

Fort Collins Science Center

Tutorial for the Software for Assisted Habitat Modeling (SAHM) package in VisTrails.

By Nick Young

Report Series XXXX–XXXX

U.S. Department of the Interior

U.S. Geological Survey

U.S. Department of the Interior

KEN SALAZAR, Secretary

U.S. Geological Survey

Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 201x  
Revised and reprinted: 201x

For more information on the USGS—the Federal source for science about the Earth,  
its natural and living resources, natural hazards, and the environment—visit   
<http://www.usgs.gov> or call 1–888–ASK–USGS

For an overview of USGS information products, including maps, imagery, and publications,  
visit <http://www.usgs.gov/pubprod>

To order this and other USGS information products, visit <http://store.usgs.gov>

Suggested citation:  
Author1, F.N., Author2, Firstname, 2001, Title of the publication: Place of publication   
(unless it is a corporate entity), Publisher, number or volume, page numbers; information   
on how to obtain if it’s not from the group above.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply   
endorsement by the U.S. Government.

Although this report is in the public domain, permission must be secured from the individual   
copyright owners to reproduce any copyrighted material contained within this report.

**Contents**

SAHM VisTrails Tutorial 1

Introduction 1

Downloading presence data 1

Creating a Workflow 3

Starting VisTrails 3

Input Modules 3

Pre-processing Modules 7

Preliminary model analysis and decision Modules 12

Correlative Model Modules 13

Output Visualization Modules 14

Running a Workflow 16

Running a Maxent Workflow 28

Provenance 29

**Figures**

**No table of figures entries found.**

**Tables**

**No table of figures entries found.**

# SAHM VisTrails Tutorial

# Introduction

The purpose of this tutorial is to get the user familiar with the VisTrails Graphical User Interface (GUI) and more specifically, introduce the features and functionality of the software for assisted habitat modeling (SAHM) package. This tutorial will use an example to demonstrate how to input data, how to add and configure models to build a workflow, how to execute a model, how to modify a workflow and how to view the multiple types of model results. This example follows the work presented in the paper by Morisette et al. (2012) that demonstrates the development of a habitat distribution model of invasive cheatgrass in Rocky Mountain National Park, USA, at two different spatial resolutions using four different niche modeling techniques and three sets of predictor variables This tutorial may be updated periodically as additional features are added to SAHM or bugs are fixed. Please visit <xxx> to download the latest version.

# Downloading presence data

Note that this section describes how and where to download the presence data used in this tutorial. A pre-downloaded version of these data are included with the SAHM download in the file RMNP\_Bromus.csv.

1. For our example, we obtained cheatgrass (Bromus tectorum) presence absence data from the National Institute of Invasive Species Science (NIISS) <http://www.niiss.org.> The National Institute of Invasive Species Science is a consortium of governmental and nongovernmental partners whose aim is to provide reliable information and advanced decision support tools for documenting, understanding, predicting, assessing, and addressing the threat of invasive species in the United States. To download data for this tutorial from NIISS, you will need have to register by selecting ‘register’ at the bottom of the left hand navigation on the home page. Fill in the required fields and click submit. You will need to confirm your registration using your email. Once registered, select ‘By Map’ under ‘Browse Data’ on the left hand navigation. This will bring you to a map with tamarisk the default species displayed. The data for this tutorial can be found in the Colorado project. To access this data, click the ‘Edit’ link at the top of the legend on the right hand side. Click ‘Delete’ for the Tamarisk organism layer and then under ‘Project layers’ click ‘New’. This will bring you to a list of projects. In the search box, type ‘Colorado’ and click ‘search’. The Colorado project will be listed in the top five results, click the name of the Colorado project which will add the project under the project layers to add to the map. Next, click ‘submit’ to view the data associated with this project on the map. Zoom to Rocky Mountain National Park by clicking on the Zoom tool and drawing a box around the northern cluster of points on the map. To download the data in Rocky Mountain National Park, click the ‘Download data to a spreadsheet’ icon and draw a box around the points. This will bring you to a download wizard. On the first page, select ‘Information on organisms’ and then ‘Continue’. On the next page, check the ‘Project’ check box and select the Colorado project. Click ‘Continue’. Leave all the defaults on the next page and click ‘Continue’. In the next page, enter “Cheatgrass\_RMNP” in the name field and choose ‘CSV file’ as the type of file to download. Click ‘Continue’. This will bring you to download page with a progress bar. The download should take less than 5 minutes to complete. Once the download has completed, a link to download the data will appear. Download the data to your computer.

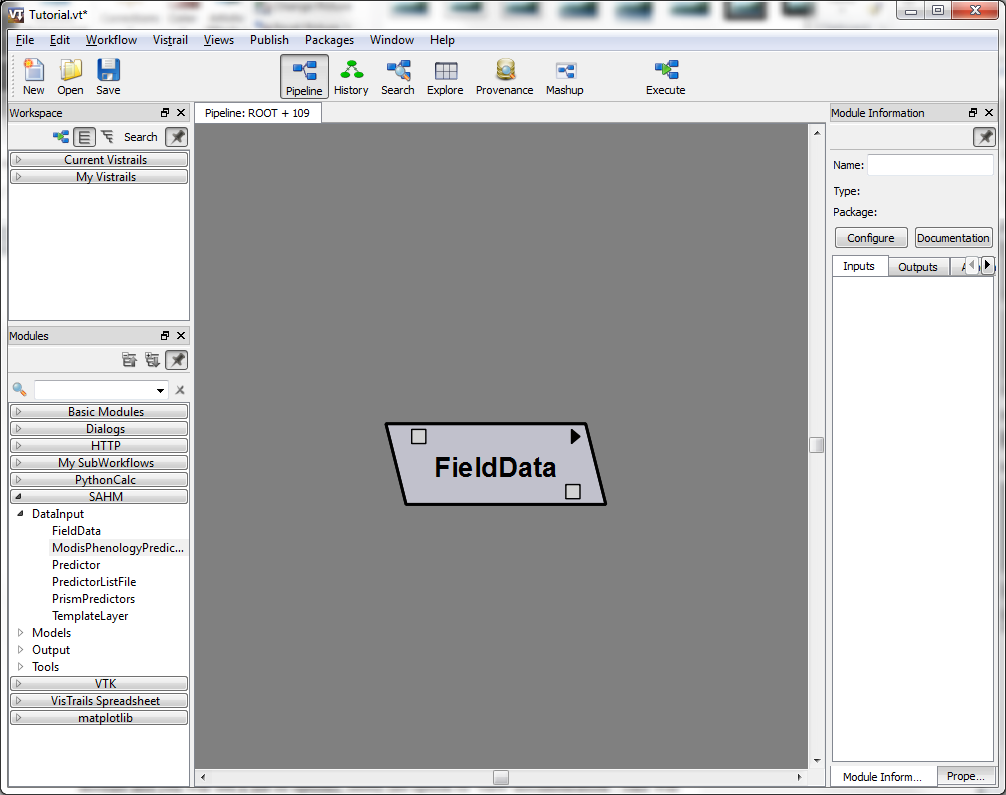
# Creating a Workflow

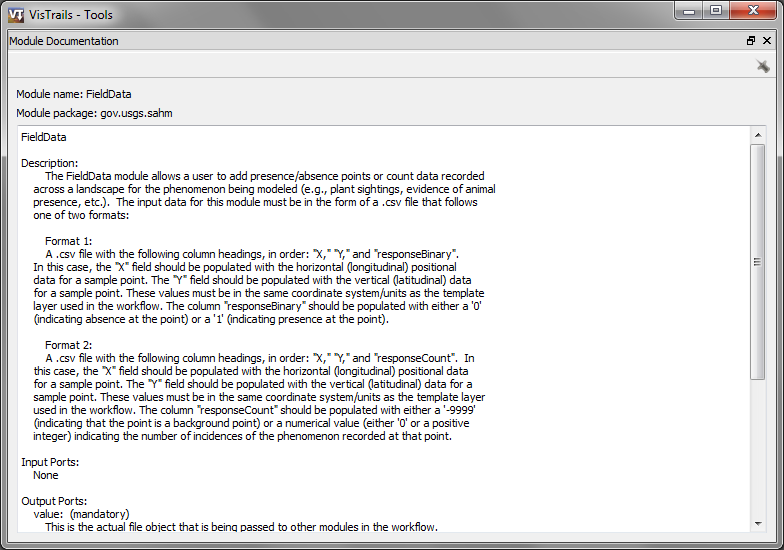
Note that this tutorial assumes you have downloaded and installed all the required components to run VisTrails and SAHM. Please refer to the user documentation with detailed instructions to install the necessary components.

### Starting VisTrails

1. Now that we have downloaded the presence data, we can start building a workflow in VisTrails. Open a new session of VisTrails by navigating to and double-clicking on the VisTrails.exe in the Central\_VisTrails\_x32 folder. This will open the GUI (see figure 1 in the user documentation). The workflow space will be empty.
2. At the top of the GUI, you will see the standard toolbar. you will also see icons that allow you to control various views with Pipeline as the default view showing the workflow canvas. On the left hand side, you can see the workspace and the modules. The workspace will show current an recent VisTrails workflows (files with a .vt extension). The Modules is where you will find packages and their associated modules. In this example, we are only interested in the SAHM package. For a full reference on the VisTrails software and associated modules, please see <http://www.vistrails.org/usersguide/VisTrails_Documentation/VisTrails_Documentation.html>

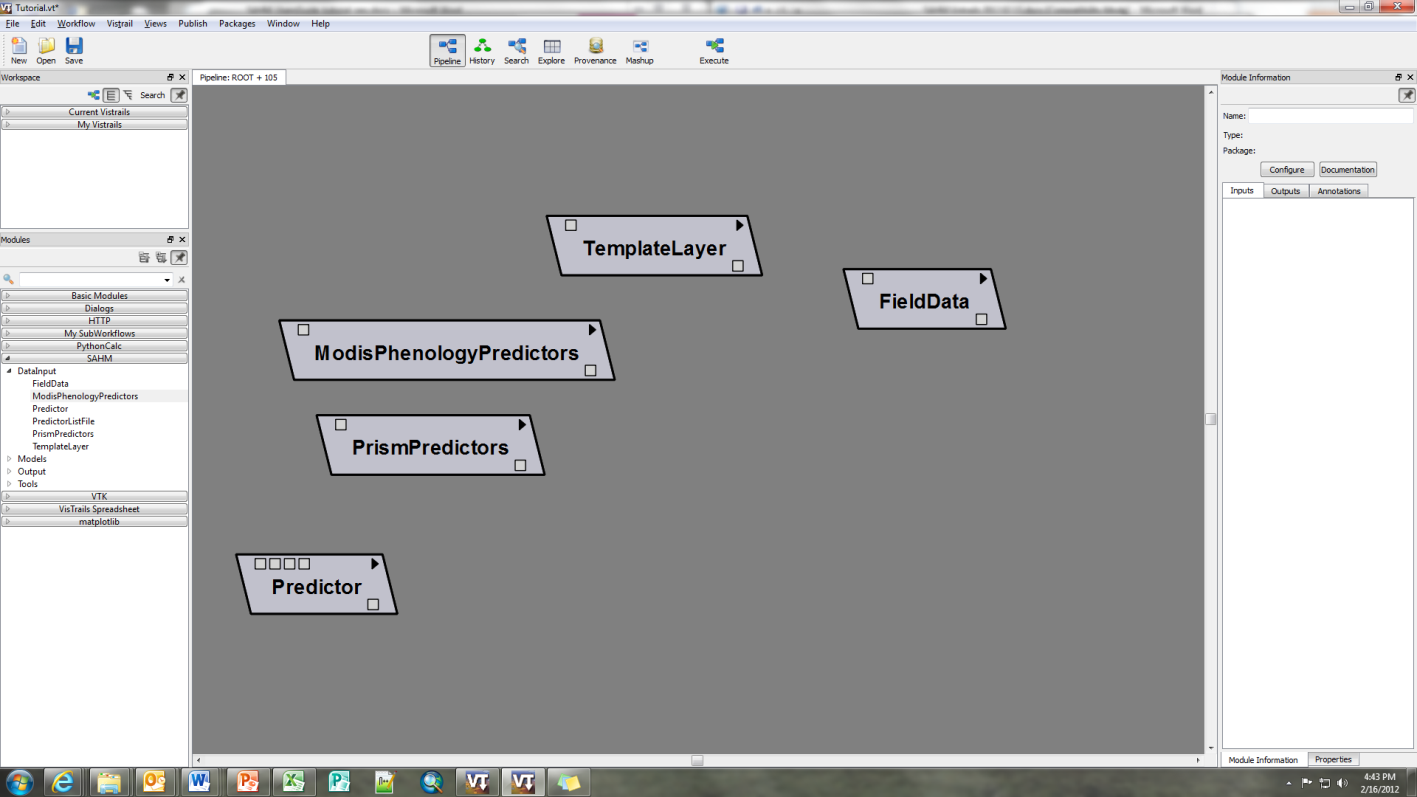
### Input Modules

1. To start building our workflow, we will need to add modules. In the pipeline view of VisTrails, drag in a FieldData Module from the DataInput section. Since this is our first module we’ve added, let’s find out more about this module. To do this, click on the black arrow in the top right corner of the module and you will see a list of options; select the option to ‘view documentation’. This will open a window that gives a description of the module, the file formats needed for inputs, any input and output ports that it may use and common modules it may connect to.



