DZnmf v. 1.0

User Guide

Characterizing sediment sources by non-negative matrix factorization of detrital geochronological data

J.E. Saylor[1]

K.E. Sundell[2]

G.R. Sharman[3]

N.M. McLean[4]

1: Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX, USA

2: Department of Geosciences, University of Arizona, Tucson, AZ, USA

3: Department of Geosciences, University of Arkansas, Fayetteville, AR, USA

4: Department of Geology, University of Kansas, Lawrence, KS, USA

Contents

1. INTRODUCTION

DZnmf is a MATLAB-based graphical user interface (GUI) designed to implement non-negative matrix factorization (NMF) of detrital geochronology data following methods outlined in Saylor et al. (in review, EPSL). The program can be run either in MATLAB, or as a stand-alone GUI with Windows or macOS.

2. SETUP

DZnmf version 1.0 requires installation of the MATLAB Compiler Runtime (MCR) version R2018a 9.4. This compiler allows the DZnmf GUI to run on the user's computer without having MATLAB.

Download the MCR for Windows here:

https://ssd.mathworks.com/supportfiles/downloads/R2018b/deployment_files/R2018b/installers/win64/MCR_R2018b_win64_installer.exe

Download the MCR for macOS here:

https://ssd.mathworks.com/supportfiles/downloads/R2018b/deployment_files/R2018b/installers/maci64/MCR_R2018b_maci64_installer.dmg.zip

Select all the default settings and install (this takes some time, usually about 15 minutes). The MCR will create a new directory on the user's machine that will be accessed in the background when the GUI is running; it will not interact with any existing versions of MATLAB or other versions of the MCR on the user's machine.

The DZnmf GUI does not require any installation other than the MCR compiler. Once the MCR is installed, download the DZnmf GUI and example data sets from the Supplemental Material in Saylor et al. (2019), or at these websites:

https://github.com/kurtsundell/DZnmf

https://www.kurtsundell.com/downloads/

The GUI file can be saved to any location on the user's computer. As long as the default settings were selected for the MCR it will be accessible by the DZnmf GUI. Occasionally there is an error in the MCR installation, which will throw the error message in Figure 1.
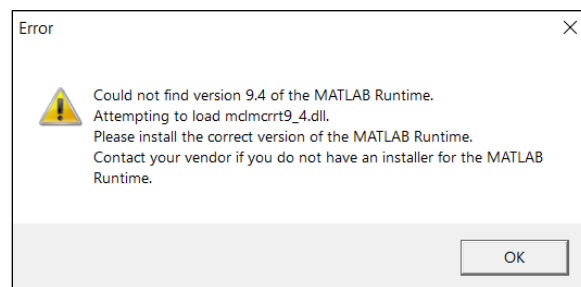
Figure 1. *Error message generated when the GUI cannot access the MATLAB Runtime compiler (MCR). Note this shows an error for version 9.4; This version of DZnmf (1.0) requires version 9.5 MATLAB 2018b.*

If the error message in Figure 1 or a similar error message is thrown, it means the DZnmf GUI cannot find and/or access the MCR. If this happens there are two possible solutions to troubleshoot: (1) restart the machine and try again, or (2) reinstall the MCR (the MCR may need to be completely removed and directory deleted before reinstalling rather than simply overwriting). If these options do not work please contact one of the first two authors.

