath the table to add a new taxon. This will open a popup asking for information on the taxa you want to add. “Short name” or Code should be unique and kept short, but comprehensible, and without spaces. The programme is generally not case-sensitive, so make sure the names are unique. “Full name” is for your own (and any collaborators’) reference and will not be used in running the MSA, so be as complete as necessary. “Fall Speed” is the fall speed as used in the PSS model. “Relative Pollen Productivity” is the (estimated) Relative Pollen Productivity as used in the PSS model.

### Communities

This table contains the data for the vegetation communities that will be used in the modelling. It gives the name of the community, followed by the percentage of each taxon. Taxa in this table must correspond to the taxa in the taxa table Empty cells are the same as zero (0). The configuration of the table is the same as the taxa table. “Add new” will open a popup asking for information on the vegetation community you want to add. “Name” is the name of the vegetation community. It should be kept short, unique, and not contain spaces, nor be called “Empty” (in that case call it “bare” instead). “add Taxon” will add a row with a dropdown with taxa from the taxa table, and a spin box where you can add the percentage of that taxa within the community. You can add as many taxa as you like, but the total percentage should not exceed 100.

## Input – Rules

The probabilistic and deterministic rules form the basis of the MSA. In contrast to the programming interface of the HUMPOL suite, MSA\_QGIS works with a menu-based interface, which allows it to present the rules in (relatively) human-readable language. A rule may therefore read:

“Rule 4: birch has 70.0% chance to be placed on heath, where Treeline - Treeline is Below, and Geology - GeologyName is Metamorph.”

The rules will only be as readable as the names you assign to your maps and their attributes and features. Help yourself, collaborators and colleagues by making the names short, descriptive, and consistent in language.

The rule number does **not** determine the order in which the rules are executed. This is determined by the [[Rule Tree](#_Input_–_Rule)]. As such the order in which you add the rules is irrelevant.

### Add new rule

Clicking this button will open a popup, wherein you can define a new rule. The rule will be added to the list with the next available number. You can find this number at the bottom left of the popup.

#### Choose a vegetation community

This is a dropdown wherein the vegetation communities defined in [[Communities](#_Communities)] are listed. This is the vegetation community that the rule will place on the map.

#### Choose a rule type

This is a dropdown wherein the types of rules are listed. Currently, there are 4:

**(Re)Place:** This will place the vegetation community within the constraints defined.

**Encroach:** This will increase the size of the area already occupied by the vegetation community by an amount defined in “spread by N distance”. This distance must be at least as big as the resolution, or the rule will not do anything. Note that this rule will not do anything if the community in question has not first been placed.

**Adjacent:** [nyi]

**Extent:** [nyi] see [[importing maps](#_Importing_maps)]

#### Chance

This sets the chance, in percentages, that the rule will apply to a point on the vector point grid within the given constraints. If it is 100%, the rule is deterministic, if it is lower, the rule is probabilistic. A maximum of two decimal points of detail are allowed.

#### Spread by N distance

This option is only available the rule is set to “encroach”, or “adjacent”. See [[Choose a rule type](#_Choose_a_rule)].

This is the distance by which a vegetation community will spread from either itself (encroach) or another vegetation community (adjacent). Keep in mind that this number must be a multiple of your resolution.

All input that uses some kind of distance or size on the map use the map units associated with the Coordinate Reference System chosen.

#### Choose a previous vegetation community

This is the vegetation community that will be replaced by the vegetation community from “choose a vegetation community”. It can be set to 1) “Empty” which means it will only apply to vector grid points that do not already have a vegetation community assigned. 2) a [vegetation community](#_Communities)- By clicking **Add conditional** below this dropdown menu, you can add more vegetation communities for this option - or 3) By checking the “Any” radiobutton, which means it can replace all of the above.

#### Choose an environmental variable

This is an environmental variable that will constrain where the rule will be applied. The environmental variables in the dropdown are the maps you selected in [[Currently Selected fields/bands](#_Available_fields/bands:)]. If you leave it set to “Empty”, there will be no environmental constraints to the rule. When you choose an environmental variable, there are two options:

**Category:** Your environmental variable is categorical, and a dropdown will appear with the categories available, equivalent to the unique entries in the attributes of the attribute table of the associated map.

**Range minimum-maximum:** Your environmental variable is numerical, and two spinboxes will appear. The left is for the minimum value of the range you would like the rule to apply to, and the right is for the maximum value. Values for range are always inclusive, both for the minimum and maximum value. If your environmental variable has a categories that are expressed in numbers, this may appear instead of the categorical dropdown. Simply insert the same minimum and maximum the select a single category (for example min=1.0 – max=1.0 to select category 1).