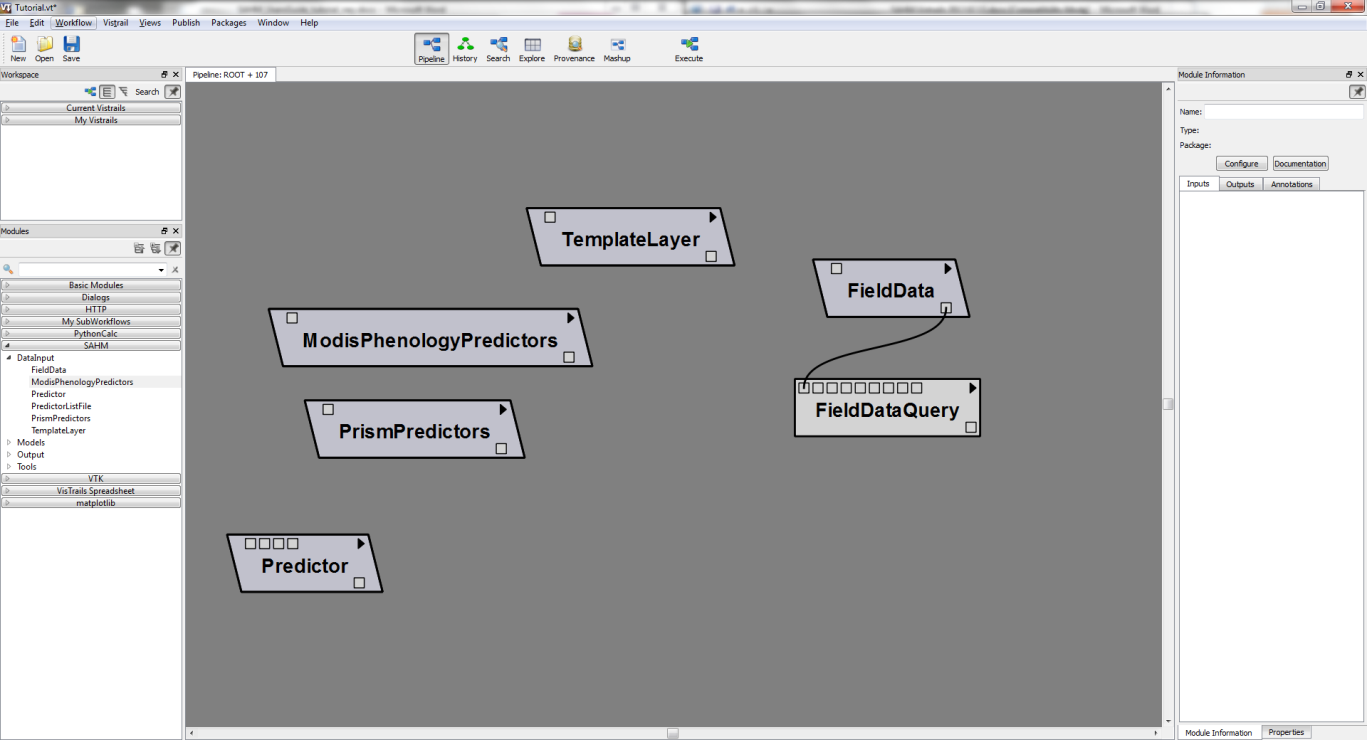
This information can be helpful when first using a module. After reading the module description, look at the Module Information pane on the right and click on the 'value' port on the Inputs tab. This will open a path box below the port name where the path to a csv file that has field data. Click on the folder icon to navigate to the file you downloaded from NIISS, or use the RMNP\_Bromus.csv supplied with the SAHM\_example.

1. Next, we want to add the TemplateLayer module by dragging it from the DataInputs in the SAHM package. The template layer will define the extent and resolution for our model (i.e., all layers will be modified to match the specifications of the template layer). For our example, we will set the template layer to be romoveg800mr which represents the vegetation cover in Rocky Mountain National Park at a 800m resolution. Click on the folder icon to navigate to the file located in examples\SAHM\_example (For grid files, you will need to navigate to the .hrd for the layer see the user documentation for more information).
2. Next, we will bring in predictor layers. There are a variety of methods to add predictor layers to the workflow and you can use multiple methods within the same workflow. The most basic method to add a predictor is to use the Predictor module. The predictor module allows you to add one predictor layer at a time to the workflow. Add the Predictor module to the workspace and set the file path to be removeg100mr. Next, bring in the PrsimPredictors and ModisPhenologyPredictors modules which are specific predictor lists for this tutorial. These modules are pre-defined to include a custom list of predictors (see the Individual Predictors selector modules section in the user documentation for more information on this type of predictor module). When you have a large number of predictors you want to include or want to have the ability to quickly change which predictors to use, you can use the PredictorsListFile module (not included in this tutorial). The PredictorListFile module allows a user to load a .csv file containing a list of rosters for consideration in the modeled analysis. Please refer the module documentation or the user documentation on the details of this module.
3. At this point, all input modules for this example should be added to the workflow canvas

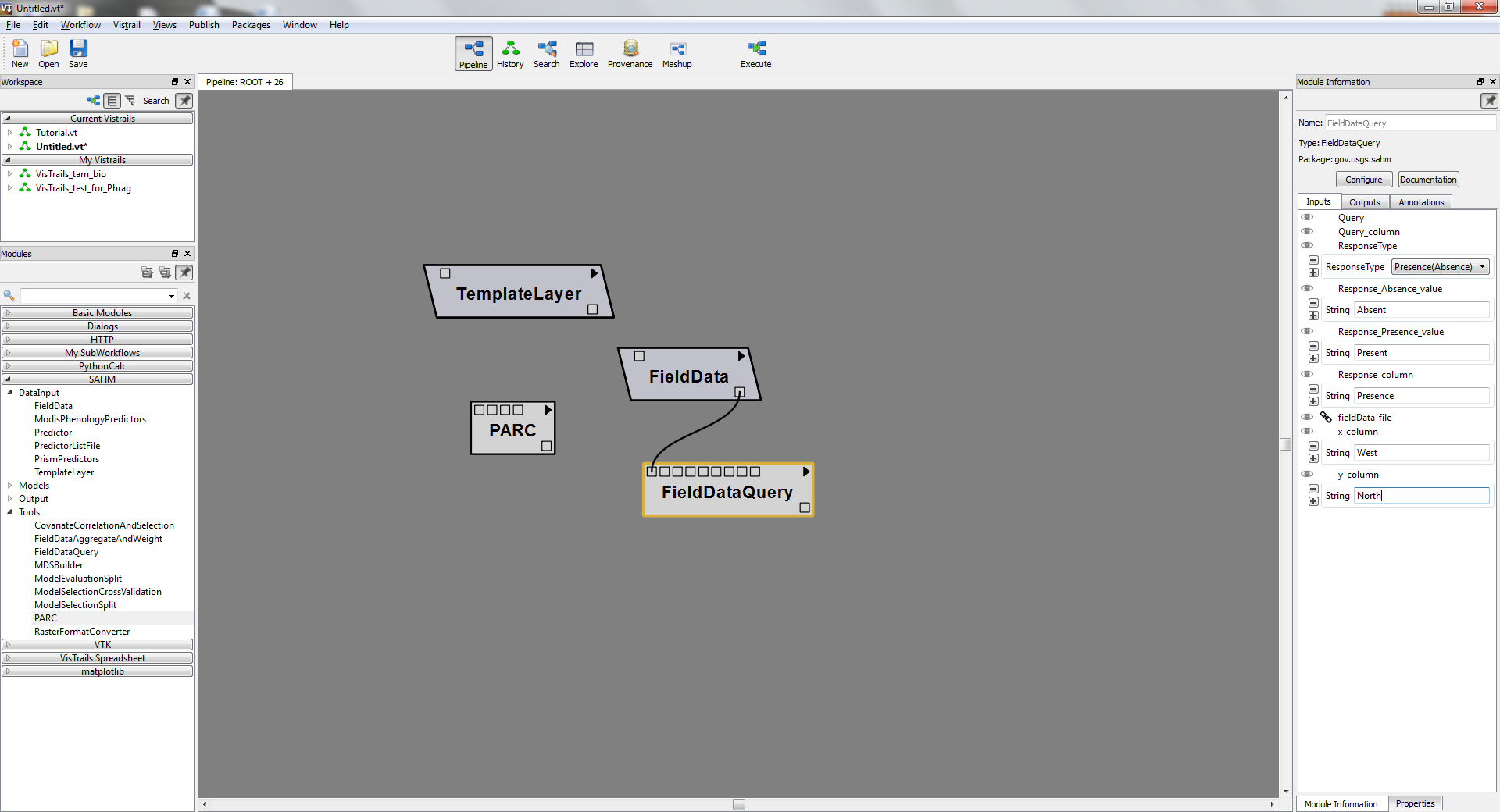


### Pre-processing Modules

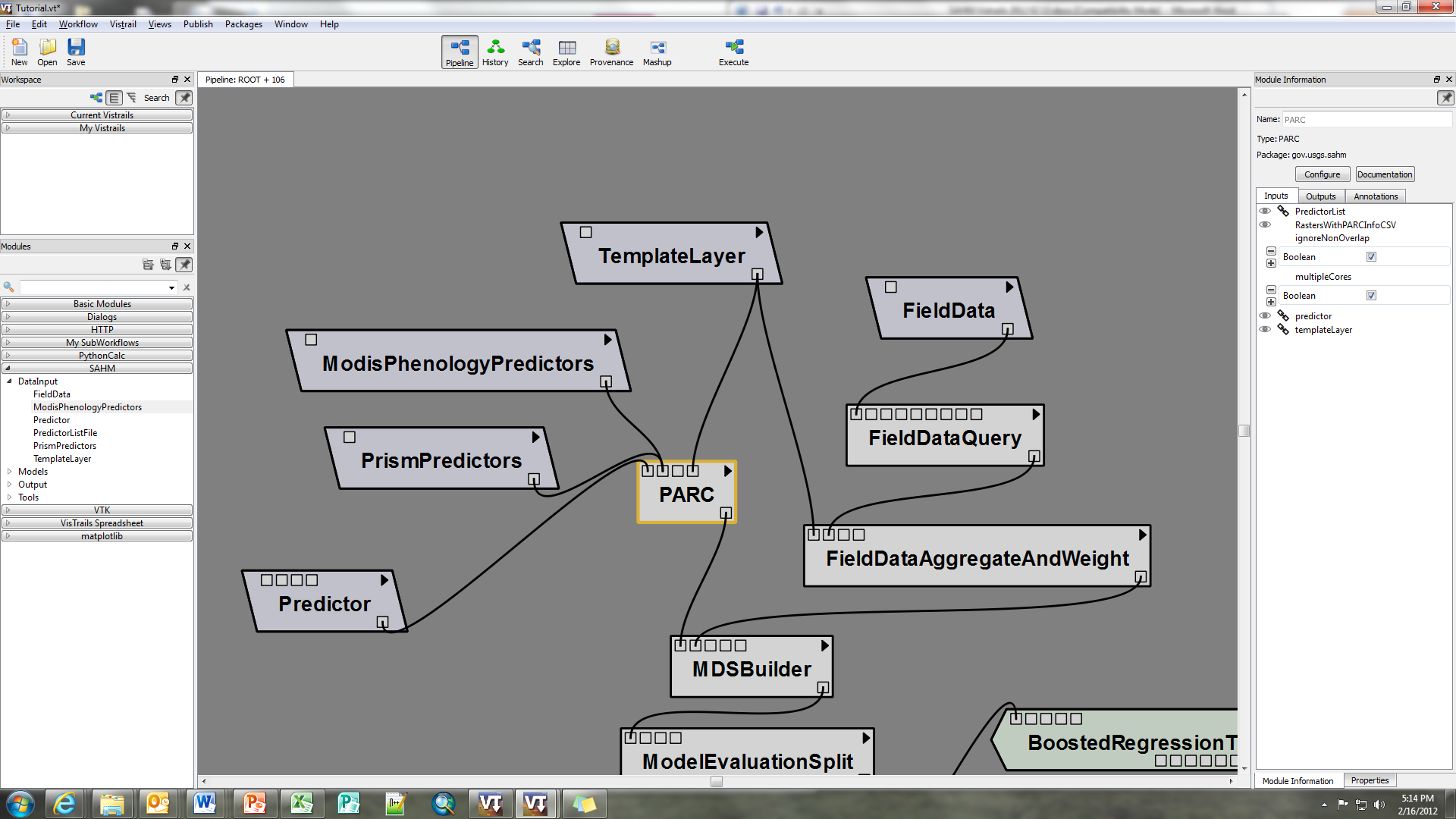
1. Since the file format that NIISS supplies does not match that used by SAHM we must reformat our input using the FieldDataQuery module (Note, instead of using the FieldDataQuery module, we could change our input csv file to match that required format). Drag a FieldDataQuery from the Tools section onto our canvas. Remember, we can always get more information on this module by viewing its documentation as we did with the FieldData module. Connect the output of the FieldData module to the FieldDataQuery module by clicking and dragging from the small output box (the box on the bottom of the module) to an input box (the box[es] on the top of the module) on the FieldDataQuery module. Note that there may be multiple input and/or output boxes on each module.