Upon opening the DZnmf GUI there will be two blank listboxes and two blank plots (Figure 2). All of the program functionality (i.e., analysis, plotting, exporting, etc.) is contained within this single main page.
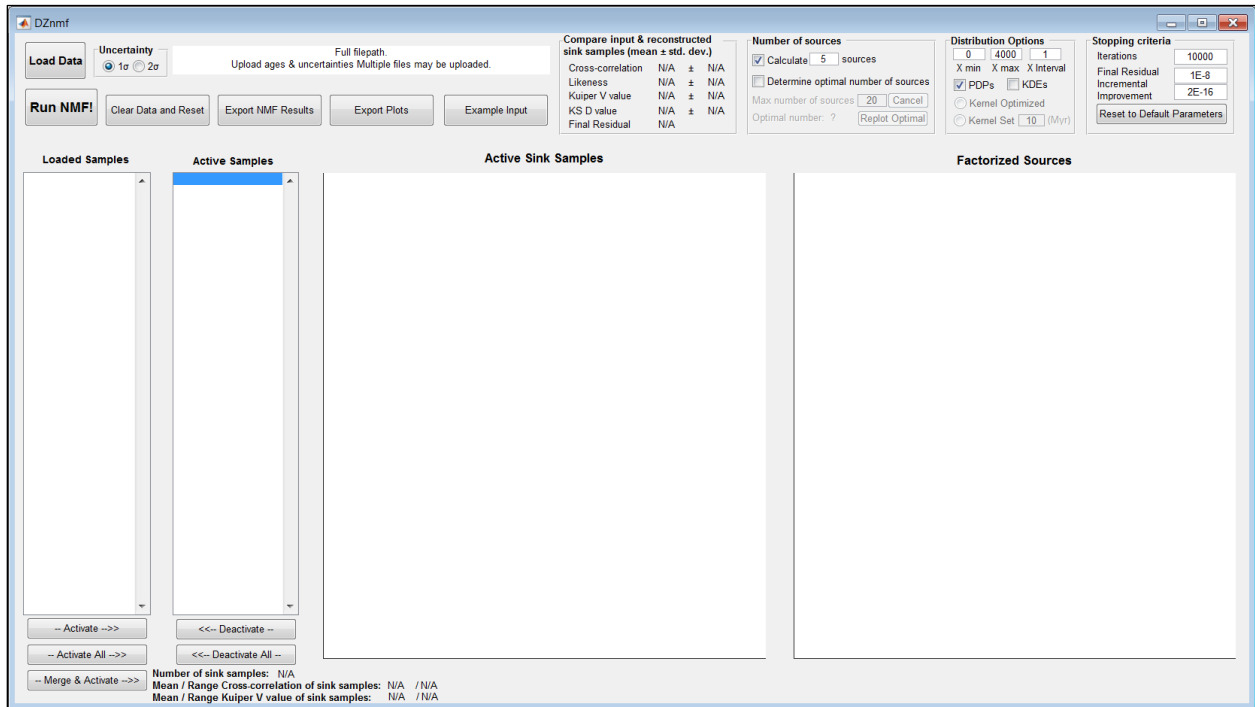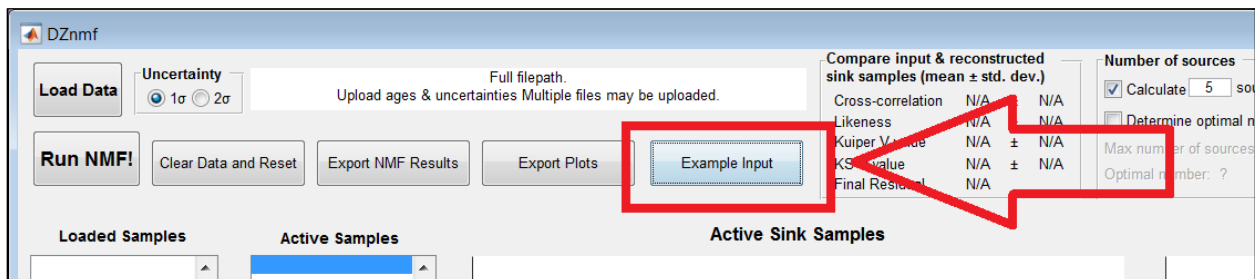

Figure 2. DZnmf GUI main page.

3. DATA INPUT

The input data needs to be set up so each sample has two separate columns with one column mean age and the other column the uncertainty; note each column must have headers (names) at the top of each sample (Figure 3). There is no limit to the number of samples that can be loaded, other than visualizing all of the distributions at the same time and limits of the computer's memory. Even if KDEs are used, each sample in the input data needs to have two columns: age and uncertainty (or kernel bandwidth). Example input can be accessed by pressing the **Example Input** pushbutton (Figure 3).