By clicking **Add conditional** below this dropdown menu, you can add more environmental constraints. These constraints are exclusive, meaning all of the constraints need to apply to a vector point in order for the vegetation community to be placed. If you want one constraint ***or*** another to apply, you need to define second rule and add them to different branches in the [[Rule Tree](#_Input_–_Rule)] instead.

#### Check rule

This will convert the rule into “human language”, so that you can check if the rule actually does what you want it to do before committing. The text will appear on the bottom left side.

### Remove rule

Clicking this button will remove (a) selected rule(s) from the list. Rules that have been removed cannot be retrieved, so make sure to [save] often.

In order to facilitate changing rules after the rule tree was made, numbers do not automatically adjust to be sequential. This means that when deleting a rule, that rule number will not be in use until another rule is made, and if no new rule is made, that rule number will remain unused (e.g., rule number 7 and 9 may be in use, but if rule 8 was removed and no new rule was made, rule 8 will not exist).

### Edit rule

[nyi]

## Input – Rule Tree

The rule tree is a very simple visual programming tool. It allows you to visually set up the scenarios for the MSA, using the rules as defined in [[Input - Rules](#_Input_–_Rules)]. It is set up as a branching structure- at the top you should put the rules that form the basis of your maps, then followed by the rules that will cause the output to vary.

The rule tree consists of blocks, each with a single associated rule that can be chosen and changed from a dropdown menu, which shows all the rules from [[Input - Rules](#_Input_–_Rules)]. Lines between blocks indicate that one rule will follow the other. The flow of the diagram is from top to bottom. A **grey** block is a regular rule block. A **green** block is the block that is currently selected. A **Red** block is part of the base group, see [[base group](#_Designate_as_base)].

### Insert rule below

This inserts a rule block with a connection to the currently selected rule block, or if there are no rule blocks yet, creates the first rule. Attaching a rule block to a rule block that already has a connection will create a branch, and thus, a new scenario. A maximum of 10 branches can be connected to a single rule block.

### Insert rule in series

This inserts a set of rules that branch off and then come back to the same point. If you have a rule in series already selected, it will add an extra rule block to the series. When you want to make a stepwise change in the vegetation communities, but every rule afterwards will be the same, using a rule in series will make editing the rule tree easier and less error prone.

[images “this” (in series) is equivalent to “this” (branched).

### Designate as base group

This will designate a rule block to be part of the “base group”. It can only be applied to rules at the top of the rule tree, cannot be a branched rule block, and must contain a rule that has a 100% chance to run. A rule designated as base group will turn **red**.

The “base group” is a set of deterministic rules at the start of the rule tree that would always result in the same output, and can thus safely be shared between iterations. This means that they are run only once, even when there are many iterations, which can save a significant amount of runtime.

### Delete branch

This will delete a branch from the selected rule block downward. It is currently not possible to remove a single rule block from between other blocks.

(Before you delete a branch, consider whether just choosing a different rule from the dropdown would be a better solution.)

### View rule list

This will show the list of rules to hold alongside the rule tree for convenience.

## Input – Pollen input

### Sample sites

“Add sampling site” opens a popup where you can add a new sampling location. Pollen loadings and counts will be generated for that site. Note that the x and y coordinates should correspond to the same Coordinate Reference System as the rest of your project. The sampling site will automatically be fitted to the nearest vector point on the vector point grid, so the coordinates do not have to fit perfectly with the resolution. Do be aware that at low resolutions this means that your sampling site may be moved away from its original location quite a lot, see [[notes on input](#_Notes_on_input)].

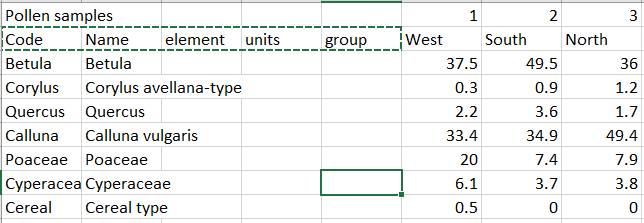
If you are using the HUMPOL mixed-basin model or similar, you can indicate here whether the sampling site in question was in a lake.

“Remove sampling site” removes the selected sampling site.

### Pollen percentage file paths

“Import pollen percentage file” will open a popup with a file explorer. Here you can choose a .csv file that contains your pollen percentages. You need to give a file per site, but they can be the same file.