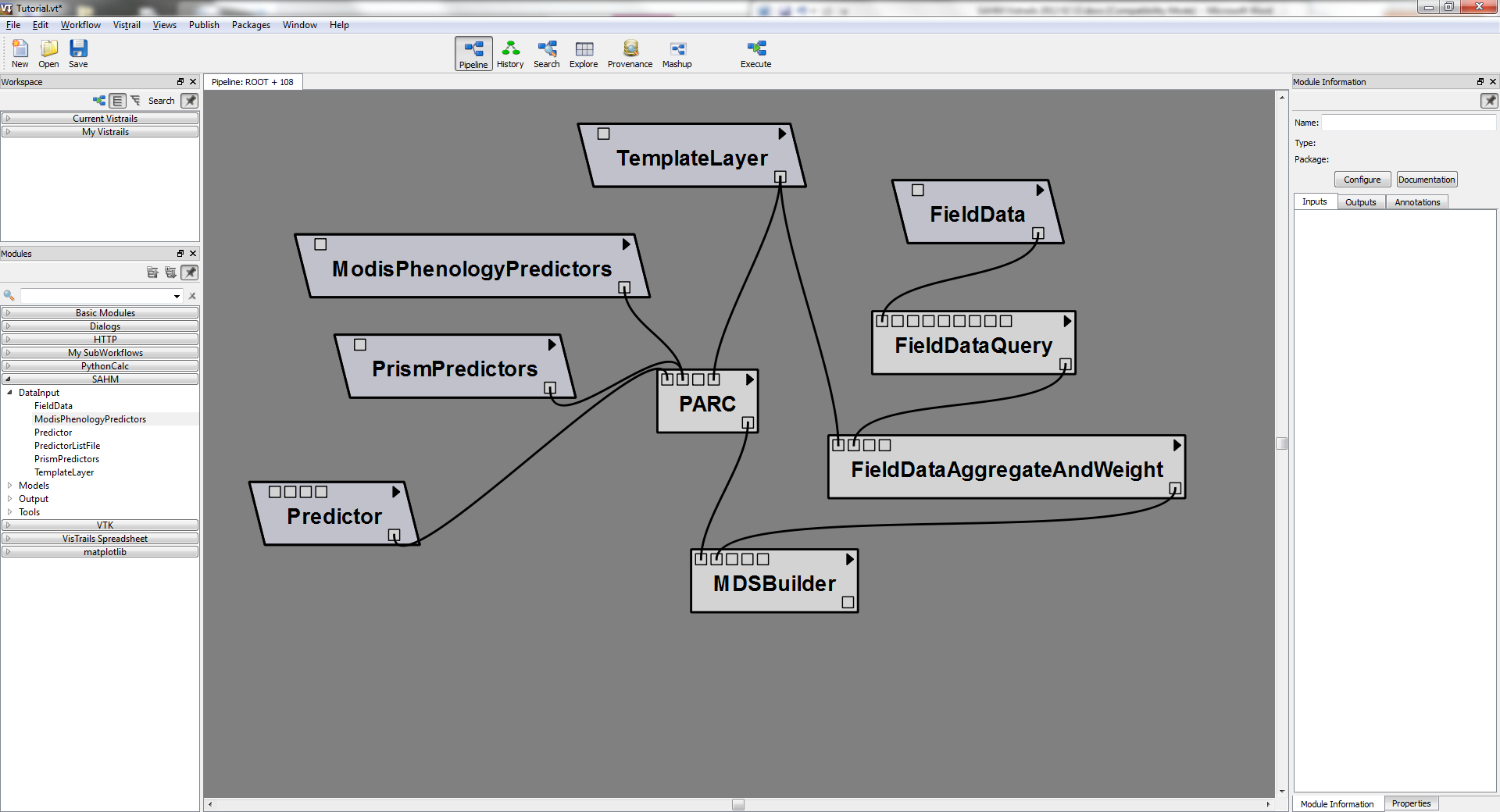
1. We will now set the parameters in the FieldDataQuery module. We know that the response we are interested is presence and absence of cheatgrass, Therefore we will define the ‘ResponseType’ to be Presence(absence) by expanding the responseType parameter and using the drop-down to menu to select the response we are interested in. Next, we will want to define what our absence and presences values based on the field data. If we examine the csv file, we can see that our input field data uses the text ‘Present’ to indicate cheatgrass presence and ‘Absent’ to indicate cheatgrass absence. Enter the text ‘Absent’ and ‘Present’ into Response\_Absence\_value and Response\_Presence\_value respectively. We will also need to define what column has the response variable which is ‘Presence’. Enter this text into the Response\_column field. Finally, we need to define which columns in our input data have the x location and the y location for our response. In our case this will the column titled ‘West’ for the x\_column and the column titled ‘North’ for the y\_column. At this point, the module information window should look like the one shown below.



1. In this example, the spatial reference of the presence locations does not match that of the template layer. Fortunately, this is easily resolved using the FieldDataAggregateAndWeight module. Add this module to the workspace and connect the output box from the FieldDataQuery to the input box an input to the FieldDataAggregateAndWeight module. Also connect the TemplateLayer module to this module. Open this module and expand the FD\_EPSG\_projection field. Enter the number 4326 in the field. This number corresponds to the EPSG code for the datum and projection that the x and y locations in the presence data are in. This will ensure that the location coordinates are re-projected to match the template layer projection. For more information on what this code is and how to find it for another dataset, refer to the documentation for this module. We will also want to change the PointAggregationOrWeithMethod to be ‘Collapse in Pixel’. Again, you can view the documentation on this module to learn more about these fields.
2. Next, add the PARC module to the workflow and connect all the predictor modules and the template model to the input boxes (Note that the connections may automatically shift to specific input boxes depending on the type of input). The PARC module is the workhorse of the pre-processing modules ensuring that all predictor layers match the resolution and extent of the temple layer provided. For this module expand the ‘ingnoreNonOverlap’ field and Check the Boolean box. This will use intersection of all covariates extents (i.e., the area of template extent will be reduce such all covariate layers extents can be completely covered by the new extent). Also expand the multipleCores and check the box to take advantage of multiple processor to speed up model processing. Notice that some of the module fields have a lock symbol next to them. This indicates that they have been defined by other module inputs.

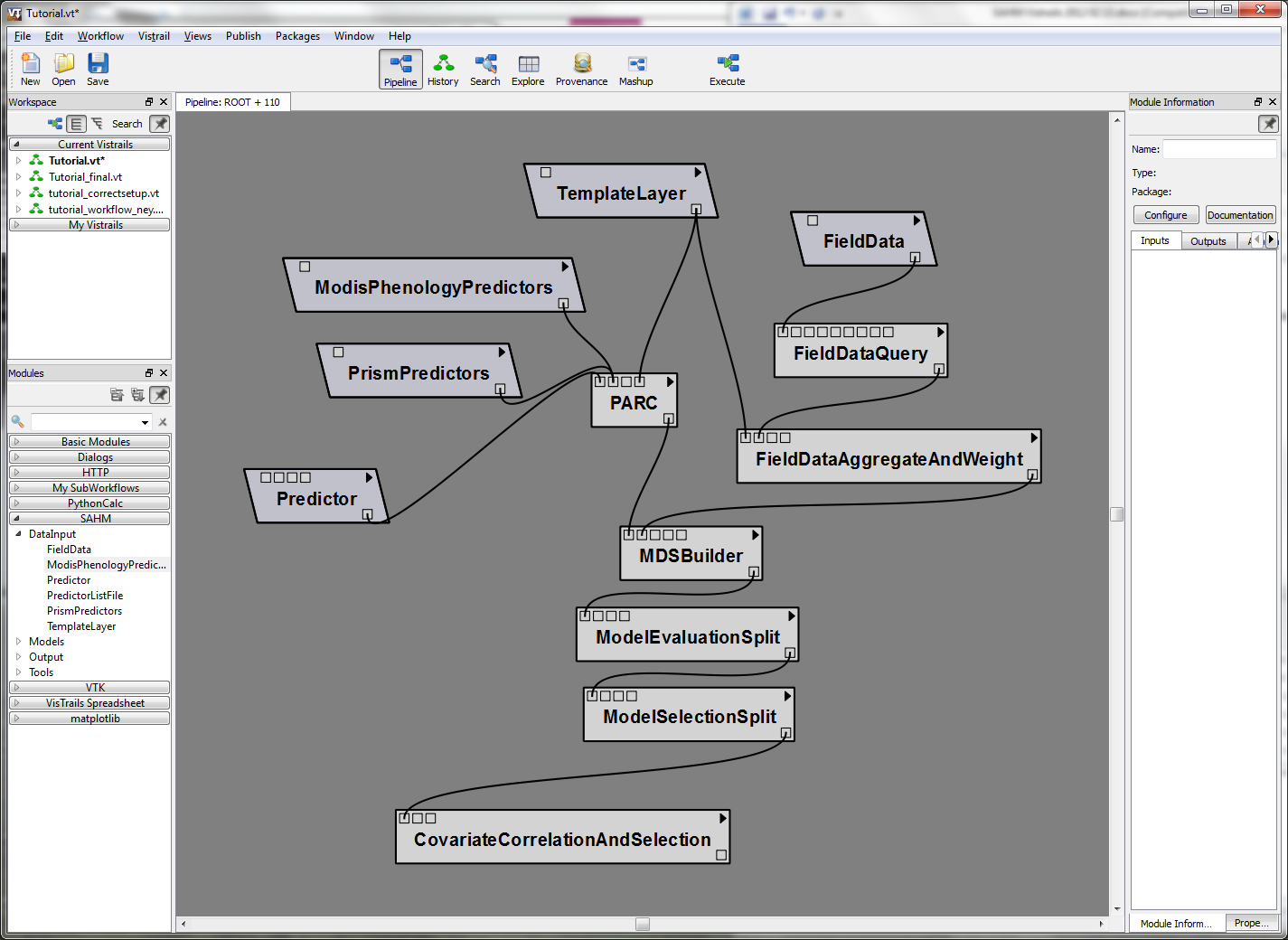


The next module we need to add is the MDSBuilder. Add this module and connect both the PARC and FieldDataAggregateAndWeight modelus to the MDSBuilder’s input boxes. Your workflow should now look similar to the one shown below:



### Preliminary model analysis and decision Modules

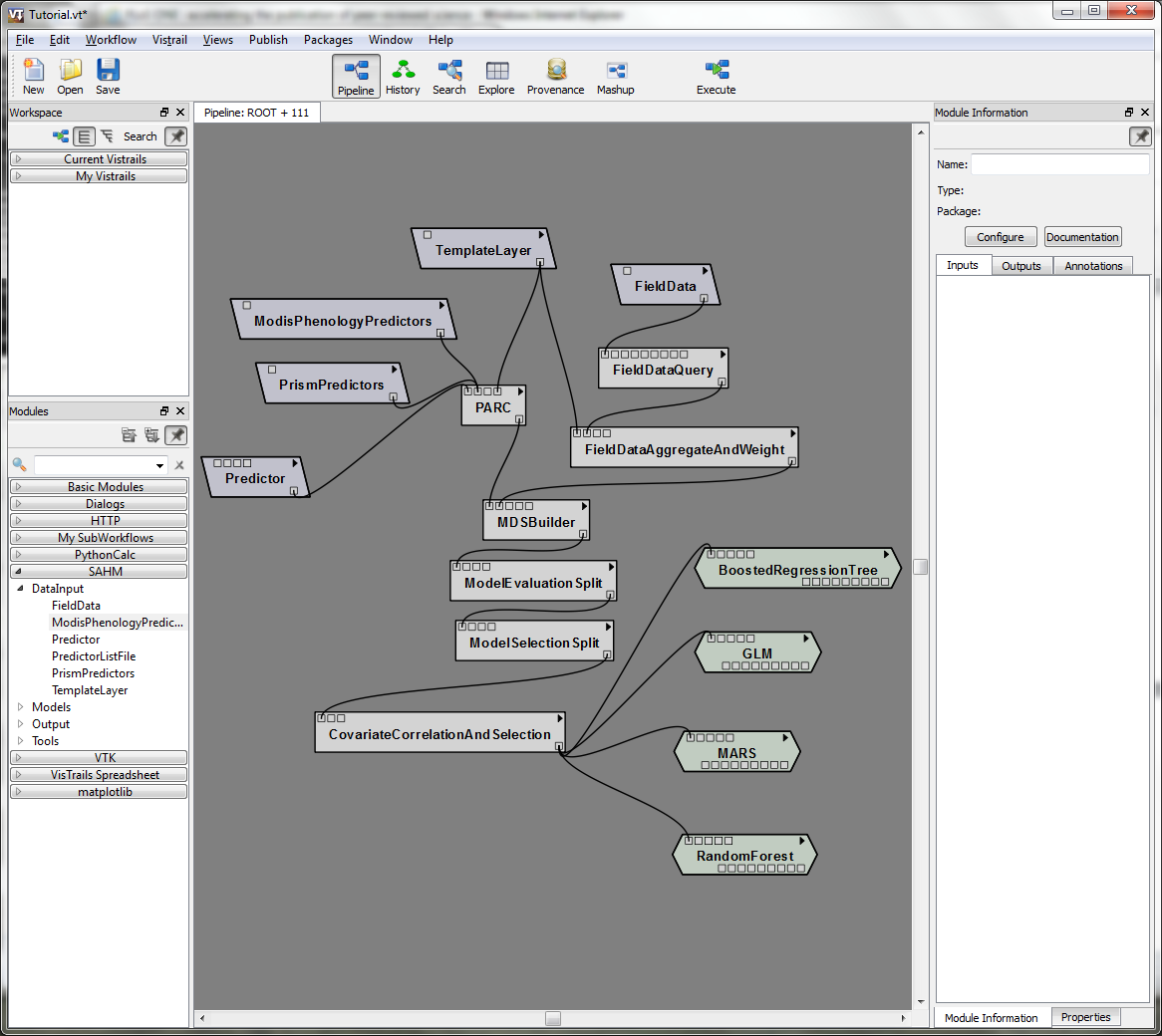
1. Now we will add the ModelEvaluationSplit module, connecting the output box from the MDSBuilder to one of the input boxes. This module provides the opportunity to reserve a specified portion of the data for producing and reporting evaluation metrics on an independent test set following model exploration and selection.
2. Next, Add the ModelselctionSplit module and connect the outputbox from the ModelEvaluationSplit to one of its input boxes. reserves a portion of the data from the model fitting process but reports the evaluation metrics on all models not just the those selected as the final models to be reported in the analysis (in contrast to the ModelEvaluationSplit). This module should be placed directly the CovariateCorrelationAndSelection. If both a ModelEvaluationSplit and a ModelSelectionSplit are specified then the training portion of the ModelEvalutationSplit will be further partitioned by the ModelSelectionSplit thus the ModelEvalutationSplit should come first in the workflow. Please refer to the documentation for more information on both the ModelEvalutationSplit and the ModelSelectionSplit modules.
3. The next module we will add is the CovariateCorrelationAndSelection will provide a breakpoint in the modeling workflow for us to assess how well each variable explains the distribution of the sampled data points and to remove any variables that may exhibit high correlation with others. After connecting the output from ModelSelectionSplit to the input to the CovariateCorrelationAndSelection module, the workflow should look similar to the one shown below:



### Correlative Model Modules

Now that we have all our input, pre-processing and preliminary model analysis modules added, we can add the correlative models we are interested. SAHM includes five of the most commonly used and best performing correlative models for species distribution modeling. These models are Boosted Regression Trees, Random Forests, Generalized Linear Models (GLM), Multivariate Adaptive Regression Splines (MARS), and Maximum entropy (Maxent). All models require absent points except for Maxent. We will first describe.

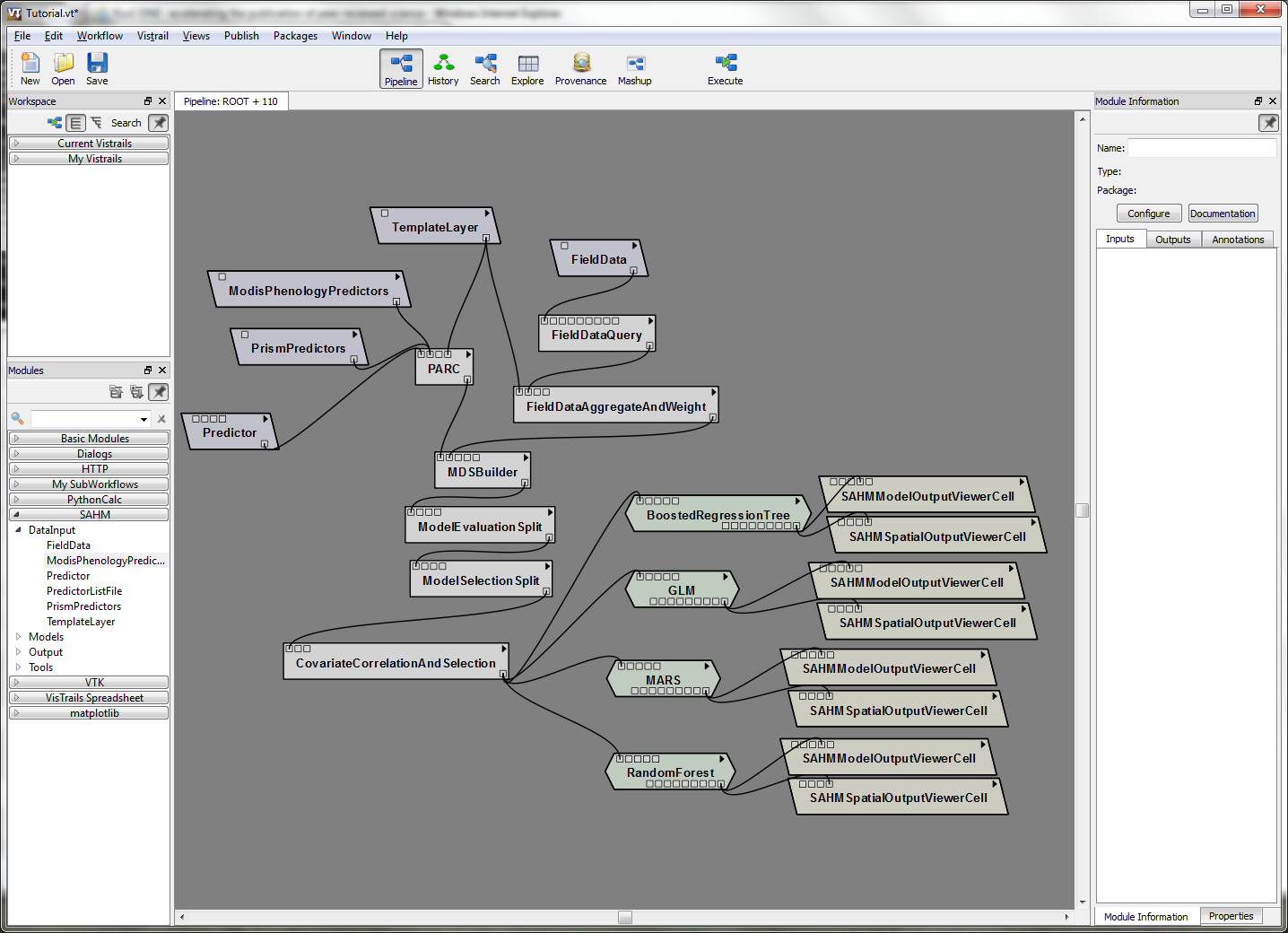
Add all correlative models to the canvas except for Maxent connecting the output box from CovariateCorrelationAndSelection to an input box for each of the models.



### Output Visualization Modules

The last type of modules we will add pertain to the visualization and organization of the output from the correlative models. Each model will produce evaluation metrics as well as spatial prediction maps.

1. For each correlative model, add the SAHMModelOutputViewerCell module. Use the Column and the Row fields to specify which cells in the output spreadsheet view you want that specific output to be viewed in. The SAHMModelOutputViewerCell will show the and the SAHM spatial output It contains tabs for the model output report, response curves, AUC plot, and calibration graph, confusion matrix and map of residuals.
2. Now add the SAHMSpatialOutputViewerCell moduel for each correlative model. Again, you can use the Column and Row fields to organize how the results are displayed. The SAHMSpatialOutputViewerCell displays the spatial outputs produced by individual model runs as well as the input presence and absence points and background points if applicable. At this point you should have a workflow that is ready to run and looks similar to the one below

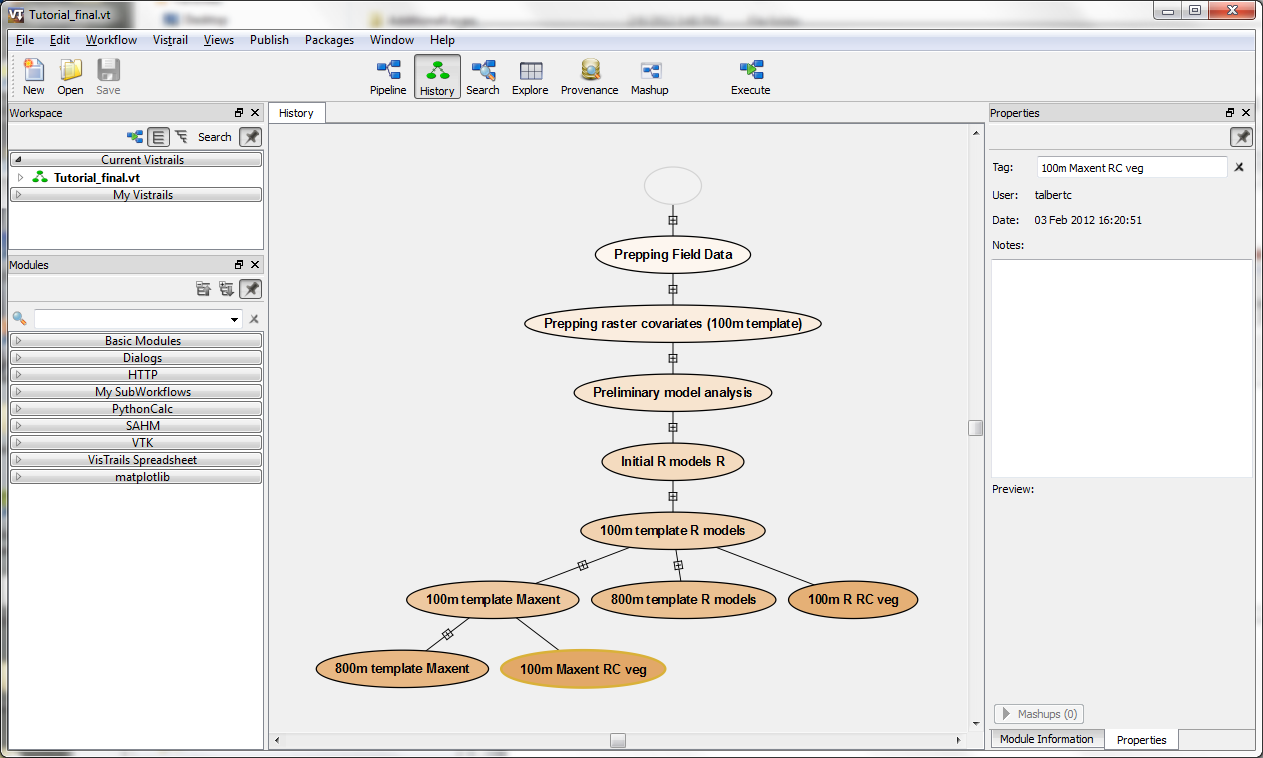


1. Continue to add, define, and connect modules, building your workflow until it matches the tutorial workflow. or load the tutorial workflow to move on to the sext section. As you add modules, continue to use the view documentation and refer to the user guide for clarifications.

# Running a Workflow

Note this section describes running the tutorial workflow included in the SAHM download.

1. Open the Tutorial\_final.vt workflow located in Central\_VisTrails\_x32\VisTrails\examples\Sahm\_example\Tutorial\_final.vt. For the purpose of this tutorial, we are going to explore model results using both climate and remotely sensed predictors, only climate predictors, and only remotely sensed predictors (See Morisette et al. 2012 for details on the specific predictors and their sources).
2. When it opens you will see that this particular workflow history is constructed to run Maxent. We can explore the history of this workflow by clicking on the history icon.