Figure 3. *Example input data can be shown by pressing the **Example Input** pushbutton (upper). Example detrital data from the Grand Canyon used in the examples below from Gehrels et al. (2011). Note that the macOS version must be either .xls or .xlsx file format.*

To load data press the **Load Data** pushbutton and selecting the uncertainty level (1σ or 2σ) (Figure 4). Windows can upload .csv, .xls, or ..xlsx file formats, macOS must be either .xls or .xlsx file format. Once the data are loaded the left listbox (Loaded Samples) will be populated with the sample names (Figure 5). The sample names must include letters.
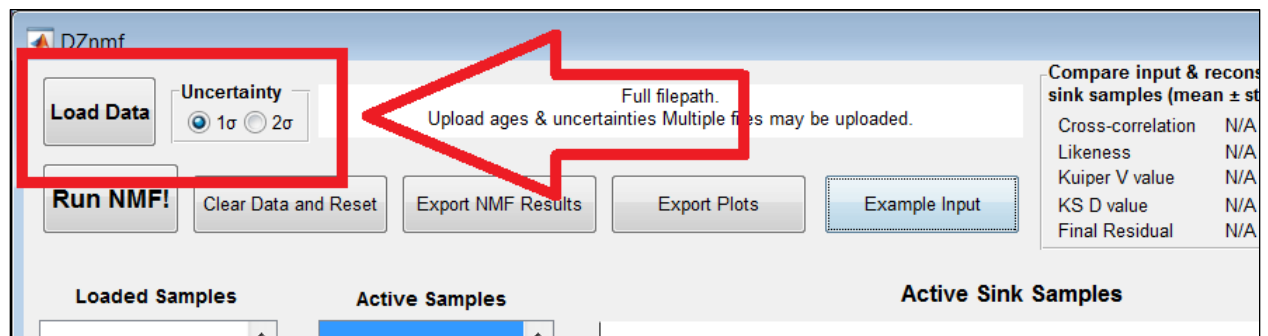

Figure 4. *Loading data and selecting the uncertainty level.*
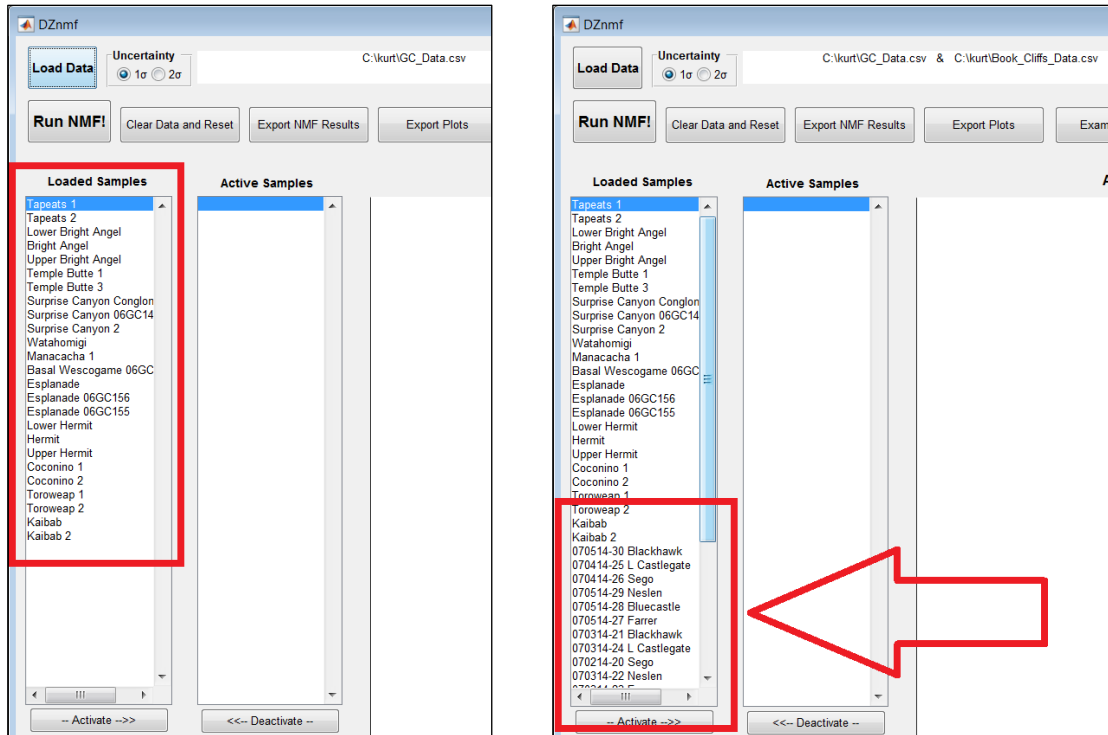
*Figure 5. Data files can be loaded one at a time. The **Loaded Samples** listbox shows the data sample names (left). Multiple files can be uploaded one at a time, the figure on the right shows a second input appended to the existing Grand Canyon data.*