The file should be in the configuration of a tilia file. This means it should first have one row with some kind of indication of the type of file, and then empty columns until the sample sites column numbers. In the second row it should have five columns called “Code”, “Name”, “Element”, “Units”, “Group”. The sixth and further columns should be the names of the sampling sites. These should correspond with sampling sites previously entered in [[sample sites](#_Sample_sites)]. The rest of the columns should then be filled with the pollen data. Column “code” should correspond with the “taxon codes” in the [[taxa](#_Taxa)] table and column “name” should similarly correspond with “Full Name”. Columns “Element”, “Units” and “Group” are ignored and may be filled with anything. The columns with sampling sites as the names should contain the pollen percentages for the taxa in columns “Code” and “Name”.



### Excerpt pollen percentages

MSA\_QGIS will automatically scan the given pollen percentage file for the site name and taxon codes. Check the tables in the various tabs to see whether the file has been read correctly.

## Input – Model Parameters

### Model for distance weighting.

You can choose from this dropdown which model you want to use. This will determine which parameters will be visible in the interface.

### Fit calculation method

You can choose from this dropdown with fit calculation method you want to use.

**Square chord distance:** This is the default version. See [Resources].

Others [nyi]

### Desired fit per site

Enter the cut-off value for what is considered an “acceptable” map for each single site. Appropriate values will depend on the fit calculation method.

### Desired fit combined

Enter the cut-off value for what is considered an acceptable map, based on the cumulative fit of all sites together. This makes sure that the map is at least somewhat appropriate for all sites, rather than good for one and bad for another. Appropriate values will depend on the fit calculation method.

### Data to keep

**Keep fit maps and simulated percentages**: This option will cause the plugin to only save maps and their associated simulated pollen percentages per site for maps that have met the desired fit per site and combined. Recommended if the non-fitted maps are not of interest as it will save space.

**Keep all maps and simulated percentages:** This option will cause the plugin to save all maps and their associated simulated pollen percentages per site. Recommended especially for exploratory runs, to determine which scenarios in the Rule Tree can be excluded.

**Keep all maps, simulated percentages and simulated pollen loadings:** This option will cause the plugin to save all maps and their associated simulated pollen percentages per site, as well as the pollen loadings per site. This is only recommended for testing newly implemented models to compare the output of the programme with the expected output. MSA\_QGIS Pollen loadings are relative values that are not comparable between most runs.

## Model – HUMPOL mire model

A version of the Prentice-Sutton model that is suitable for use with the MSA. It can optionally be extended it with a wind rose. See [resources] for discussions of the variables.

## Model – HUMPOL mixed-basin model [nyi]

A version of the Prentice-Sugita-Sutton model that is suitable for use with the MSA. See [resources]. It can optionally be extended with a wind rose. See [resources] for discussions of the variables.

## Model – 1/d [nyi]

## Model – 1/d2 [nyi]

## Model – Use Lookup Table [nyi]

Instead of importing a model directly, a .csv file with a lookup table can be imported.

## Input – Metadata

This data is relevant for when you want to share your save files. In keeping with the open source spirit of the software, please make sure to fill this in for any runs that will be published or otherwise made available to others.

### Author(s) and Project name

Speak for themselves.

### Description

Wherever relevant, include any considerations made for how the model was set up, and sources and publications used to generate the input.

### Project changelog

This is a useful feature for when you are sharing files back and forth between authors. Changes can be described here, and will be automatically assigned a date and time so that it’s easy to keep track of versions.

### Notes

Use this however you like.

## Input – Checklist

The MSA requires quite a number of inputs, and it’s easy to forget something. The checklist will stop you from starting a run with incomplete input, so that you don’t accidentally waste minutes (or sometimes even days!) of your time on an ultimately unsuccessful run. It will not, however, stop you from making mistakes! Check your input for missing or extra zeros and decimals, typos and other human mistakes before running, do trial runs with low iterations and/or resolutions to check the output, and don’t forget to [save]!

The checklist is divided into four columns, with requirements for the different run types (see [running the plugin]).

## Input – Order of Operations

It is recommended to fill in information going from left to right in the tabs, and top to bottom per tab. While things may be edited and changed after being added, please be aware that most things do not get automatically adjusted. So, for example, if you have an encroach rule that was written with a resolution of 20 in mind, and you change the resolution to 50, this rule will likely no longer work as intended, or if you remove rule 9 and replace it with another rule 9 in the Rules, this will not automatically change the rule in the Rule Tree. So, when you edit something, make sure you think of what else may need to be changed.