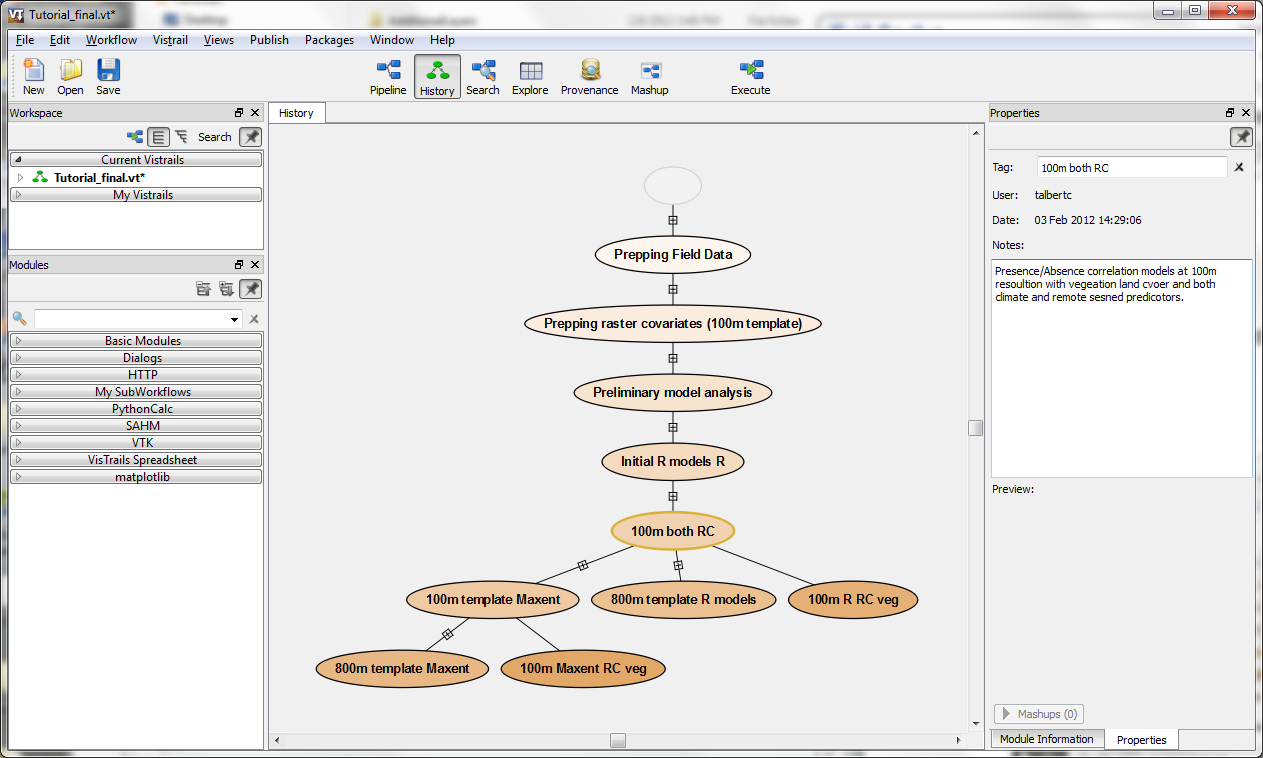
Notice that we see many different workflow changes that were create and documented in this workflow history. Currently, we are looking at the 100m maxent RC veg workflow.

1. For this example, let’s go back to look at the ‘100m template R models’ by selecting the associated oval and then going back to the pipeline view by clicking the pipeline icon. We can see that this workflow includes multiple types of predictor data including climate, remotely sensed and landcover represented by the PrismPredictors, ModisPhenologyPredictors and the Predictor modules respectively. The way this workflow is set up includes both climate and remotely sensed predictors. We can confirm this by clicking on the black arrow of the PrismPredictors and selecting edit configuration. You will see a list of predictors with checkboxes that are all checked which means they will be included in the workflow execution.

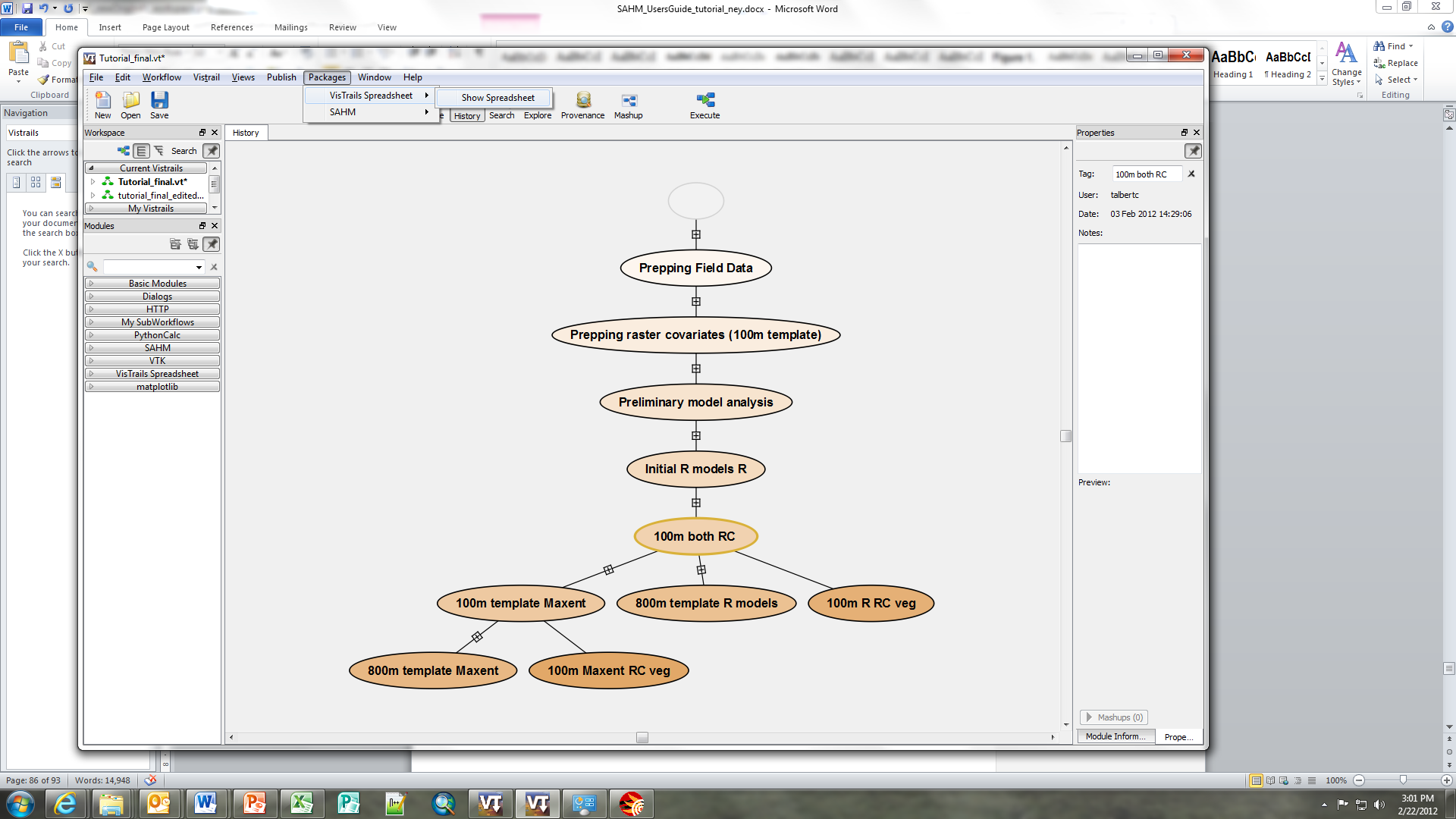


The same will be true for the ModisPhenologyPredictors. We can also look at the template layer used for this workflow by clicking on the TamplateLayer module. The template layer that is used is at a 100m resolution.

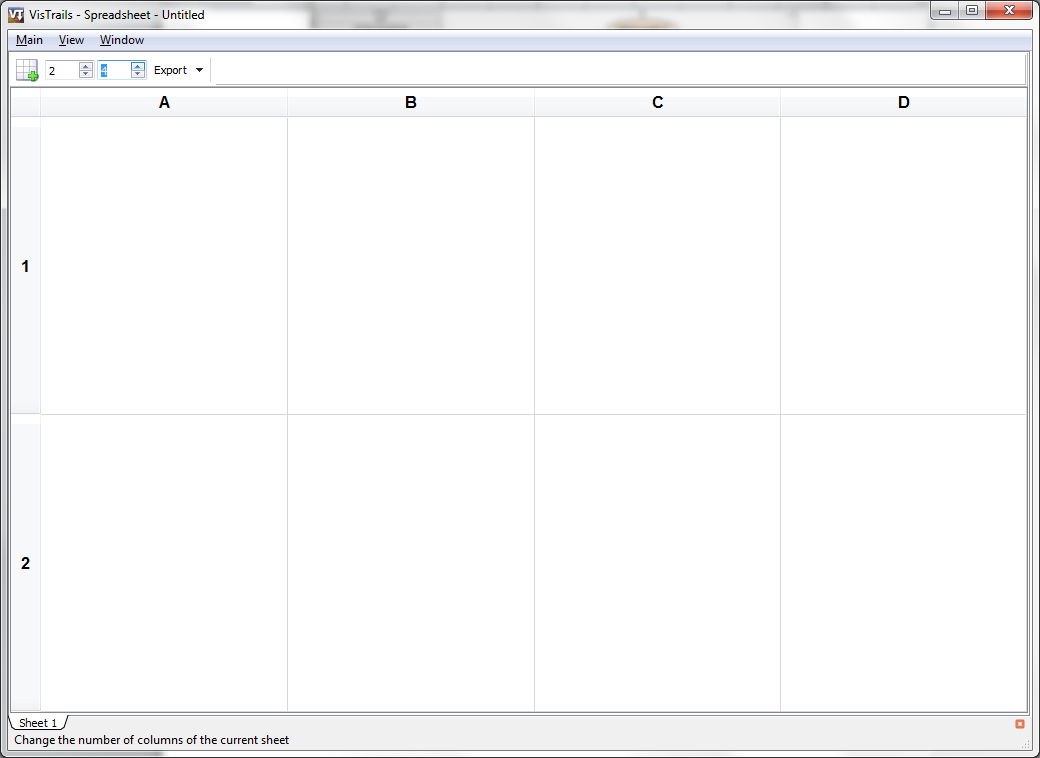
1. Since we know that we will be running multiple workflows under with varying predictor sets, we want to make sure we document the details of each workflow. We can do this by going back to the ‘History’ icon on the top navigation bar. We will want to change the name of the workflow by clicking in the oval and renaming it to 100m both RC and add documentation in the Note window on the right. Note that the history will keep track of any changes to the workflow including removing or adding modules and even the movement of modules so it best to start documenting the workflow sessions only when all modules are added and arranged as needed.

