# Outline of the Data Processing Workflow for the EASTWeb Software System

*DRAFT* – Prepared by Michael C. Wimberly and Yi Liu, November 19, 2014

The following document presents a general outline of the major data processing steps and input and output data that will be incorporated into the new version of the EASTWeb software being developed through the NASA ACCESS and NIH EPIDEMIA projects. The current version is a draft and should be considered a working document that will be updated as we continue to develop and clarify these procedures.

## Data Processing Steps

### Download

Access scientific data files from multiple online earth science data archives and store them locally for processing.

### Extract

Extract the necessary data layers from the scientific data files and convert them to a geospatial data format suitable for subsequent processing (e.g., TIF).

### Projection

Reproject the geospatial data into the projection chosen by the end user. Reprojection will always need to be done, but could potentially done at different stages in the processing workflow depending on the data source.

### Mosaic (optional)

For tiled data such as MODIS products, combine multiple tiles into a single geospatial data layer.

*The developer will select whether or not mosaicking will be implemented for each product.*

### Clip (optional)

Clip the spatial extent of the downloaded data to a smaller area of interest.

*The user will select whether to clip or not as an option. When we clip we should clip to the rectangular bounding box of the shapefile – not to the shape of the polygon itself.*

### Screening (optional)

For data with QC flags or other screening information, determine which pixels have “bad” data and convert them to a NoData value. Note that depending on the specific product, QC information may be accessed from the same scientific data files as the data layers (as is the case with MOD11A2) or accessed from a separate scientific data product (as is the case with

MCD43A4/MCD43B4, where the QC information is provided via a separate MODIS product MCD43A2/MCD43B2).

*The developer will select whether or not to implement screening for each plugin that is developed for a specific remote sensing product. For example, we are currently planning to implement QC screening for MODIS products, but not for TRMM or NLADS product.*

*For products that have QC screening implemented, the user will decide whether or not to implement QC screening in a project. User options for QC screening may be different for different products. For example, we may have different levels of QC screening for MODIS BRDF-adjusted reflectance versus MODIS land surface temperature.*

### Index Calculation

Calculate one or more environmental indices. Depending on the data source, the index calculation procedure may involve temporal summarization (for example, summarizing hourly NLDAS data to compute mean, minimum, and maximum daily temperature), or other types of summarization (for example, calculating spectral indices using reflectance data from an 8- or 16-day composite period).

*The developer* will determine the specific indices that can be calculated for each remote sensing product.

*The user will be able to select which of these indices will be calculated and included in a particular project.*

*Note: One major different from the previous version of EASTWeb is the implementation of cumulative indices, such as growing-degree days (GDD) or freezing-degree days (FDD) that will need to incorporate information from the previous time step as well as the current time step. For example, to compute GDD from NLDAS data, we would need to take the GDD from the previous day and add the number of degree above the GDD threshold for the current day's temperature.*

### Temporal Summary

Convert the index grid to the specific time periods selected by the user (e.g., Epi Weeks or months). If the environmental indices are computed on a daily basis, as with NLDAS and TRMM, then this should be a fairly straightforward processing of summarizing these daily values as mean, sum, or other statistic. We will only implement temporal summarization for hourly or daily products such as NLDAS and TRMM that can be aggregated to coarser temporal resolution in a straightforward manner.

*The developer* will determine whether or not temporal summarization (e.g., CDC epi-weeks, WHO epi-weeks, or months) will be enabled for a particular remote sensing product.

The user *will be able to choose the level of temporal summarization* (e.g., CDC epi-weeks, WHO epi-weeks, or months) *for products where it is allowable. The user will also have the option of choosing no temporal summarization, in which case the data will be kept at a daily resolution.*

### Masking

Pixels that we do not want to include in the spatial summaries will be temporarily converted to NoData before running the spatial summaries. For example, we will typically want to mask out areas dominated by water as we have done in the past. However, may also want to mask out other areas as well, such as non-malarious portions of districts in Ethiopia. The key differences going forward are (1) we don't want to permanently convert masked pixels to NoData, just temporarily convert them for the spatial summarization, and (2) we should start thinking of this as more of a generic mask than just a water mask.

Note*: We have previously assumed that the mask layer will be the same resolution as the MODIS data (1000 m). However, we need to go back and take a look at how masking is implemented when the datasets have a coarser resolution (for example, ~ 12.5 km for the NLDAS and ~ 25 km for TRMM). For example, we don't want to have a situation where an entire NLDAS or TRMM pixel is masked out because it overlaps with a single mask pixel. In these cases, we might want to just turn masking off or come up with a way to do the masking.*

### Spatial Summary

Compute zonal statistics (count of good pixels, mean, median, sd, min max) by overlaying the shapefile of summary zones on the masked temporal summary grid. Results will be output to the tabular database.

## Input Data

### Scientific Data Files

Files downloaded from online archive. May be in a variety of formats including HDF, NetCDF, GRB, or raw binary.

The user *will have the option of saving the scientific data files or deleting them after the necessary data has been extracted and processed.*

### Zone Shapefile

Contains the polygons that will be used to compute spatial summaries.