## Notes on input

Notes on any input

Behind the screens, MSA\_QGIS makes use of SQLite, which has its own particular language. As such there are some words that need to be avoided at any time for any of the input. Using words like these (such as SELECT, DROP TABLE, DELETE) as (part of) your input (such as a name or table header of one of your input maps, or a name of a taxon) can cause the run to crash, or the output to be unusable. (It will not cause your computer to crash, and cannot alter your input files.). This is called a SQL injection. Since all the databases are local to the device, doing something like this will not cause any great harm.

* Please check your inputs with this list in the Sqlite documentation: <https://sqlite.org/lang.html>

Additionally, some names are already in use as default names within MSA\_QGIS and these should similarly be avoided, or else will cause the run to crash or output to be unusable. These are:

* Basemap
* Chance\_to\_happen
* Direction
* Dist\_dir
* Distance
* Empty\_basemap
* Fall\_speed
* Full\_name
* Geom\_X
* Geom\_y
* Iteration
* Like\_thres\_cumul
* Like\_thres\_sites
* Likelihood\_cumul
* Likelihood\_met
* Maps
* Msa\_id
* PollenLookup
* Pseudo\_id
* Pseudo\_points
* RelPP
* Sample\_is\_lake
* Sample\_x
* Sample\_y
* Sampling\_sites
* Site\_name
* Snapped\_x
* Snapped\_y
* Taxa
* Taxon\_code
* Taxon\_code
* Taxon\_percentage
* Temp
* Veg\_com
* Vegcom
* Vegcom\_percent
* Windrose
* Windrose\_weight

Additionally, you should not give any input with the following syntax

* Percent\_[any of the vegetation communities you also gave as input] (for example Percent\_bare)
* Likelihood\_[any of the site names you also gave as input] (for example likelihood\_midden1)
* [any of the taxon names you also gave as input]\_DW (for example poaceae\_DW)
* [anything]\_idx or idx\_[anything]

You should also not repeat any names of columns, attributes, map names, etc. This includes all overlap between these (so you should not have a feature “basin” that can also have an associated value of “basin”.)

notes on determining the size of the nested area.

The size of the nested area should be primarily defined by (an estimate for) the relevant source area of pollen (or a similar measure). See [[resources](#_Resources)]. However, due to the manner in which the outer grid is defined, and the fact that multiple sampling sites are possible, sampling sites may not be situated perfectly on a point of the grid. If the sampling point falls perfectly within the grid, the nested grid can be the size of the relevant source area of pollen. If the sampling point does not fall within this grid, the maximum deviation of the sampling point from the grid is half the resolution of the outer grid. Therefore (**RSAP + resolution of outer grid)** is always a safe value to use for the size of the nested area. Having too high a resolution in (part of) the area that falls outside of the RSAP is not a problem, but having too low a resolution within the RSAP is, so play it safe.

Notes on determining the resolution

For the most detailed possible vegetation map, the resolution would need to be at the level of a single plant, circa 1-10 cm. However, while this would undoubtedly result in very beautiful and detailed vegetation maps this is not a level of detail that the pollen data can realistically represent, and so would misrepresent to the final audience the level of detail that the model has achieved (not to mention that it would take very long to run!). Pollen data has been shown to be sensitive to more general vegetation units (source) and the amount and sizes of the patches of the vegetation units (patchiness)(source). This is relevant especially close to the sampling site. The resolution should therefore be equal to or smaller than the smallest “patch” of vegetation. Seeing as the modelling is generally done about past scenarios, this patchiness can be hard to predetermine. It is generally safe to use a resolution of 1-10 meters in nested areas (approximately the size of 1 tree), whereas in background areas a resolution of 50-200 meters would probably be sufficient (a small stand of trees, a small agricultural field). Don’t forget to adjust these numbers to the appropriate map unit, see [[Setting the Coordinate Reference System](#_Setting_the_Coordinate)].

Note on placing sampling points at exact distances on the grid

Sampling points will be snapped to the nearest vector point to it. In most cases, there is only one point that is “the closest” and this will happen without error. The exception to this is when a sampling point is precisely between two (or four) vector grid points. For example, in a grid with a resolution of 50, there is a vector point at x=125 and one at x=175, and a sampling point at x=150. (This is especially likely to happen when the dimensions of the grid were determined based on the location of a sampling point). In such a case, the point that is snapped to will be random, although it is consistent between runs. In rare cases, this may mean that the sampling point will be within the wrong vegetation community (for example, when it was supposed to be in water, but is now in the grass next to the water) which can significantly affect the resulting pollen percentages generated for that site. In cases like these, you can change the x and y input for the sampling point by less than half the resolution (or half of the nested resolution if the map is nested) to “force” it to snap to the correct point. Be sure to add a note of this to the metadata if you do.