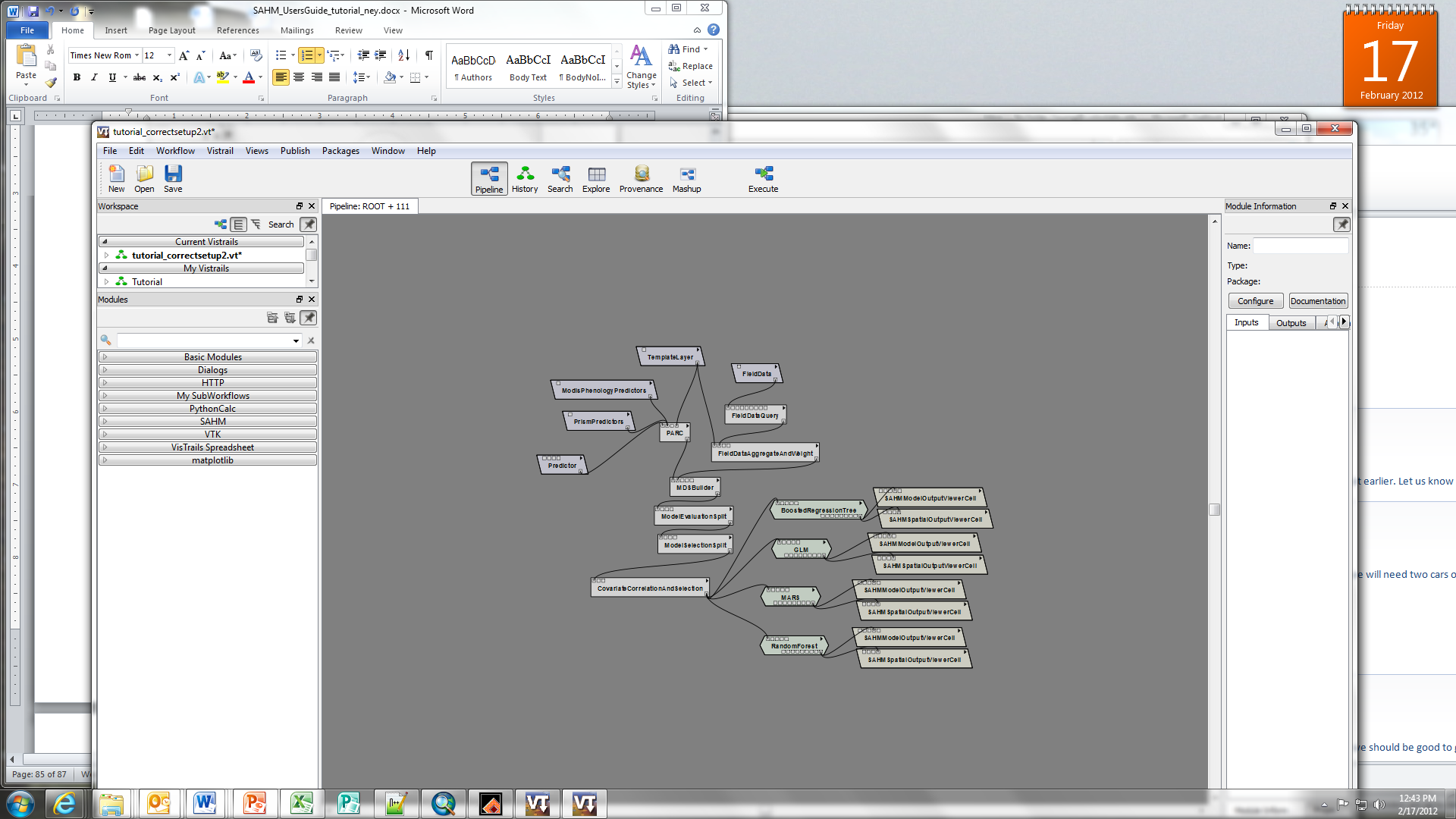


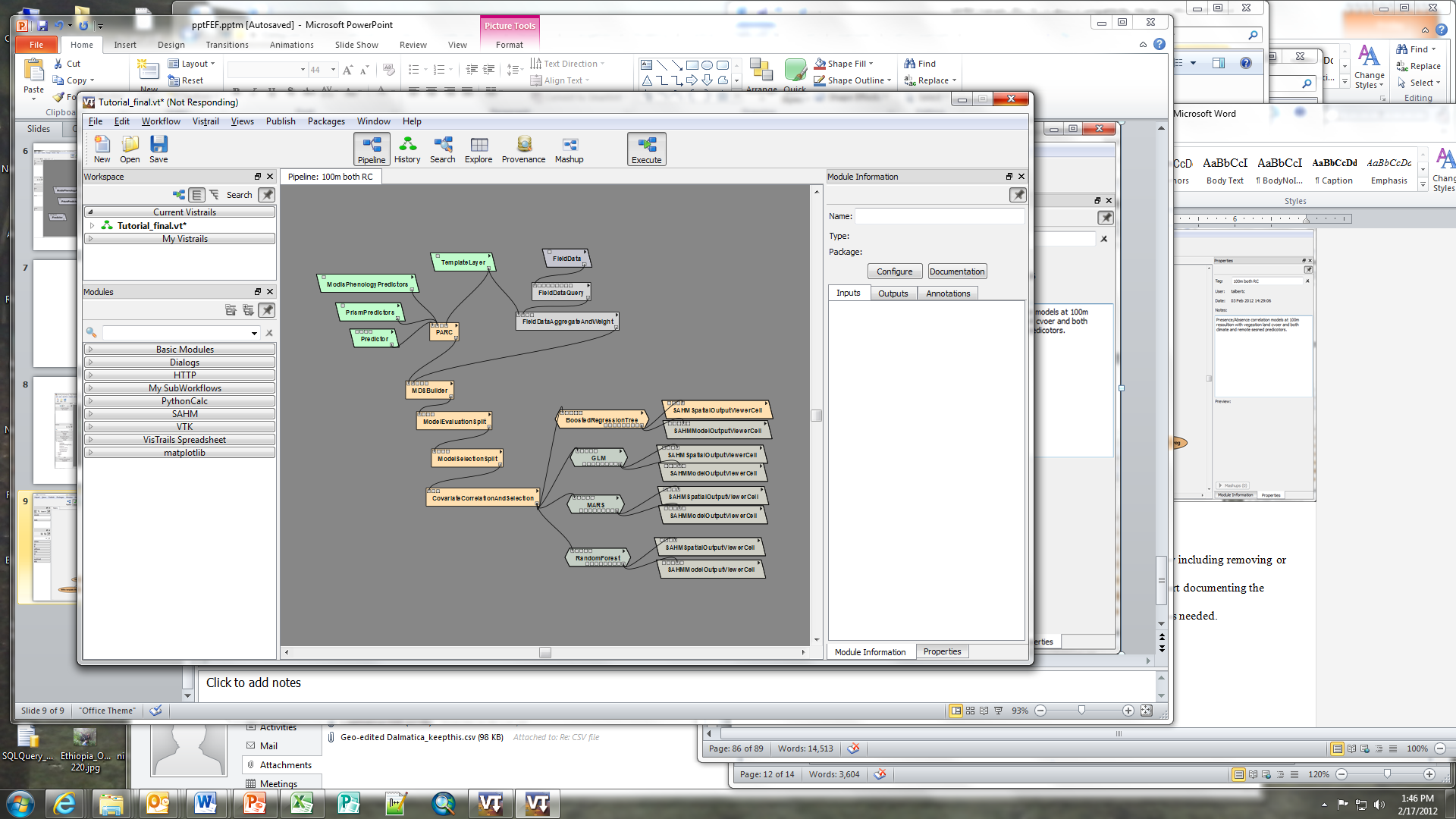
1. Before we run the workflow, we can organize how the results will be displayed in VisTrails using the Spreadsheet. To do this go to ‘Packages’ in the top navigation bar, then scroll over VisTrails Spreadsheet and select Show Spreadsheet



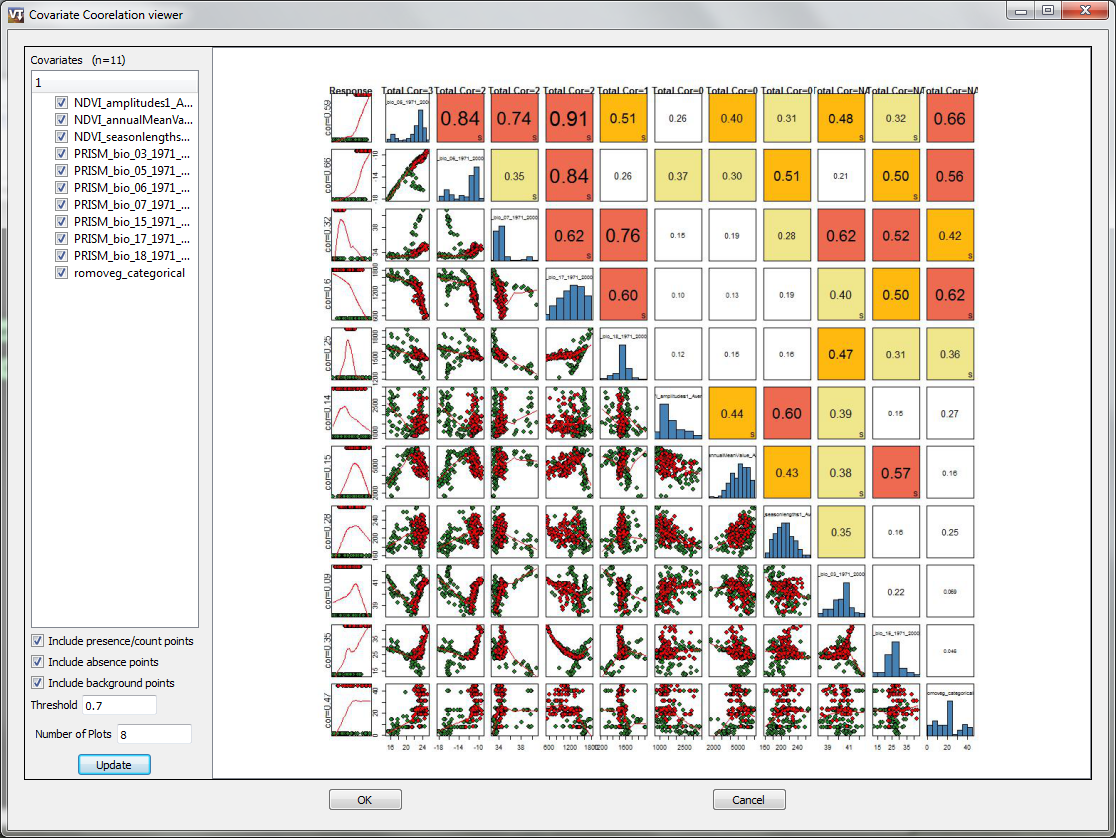
This will open the Spreadsheet that will hold all the output from the models (both general output and spatial output). Since we are running four different correlative models, we will want to create a spreadsheet that will have enough columns and rows to display all the output. Increase the number of columns to four and leave the number of rows at two.



1. We can now go back to the pipeline and run this workflow by clicking the execute icon. As its running, the modules change colors to indicate their status.

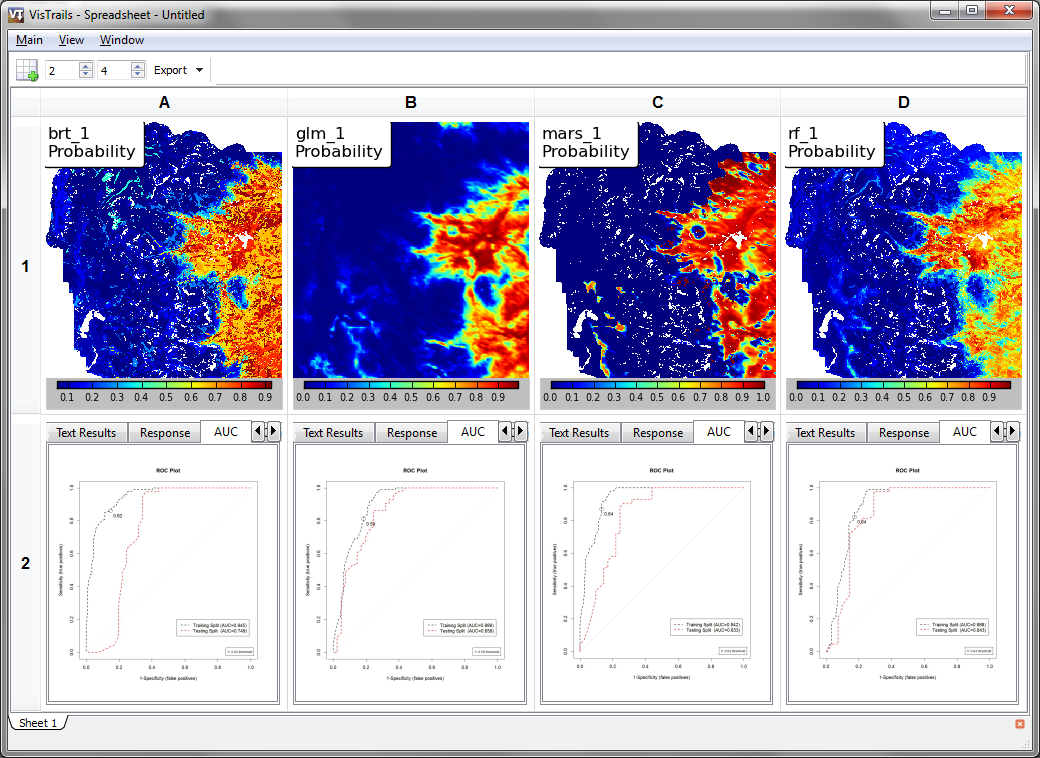


The workflow will continue running until it reaches the CovariateCorrelationAndSelection module. At this point, a new window should appear that displays the correlation values between all predictors. This allows us to assess how well each variable explains the distribution of the sampled data points and to remove any variables that may exhibit high correlation with others (refer to the User documentation to learn more about this module and its functions).

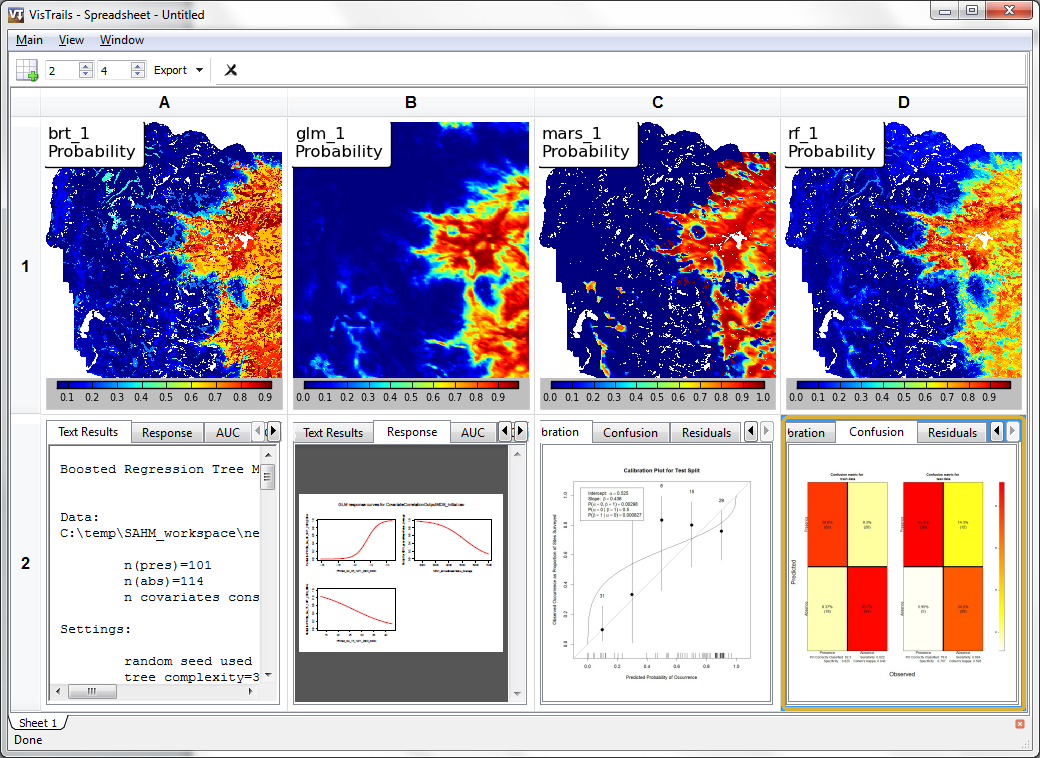


1. At this point, we could choose to remove highly correlated variables or only variables of a certain type and update to the display to see how these impacted the correlations. For the purpose of this tutorial, we will continue by clicking ‘ok’ and include all predictors.