### Mask Grid

Indicates the areas that will be excluded from the spatial summaries.

The user *will be able to specify the mask file when choosing the project, or choose not to have a mask file associated with a project.*

## Output Data

### Index Grids

Environmental indices summarized at their “native” temporal resolution (e.g., daily climate summaries from NLDAS or 8-day indices computed from MODIS products).

*The user will have the option of saving the index grids or deleting them after the necessary data has been extracted and processed. However, we also need to keep in mind the previous point about cumulative index calculations. I think we would always want to retain he index grid from the previous time step to facilitate cumulative index calculations.*

### Temporal Summary Grids

Environmental indices summarized either summarize or interpolated to the user-specified temporal summary period (e.g., weeks or months). *Users should be able to choose an option of whether or not to save or delete these intermediate files.*

*The user will have the option of saving the temporal summary grids or deleting them after the necessary data has been extracted and processed.*

### Spatial Summary Table

Contains tabular data on the summarization of environmental indices for each spatial unit (e.g., county or district) and user-specified temporal summary period (e.g., weeks or months).

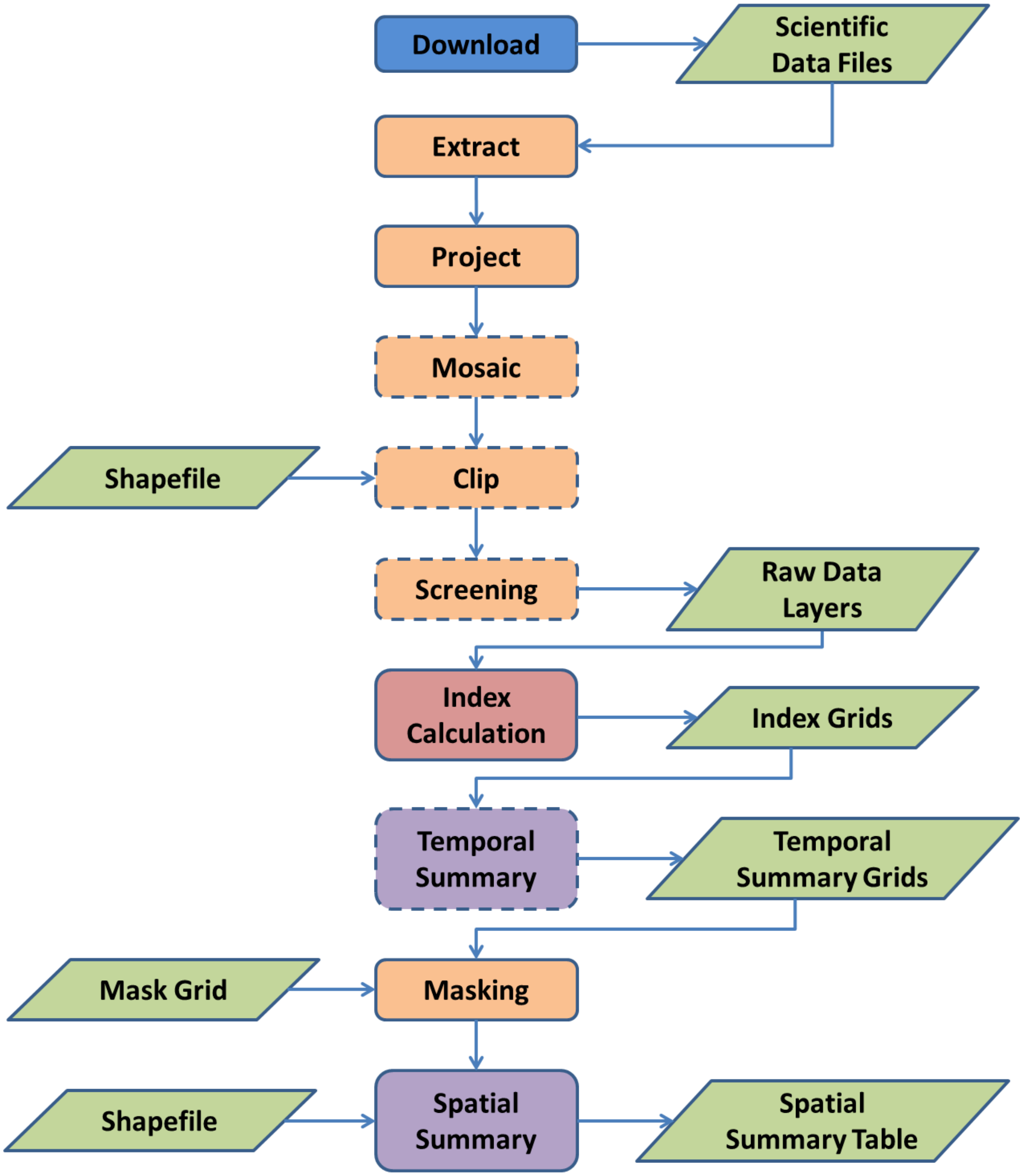

Figure 1: Workflow diagram for the EPIDEMIA system. Dashed lines represent optional processing steps.