## Save

All of the input for MSA-QGIS can saved, and these files should ideally be provided openly with any work or publication that has used MSA-QGIS, to aid reproducibility. Saving multiple files together will put them in the same directory. Ideally, the files required for loading ([input state](#_Input_state_(.csv)), [rule list,](#_Rule_list_(.pkl)) [rule tree](#_Rule_tree_(.pkl))) should be kept together in a clearly named directory, and the names of the files themselves not changed. Ideally, the [input state](#_Input_state_(.csv)), [rule list,](#_Rule_list_(.pkl)) [rule tree](#_Rule_tree_(.pkl)) and [image of the rule tree](#_Image_of_Rule) should be provided with publications.

#### Input state (.csv)

This produces a .csv file called “inputstate.csv” in the chosen directory.

This contains all of the input except for the Rule Tree, in an easily accessible format that can be read without having to load the file into the plugin. The list of rules is included for completeness, but cannot be loaded from this file, and should instead be loaded from Rule\_list.pkl.

Keep this file together in the same directory with the [rule list](#_Rule_list_(.pkl)) and [rule tree](#_Rule_tree_(.pkl)) of the same project.

#### Rule list (.pkl)

This produces a .pkl (pickled python) file called “Rule\_list.pkl” in the chosen directory. This contains all of the information needed to load the rule list. The rule list will load, but not be functional without also loading “inputstate.csv”. Pickled files are not as readily open-able as .csv files, so the rules are included in “inputstate.csv” as well so people can examine the input without the plugin.

Keep this file together in the same directory with the [input state](#_Input_state_(.csv)) and [rule tree](#_Rule_tree_(.pkl)) of the same project.

#### Rule tree (.pkl)

This produced a .pkl (pickled python) file called “Rule\_Tree.pkl” in the chosen directory. This contains all of the information needed to load the rule tree. The rule tree will load, but not be functional without also loading “rule\_list.pkl” and “inputstate.csv”. Pickled files are not as readily open-able as .csv files, so the rule tree can also be saved as a .jpg image so people can examine the input without the plugin.

Keep this file together in the same directory with the [input state](#_Input_state_(.csv)) and [rule list](#_Rule_list_(.pkl)) of the same project.

#### Image of Rule Tree (.jpg)

This produces a .jpg file called “ruletreeimage.jpg”. This is not a file that can be loaded, but simply an image representation of the rule tree. See above.

#### HUMPOL file (.hum)

[nyi] This is a simple text file that is compatible with the previous software for the MSA, the HUMPOL suite. It contains the information on the taxa, vegetation communities, sampling sites, and windrose. This is not the complete input for the HUMPOL suite, but makes comparison a little easier and less error-prone.

## Load

In order to load a previously made, or shared, project, select the directory that contains the file you have checked (see [[Save](#_Save)]). You will not be able to select the file directly, only the directory that contains it. If multiple files are selected for loading, they must be in the same directory.

## Running the plugin/ Output

When one or more columns of the checklist has been completely checked, it becomes possible to click “OK” at the bottom of the MSA QGIS screen. This will open a popup called the “run dialog”. There are four options for types of runs:

#### Point Sampling (Create map with point sampled environmental variables)

**This option should not take much more than an hour to run (and more likely much less). Time taken increases with resolution.**

This option creates the grid of equally spaced, possibly nested, vector points as defined in the Spatial and Environmental Input-tab. This option will not be available when the starting point is not [[create a map from scratch](#_Create_a_map)].

The created file is called “pointsampled\_basemap.sqlite”, and is a SQLite database that contains a single table called “empty\_basemap”. This file can be loaded as a [[starting point](#_Load_Point_sampled)]. The table has 6+ columns, which are:

**Msa\_id:** Unique ID of the vector point.

**Geom\_x:** Easting/x-coordinate of the vector point, in the CRS used for the project.

**Geom\_y:** Northing/y-coordinate of the vector point, in the CRS used for the project.

**Veg\_com:** Vegetation community assigned to the vector point. In the case of running Point Sampling, this will contain only “Empty” as a value, since the MSA has not yet been run.

**Chance\_to\_happen:** Only used in the background while running a thought experiment or full reconstruction, this column needs to be available for loading, but can otherwise be ignored. Values are 0 or 1. In the case of Point sampling, all values should be 0.