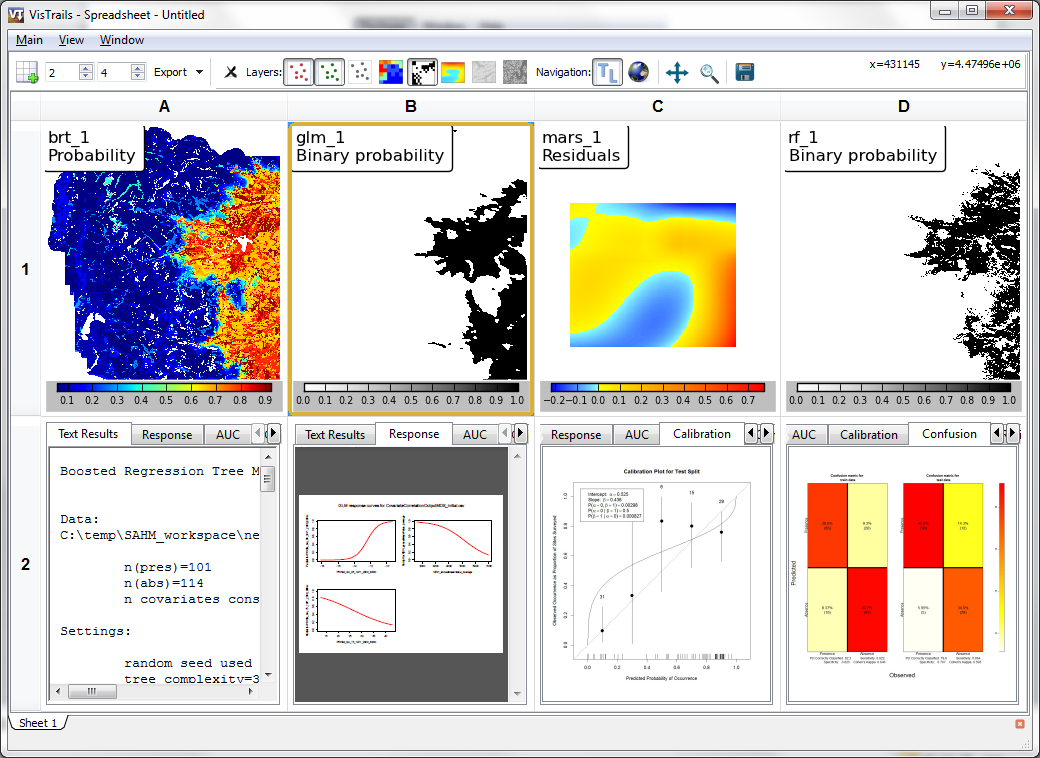
As the model continues you will see a VisTrails spreadsheet that we organized earlier appear with output from the models organized by cell.



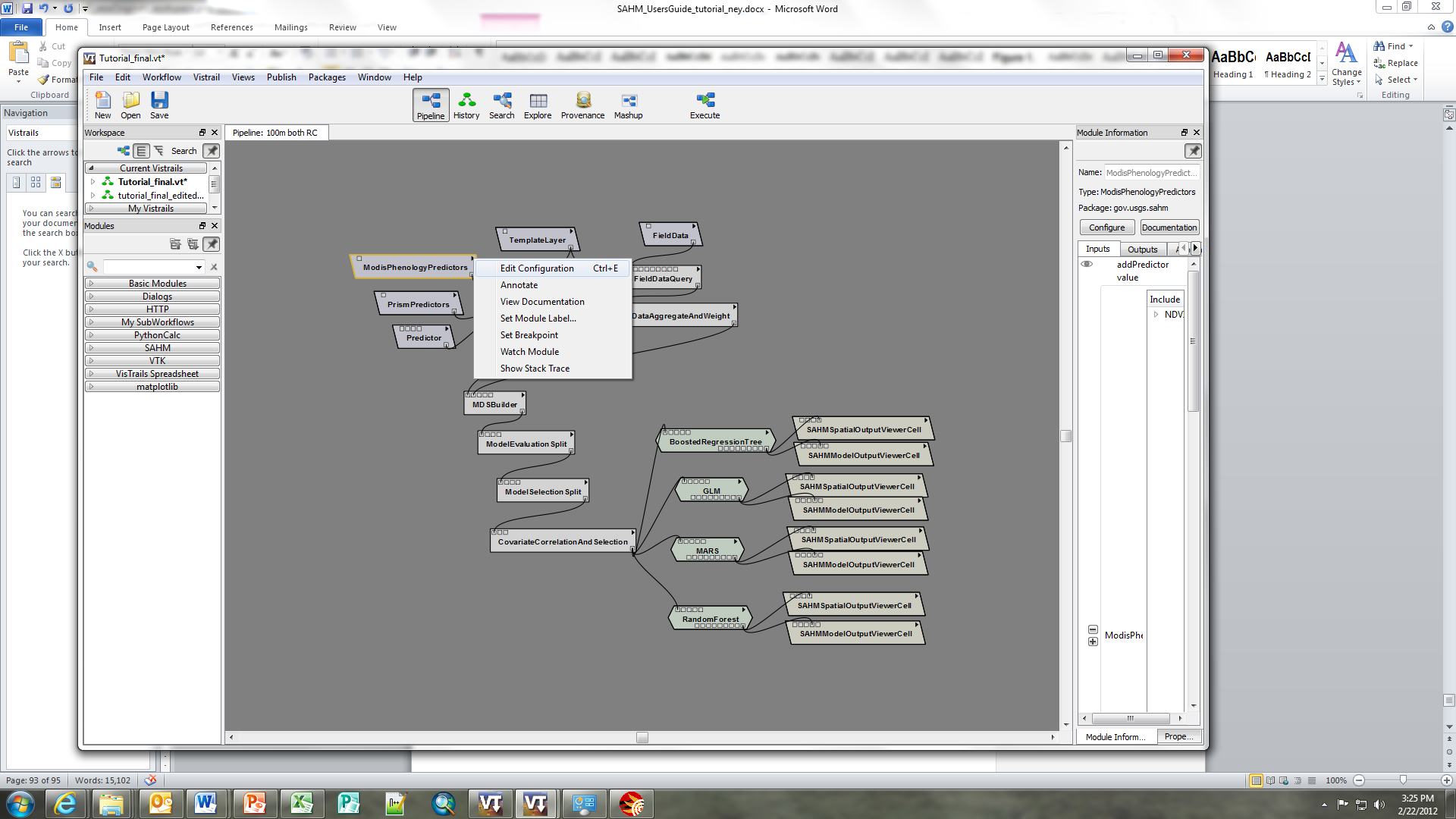
The cells that represent model results have multiple tabs that describe the results of that particular model. This is where you can find evaluation metrics, variable response curves, environmental predictor importance among other informative output.



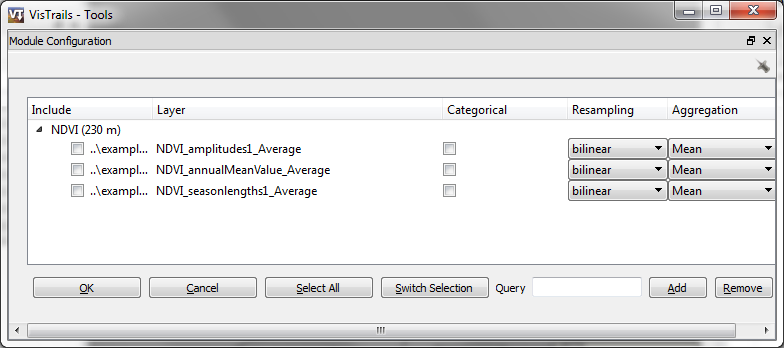
1. We can also explore the spatial output of a specific model by clicking on the spatial results cell. This will add a toolbar that will allow you to add or remove different spatial results specific to the model. This allows us to look at binary predictions, the presence locations, the residuals and other informative spatial display.



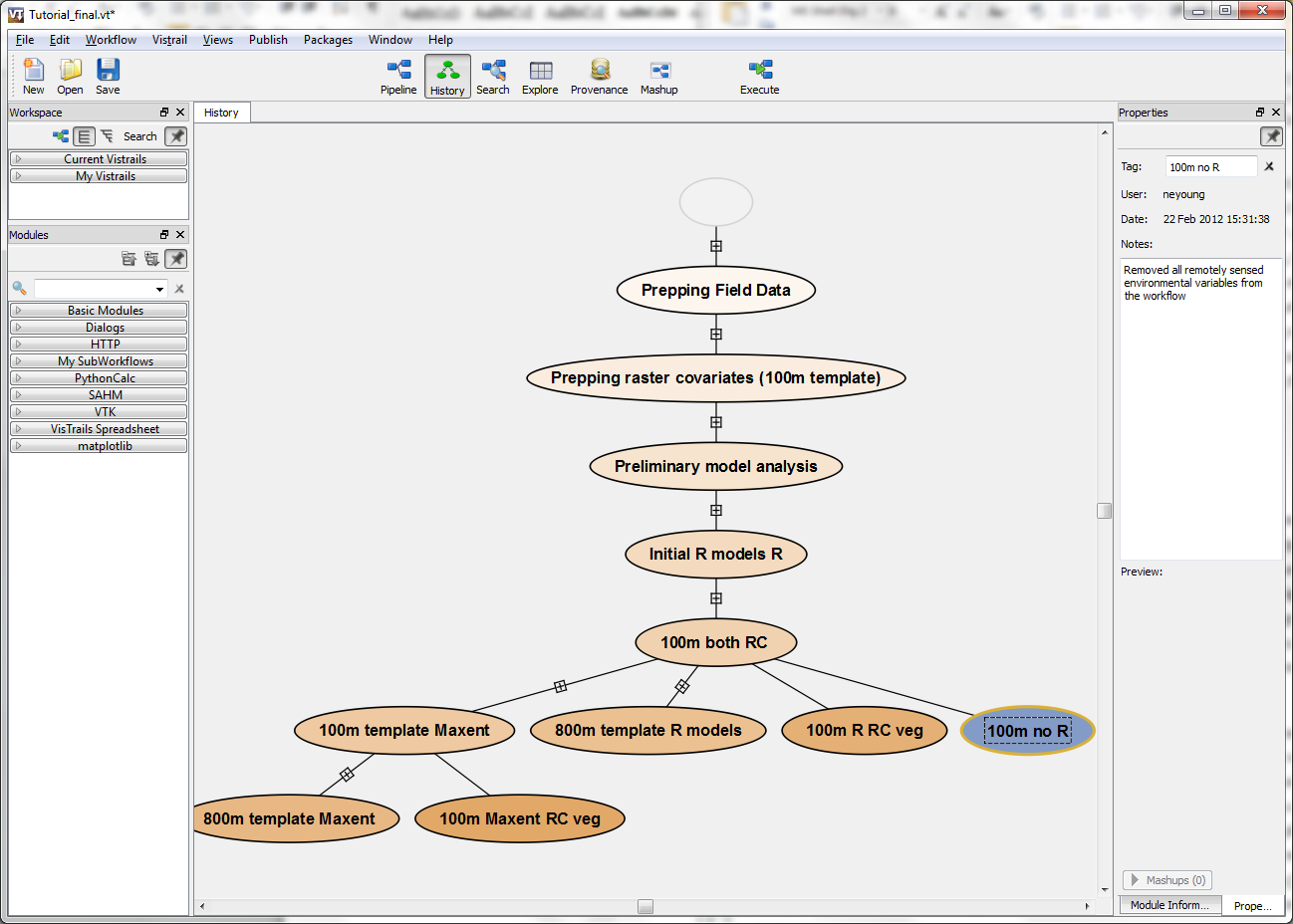
Now that we have the results for 100m resolution with both climatic and remotely sensed variables, lets change the workflow to only include climate variables. We can do this by going back to the history and selecting the ‘100m both RC’ workflow. If you double click on this oval, it will open the pipeline view. From the pipeline of this workflow, we can click the black arrow on the ModisPhenologyPredictors module and select ‘Edit Configuration’.