4. AGE DISTRIBUTIONS

The next step is to specify how the distributions will be made. There are options to make probability density plots (PDPs) or kernel density estimates (KDEs) (Figure 6). This needs to be specified before activating the samples. Once samples are activated these options will be grayed out to ensure all distributions are generated the same way.
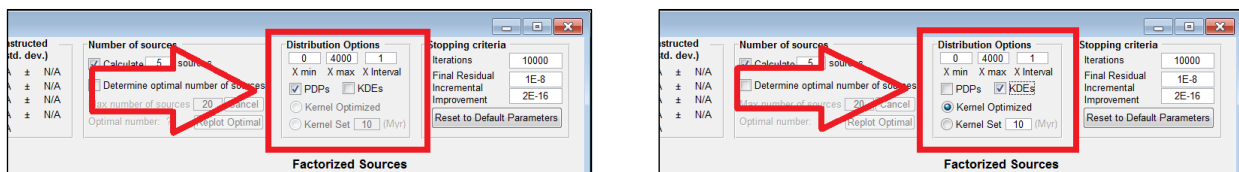
Figure 6. Options to use probability density plots (PDPs, left) or kernel density estimates (KDEs, right) before activating samples and running NMF.

5. ACTIVATING, MERGING, AND DEACTIVATING SAMPLES

Next, select the samples from the **Loaded Samples** listbox and press the **Activate** pushbutton. Alternatively, all samples may be activated by pressing the **Activate All** pushbutton (Figure 6).
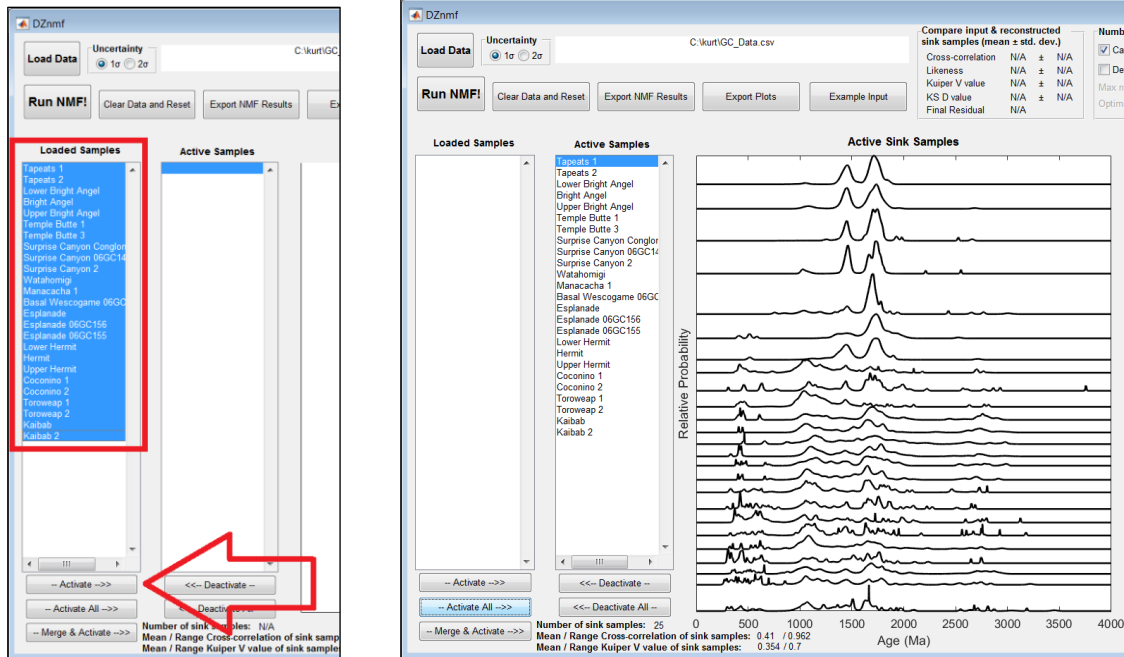
6