**Other columns:** All of the other columns will contain the names of the fields and bands that were selected in [[Available fields/bands](#_Available_fields/bands:)], with the value corresponding to the value of that field/band at the location of the vector point.

#### Basemap (Create basemap with only base rules from rule tree)

**This option should not take much more than an hour to run (and more likely much less). Time taken increases mainly with resolution.**

This option applies the rules of the [[base group](#_Designate_as_base)], either to a given point sampled map, or to a newly created vector point map. The option will not be available when the starting point is [[Load basemap](#_Load_basemap)].

The created file is called “output\_basemap.sqlite”, and is a SQLite database that contains 1 table called “basemap”. The “basemap” table is the same as the “empty\_basemap” table, except the **veg\_com** column has now been filled according to the rules in the base group. This file can be loaded as a [[starting point](#_Load_Point_sampled)].

This will also create the “pointsamples\_basemap.sqlite” file described in [[Point Sampling](#_Point_Sampling_(Create)] when creating a map from scratch.

#### MSA thought experiment (Run MSA thought experiment (without fit))

**This option can take a significant amount of time to run. Time taken increases mainly with resolution and number of iterations.**

This option runs the MSA without calculating fit compared to actual pollen counts, such as would be useful when running thought experiments.

Upon finishing, a popup will open that shows how many maps were made and the fit results. In the case of the thought experiment, the fit results will be empty. From this menu, the resulting maps can be loaded directly into QGIS, if desired, but these files can also be loaded into QGIS afterwards from the output folder.

The output files are as follows:

For each map created, there will be a .csv file, with a naming convention of scenario\_iteration. (For example, 1000\_1 or 11000110000\_10). These will be the same as “basemap” and “empty\_basemap” except **veg\_com** has now been filled with vegetation communities according to all of the rules in the rule tree.

“Simulated\_likelihood\_and\_landscape.csv” is a .csv file that contains 7+ columns, which are:

**Map\_id:** The name of the generated map, which corresponds with the previously mentioned .csv files.

**Iteration:** The iteration for which the map was made, which will be the same as the number after the underscore in map\_id.

**Likelihood\_met:** A boolean indicating whether both the cumulative and the site-specific likelihood thresholds were met. Values can be yes, no or null. In the case of a thought experiment, it will be null as no fit was calculated.

**Like\_thres\_sites:** A copy of [[desired fit per site](#_Desired_fit_per)].

**Like\_thres\_cumul:** A copy of [[desired fit combined](#_Desired_fit_combined)].

**Likelihood\_cumul:** Contains the value calculated for fit compared to the actual pollen counts cumulative for all sites. In the case of a thought experiment, this column will be empty.

**Likelihood columns:** Columns named with a naming convention of likelihood\_site (e.g. likelihood\_samplingpointnorth) that contains the value calculated for fit compared to the actual pollen counts for the specific site. In the case of a thought experiment, these columns will be empty.

**Percentage columns:** Columns named with a naming convention of percent\_vegetation community (e.g. percent\_woodland). This will contain the percentage of coverage per vegetation community for the map in question. This will be a value between 0 and 100. Note that this value can be helpful, but does not tell the entire story, and the spatial distribution of the vegetation communities should be examined as well.

“Simulated\_pollen\_output.csv” is a .csv file that contains 3+ columns, which are:

**Map\_id:** Same as for previous

**Site\_name**: Name of the site for which the percentages were calculated.

**Simulated percenage columns:** Columns names with a naming convention of sim\_taxon\_percent (e.g. sim\_birch\_percent). These contain the simulated pollen percentages for the sampling site, per scenario, per iteration.

“MSA\_output.sqlite” is a .sqlite database that contains many tables, including copies of the output .csv files described above. Since .sqlite files are small and fairly timeless, this file is most suitable for archiving if you don’t want to keep the whole host of .csv files. In addition, this is where you go to check if any mistakes were made by the software. As such, it is mostly relevant for developers on MSA-QGIS, rather than the end users. Most data analysis concerned with the output of the MSA will likely only involve the .csv files. However, should you find it useful, the file will have the following tables:

**Basemap:** This is the basemap, which is the same as the basemap in “output\_basemap.sqlite”. If your run had no basemap, then it will be the same as the empty\_basemap from “pointsamples\_basemap.sqlite”.

**Dist\_dir:** This contains the distance and cardinal direction for every point on the vector grid to every site.

**Maps:** This is the equivalent of the “Simulated\_likelihood\_and\_landscape.csv” file.

**PollenLookup:** This is the lookup table for distance weighted plant abundance as calculated with the chosen dispersal and deposition model.