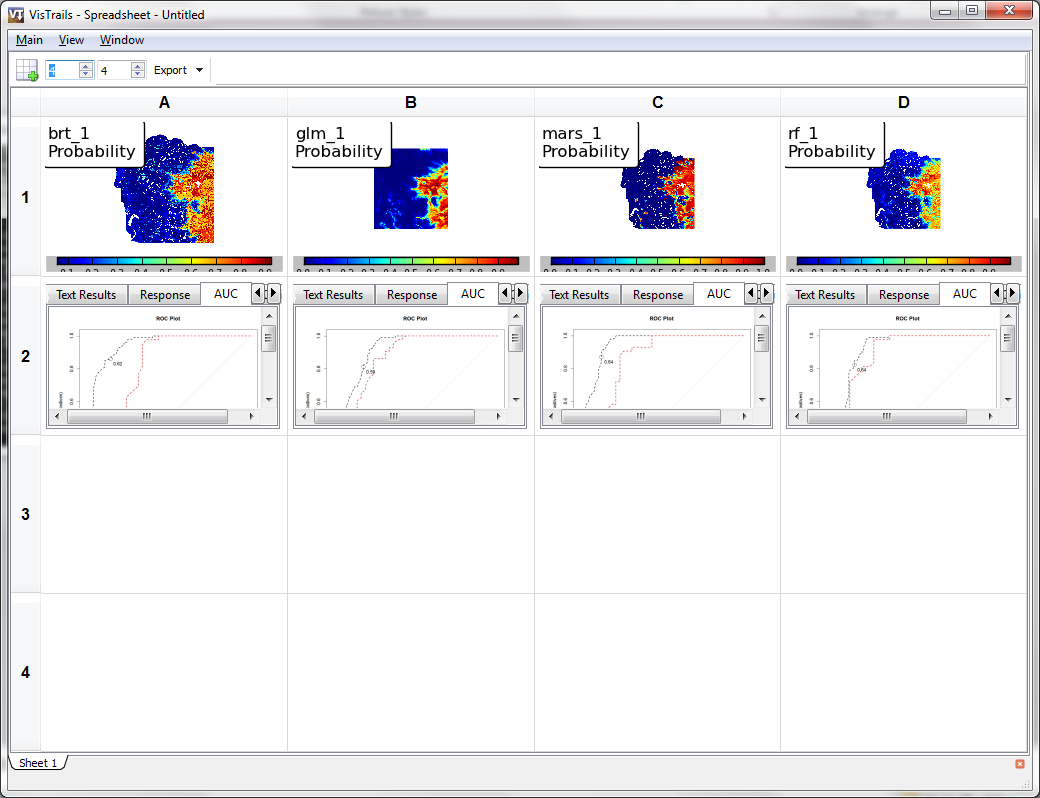
In this window, we will de-check all the layers and press ‘ok’ which will remove them from the workflow process.



Since we have made a change to this workflow, we will want to document this in the workflow history. Change the view to show the history. You will notice that there is a new oval that represent the change that we made. Rename this to represent the change that was made such as ‘100m no R’.

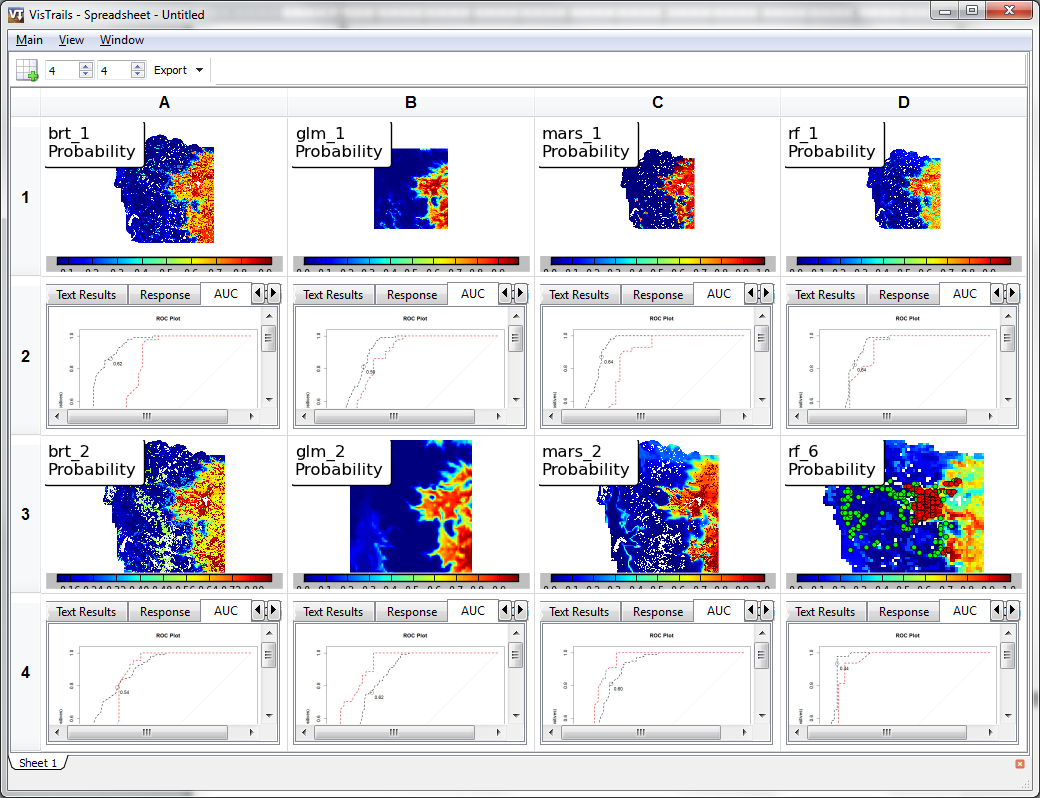


We can go back to the VisTrails spreadsheet and modify our spread sheet to include two more rows to hold the output under our no remote sensing scenario. In addition to adding additional rows to the existing spreadsheet, you could add a new spreadsheet in which to display the output of this scenario. The new spreadsheet would show up as another tab and you could then organize the output by tabs scrolling though different tabs to compare outputs.

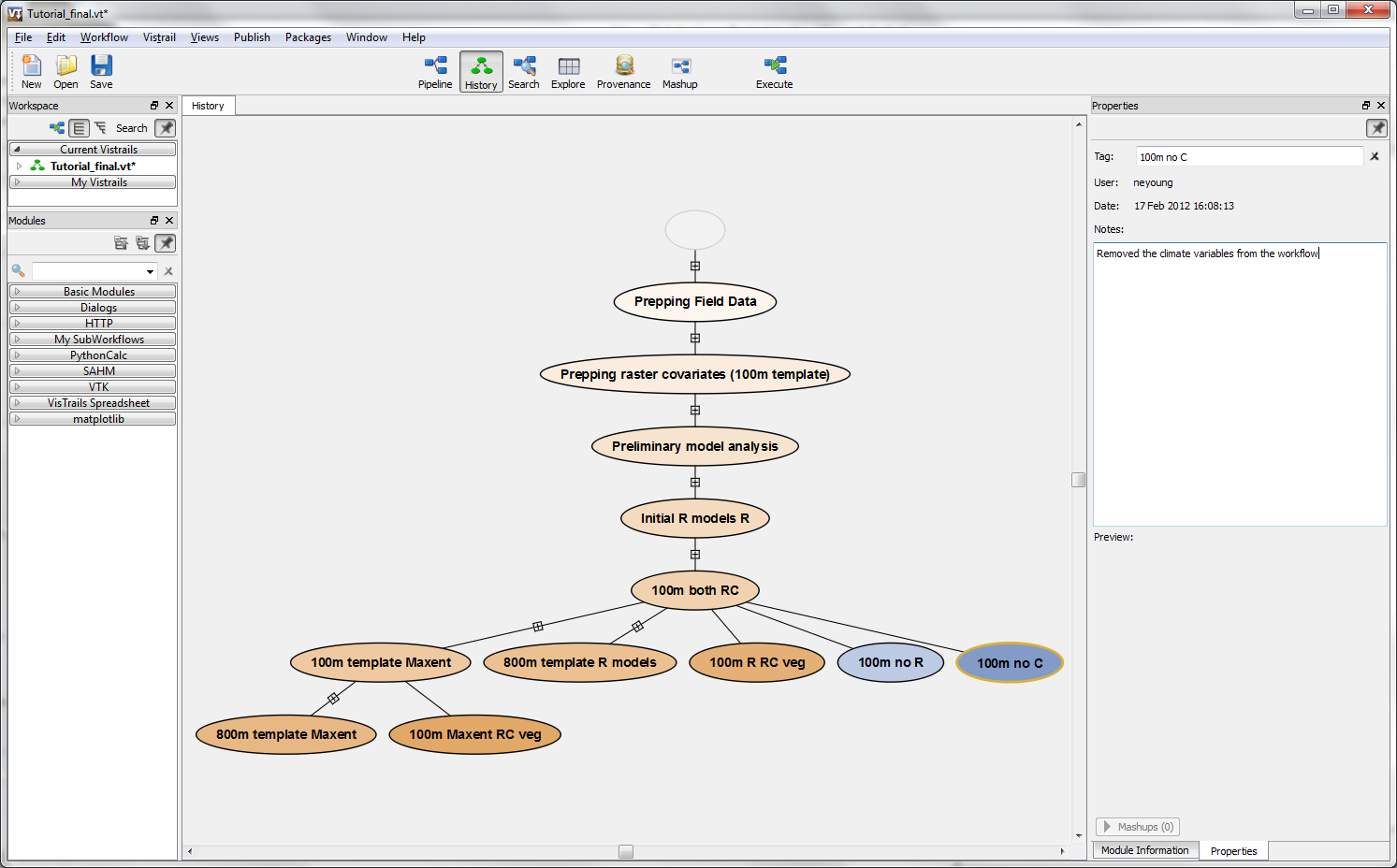


Now we can go back to the pipeline and run the workflow again. Notice how VisTrails will only re-run modules that are influences by the change we made which will speed up the time to finish the workflow.

Now we have output for four models under two scenarios that we can compare side-by-side. We can look at model evaluation metrics to see if including both the climate layers and the remotely sensed layers or just the climate layers.



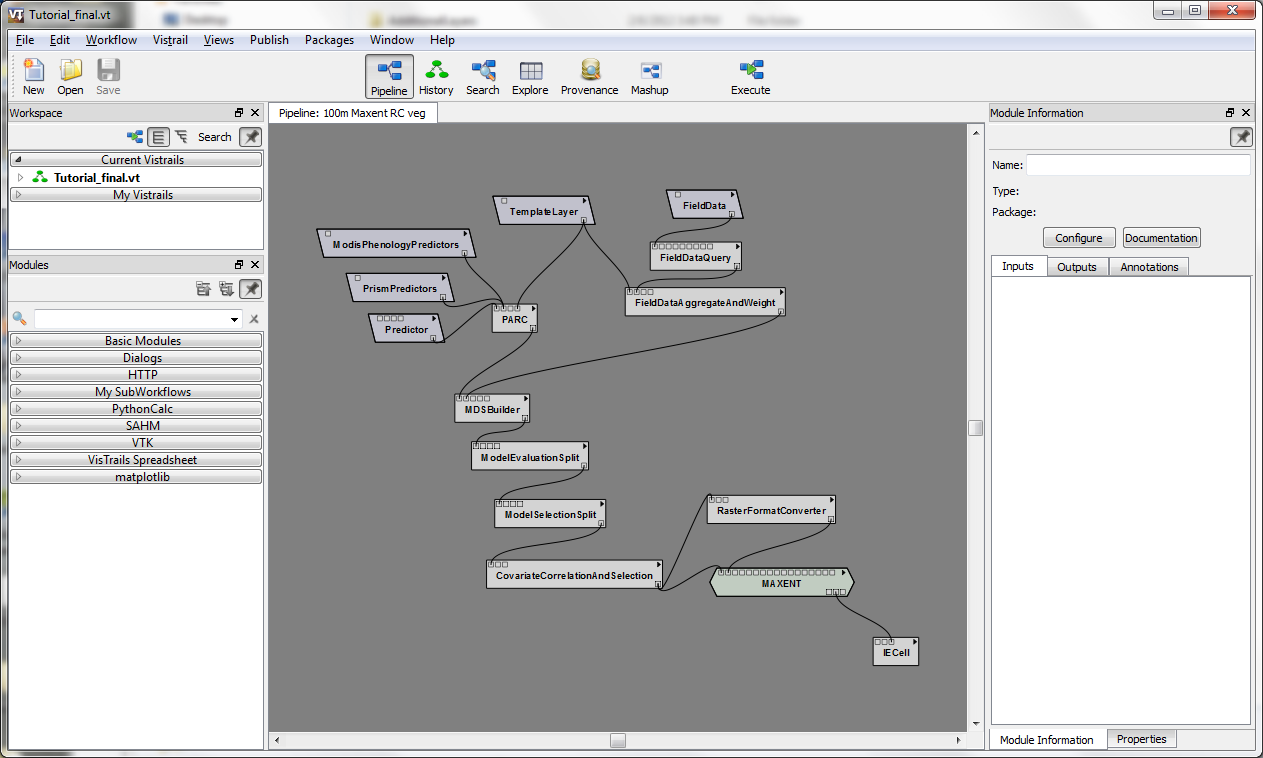
Next we can click on the past workflow ‘100m both RC’ in the history and go to the pipeline and instead of removing the remotely sensed predictors, we can remove the climate layers. Again, we will want to document this change in the history naming this workflow ‘100m no C’. And again, we will want to increase the number of rows in our spreadsheet to hold the model outputs under this scenario.



We can now run the workflow under these settings and compare the output with the previous two scenarios.

# Running a Maxent Workflow

Often, the presence data we have for a species of interest does not have absences data. In these cases a presence-only correlative method will have to be employed. The four correlative models discussed above are all designed for datasets that have both presence and absence data. Fortunately, SAHM also includes the presence-only method Maxent that allows users to use data that include only presence locations. If we go to the history view of the tutorial, we can see that there are workflows that include using Maxent. Open the 100m template Maxent workflow and view the pipeline.



# Provenance