Figure 7. Activate samples by selecting the loaded samples and pressing activate.

The user has the option to merge (combine) data sets (Figure 8). When samples are merged their sample names will be horizontally concatenated (Figure 8). This is the same as combining ages and uncertainties into a single age distribution. The new merged distribution will be added to the **Active Sink Samples** plot on the bottom. Note the plotted distributions are always plotted in the same order as the **Active Samples** listbox (Figure 8).
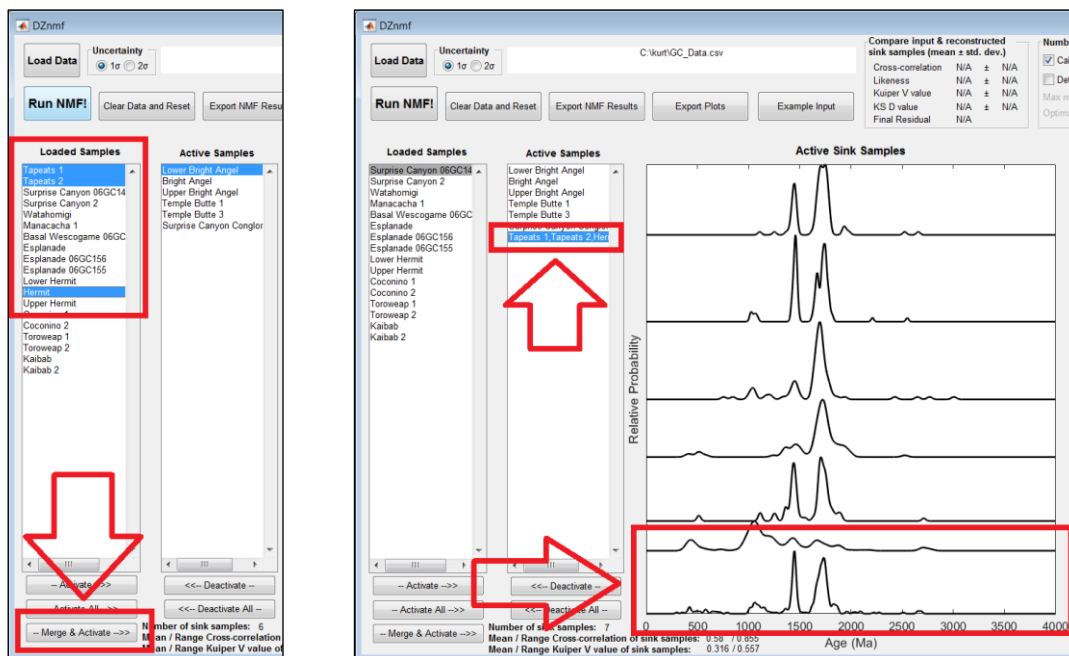