**Pseudo\_points:** This contains information on the points that are created around the sampling point to use instead of the point that is snapped to by the sampling point.

**Sampling\_sites:** This contains the information the user has given about the sampling points, as well as information on which point on the vector grid they have snapped to.

**Taxa:** This is a direct copy of the input the user has given about the taxa.

**Veg\_com:** This is a copy of the input the user has given for the vegetation communities, presented in a way that makes it suitable as a sql lookup table.

**Vegcom\_list**: A list of the names of the vegetation communities.

**Pollen loading tables:** Tables with a naming convention of SitenameMapID (e.g. samplingpointnorth10000\_1) contain the pollen loading calculated for each taxon, for each vector point and pseudo point. The pollen loading values are not comparable between different projects.

**Maps**: Tables with a naming convention of scenario\_iteration (also known as the map id) are exactly the same as the map .csv files, with the same names and columns.

**Optional columns:** Depending on which pollen dispersal and deposition model was used, additional tables may appear with data that facilitates these calculations. An example is a “windrose” table that will appear if the HUMPOL mire or HUMPOL mixed basin model was run with windrose weighting.

#### MSA reconstruction (Run MSA reconstruction (with fit))

**This option can take a significant amount of time to run, slightly longer than without calculating fit. Time taken increases mainly with resolution and number of iterations.**

This option runs the full MSA reconstruction, including calculating fit compared to (actual, real life) pollen counts.

Upon finishing, a popup will open that shows how many maps were made and the fit results as compared to the [[desired fit combined](#_Desired_fit_per)] and [[desired fit per site](#_Desired_fit_combined)]. Maps can be opened into QGIS, selecting maps with various fit results, , if desired, but these files can also be loaded into QGIS afterwards from the output folder.

The output files are the same as those described in [[MSA thought experiment](#_MSA_thought_experiment)], with the following differences:

In “Simulated\_likelihood\_and\_landscape.csv”, the likelihood\_met, like\_thres\_sites, like\_thres\_cumul and likelihood columns will now be filled with the appropriate values based on the chosen [[fit calculation method](#_Fit_calculation_method)].

If “Keep fit maps and simulated percentages” was chosen under [[data to keep](#_Data_to_keep)], there will be no .csv files of the maps that did not make the likelihood threshold. These maps will, however, still be available as tables in “MSA\_output.sqlite”.

“MSA\_output.sqlite” will now contain additional tables that are copies of the pollen percentage data given per site by the user in [[Pollen percentage file paths](#_Pollen_percentage_file)]. The name of the table will be the name of the site.

## Notes on Running the plugin/Output

What to do if the MSA run has been/needs to be interrupted.

The MSA thought experiment and MSA reconstruction can take significant amounts of time to run- sometimes up to weeks. While the process does not lock down the computer and you will be able to keep working (except in QGIS) while it runs in the background, there are many situations in which a computer may shut down by itself, or in which a computer needs to be shut down. Fortunately, in most cases this does not mean all is lost. The iterative nature of the programme means that any iteration that managed to finish does not need to be repeated. However, the programme cannot clean up interrupted runs (yet…) and so some data clean-up is needed in order to get the correct files and potentially continue running. This obviously requires some manual data handling, and that can take some active time, whereas simply re-running the entire thing will be happening passively in the background. Therefore, it is probably not wise to do this unless you have a close deadline for when this data needs to be available. Nevertheless, while somewhat tedious, the process is not very complex, and here are the steps to follow in such a case:

1. Determine how many maps are made each iteration.
2. Check in the output folder if the correct number of map .csv files were made for each iteration. Map naming conventions means the map names will always end with an underscore and then the iterations (e.g. \_1 or \_12). If no maps were generated yet, then there is no other option than to simply run the whole thing again. (Note that the programme does not necessarily compute the iterations in order! It may be the case that iteration 3 is complete, but iteration 2 is not).
3. Make note of which iterations did not finish.
4. Remove all of the map .csv files for iterations that did not finish/ did not have the correct number of maps.
5. Download a programme that can read SQLite databases (T.W.B recommends [DBeaver](https://dbeaver.io/)).
6. Extract the simulated\_pollen table to a .csv file. This is the equivalent to the simulated\_pollen\_output.csv file.
7. Remove all rows where map\_id is of an iteration that did not finish.
8. Extract the “maps” table to a .csv file. (This is the equivalent to the simulated\_likelihood\_and\_landscape.csv file).
9. Remove all rows where map\_id and/or iteration is of an iteration that did not finish.
10. Determine how many iterations of your original number did not finish.
11. Load the save files (you absolutely should have saved before running! This will be automated later [nyi]…), set the [[starting point](#_Starting_Point)] to the output\_basemap.sqlite file in the output folder, and set the number of iterations to how many iterations did not finish. Change nothing else.
12. Run the MSA in the same way you did before, with the output going to a different output map (otherwise the previous output will be overwritten).
13. When the run has finished, you will have repeated names for some of the map\_ids, these names (of the map .csv files, and of the rows within the simulated\_pollen\_output.csv and the simulated\_likelihood\_and\_landscape.csv files) will have to be changed manually before combining the files. Make sure the name changes are the same between all of the files and rows.
14. Combine the extracted “simulated pollen” table .csv file with the new “simulated\_pollen\_output.csv” file.
15. Combine the extracted “maps” table .csv file with the new “simulated\_likelihood\_and\_landscape.csv” file.
16. If you want, you can combine the two output folders completely so thatyou’re your data is in a single directory, but this is not necessary. Congratulations, you have rescued your operation.

## Warnings and Errors

### Foreseen Input warnings and errors

These are errors and warnings that occur while you are writing the input. They are generally descriptive and do not indicate any serious problems, but may stop you from completing an action to prevent problems further down the line. They will show up as yellow (warning) or red (error) banners in the QGIS interface. [to be changed to popups[nyi]].

### Foreseen Runtime errors

These are errors that occur after starting the run (you’ve pressed ok in both the main and run dialogs). They are generally more serious and will likely prevent a run from completing.

Foreseen Runtime errors have been anticipated in the code, and if the run needs to stop, the data produced until that point will be clean and files produced will not be broken (although they will likely be incomplete). The error will be stored in the “MSA\_QGIS\_log.txt” file, and can be read in the QGIS log if the log panel is open. The error message should be descriptive enough to fix the problem yourself. If this is not the case, contact the author or [open an issue on GitHub[nyi]] that includes the “MSA\_QGIS\_log.txt” file and a description of what you were doing.

### Unforeseen errors

Unforeseen errors have not been anticipated in the code, and if the run needs to stop, the data produced until that point will not be clean and files produced may be broken. The error will be shown in the QGIS interface as “**Python error:** An error has occurred while executing python code: See message log (Python error) for more details”. Two buttons will open the log panel (if it was not already open) or traceback, both of which will contain the traceback information on what went wrong in the code. Please copy this and send it, along with the “MSA\_QGIS\_Log\_txt” file and a description of what you were doing to the author, or [in an issue on GitHub[nyi]]. You will then have to re-start QGIS before you can make another attempt at running MSA-QGIS.

### Crashes

This is when QGIS (or your computer!) fully stops responding or shuts down (crashes). If this happens before running the plugin, it is most likely a problem with QGIS itself, and should be reported to QGIS rather than the plugin author. If this happens during runtime, it could be either the plugin or QGIS itself. Check “MSA\_QGIS\_log.txt” for errors. QGIS itself may also generate an error report. If you suspect the error was caused by the plugin, please raise an issue on the MSA\_QGIS GitHub [nyi], or contact the authors. If you do not think it was caused by the plugin, you can raise the issue with QGIS, or simply try again.

### List of foreseen errors

[…]

## Credit

### Licence

This manual for MSA\_QGIS is licenced under the GNU Free Documentation License (FDL) © 2023 Thya W.B. van den Berg & M. Jane Bunting

The MSA\_QGIS software is licenced under the GNU General Public License (GPL) version 3. © Thya W.B. van den Berg. You can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.

### Cite MSA\_QGIS

[nyi]

While not required according to the GPL, for the purpose of reusability and reproducibility, and according to academic convention, we recommend that you cite the software and version in publications and data that have used it.

Please cite [thesis] and/or [publication], along with the software itself, version used, and a link to the repository, in whichever citation style your work requires. As nothing is published yet, please cite Van den Berg, T.W.B. & Bunting, M.J. (Unpublished software 2023) MSA in QGIS, vx.x.x.

## Contact

Email w.b.van-den-berg-2020(at)hull.ac.uk, for T.W.B. van den Berg, for any questions. (If no longer available, try thya(at)live.nl. instead).

If Thya is not available, you can email M.J.Bunting(at)hull.ac.uk for Jane Bunting, for questions about the (history of and maths behind) the MSA, and pollen modelling in general.