*Figure 8. Merging samples. Left: Select samples to merge. Right: Merged sample is added to the **Active Sink Samples** plot, which is always in the same order as the **Active Samples** listbox.*

Active samples may be removed by highlighting individual sample names and pressing the **Deactivate** pushbutton (Figure 9). The user also has the option to **Deactivate All**. This will clear the **Active Sink Samples** plot and return all of the samples to the **Loaded Samples** listbox in their original order, and the distribution options (Figure 6) will be made available.
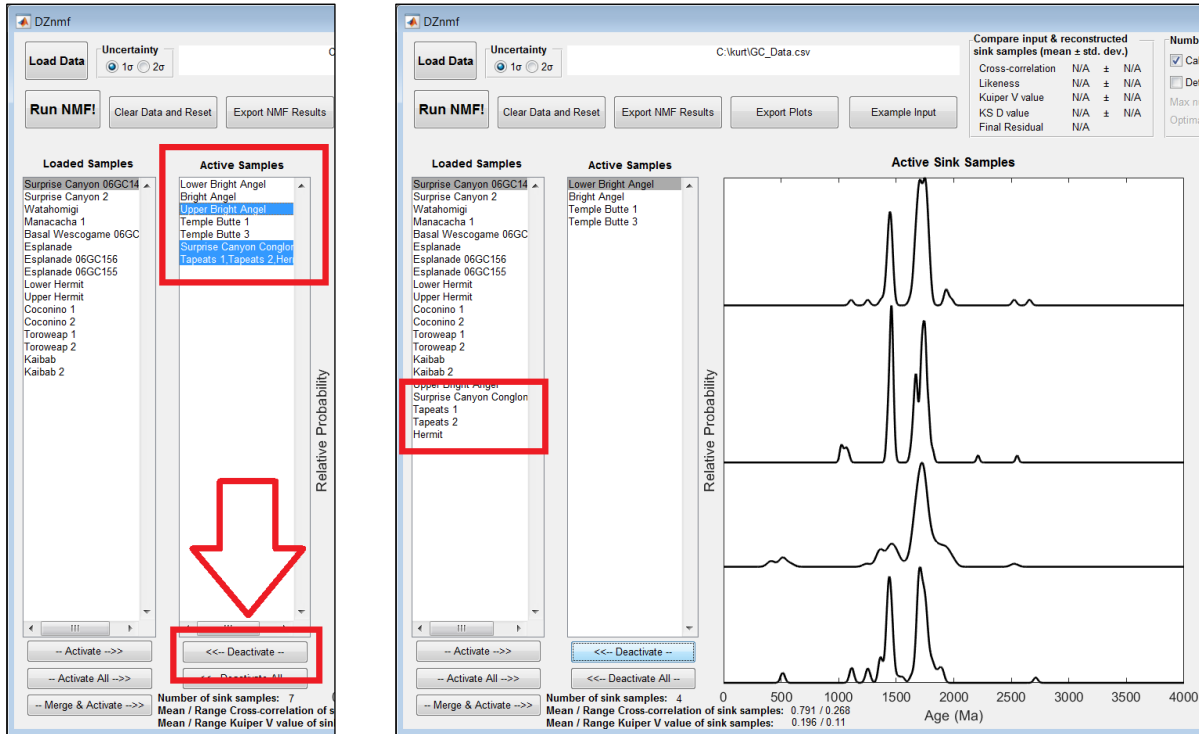
Figure 9. *Left: Example of deactivating selected samples by highlighting and pressing the **Deactivate** pushbutton. Right: Sample names are returned to the bottom of the **Loaded Samples** listbox and the distributions are removed from the **Active Sink Samples** plot.*

6. STOPPING CRITERIA

The default stopping criteria is set to 10,000 iterations, 1E-8 final residual, and 2E-16 incremental improvement (Figure 10). These parameters are suitable for most applications to detrital geochronology, and will rarely need to be changed. However, the user has the option to change these parameters and to reset to default parameters (Figure 10).
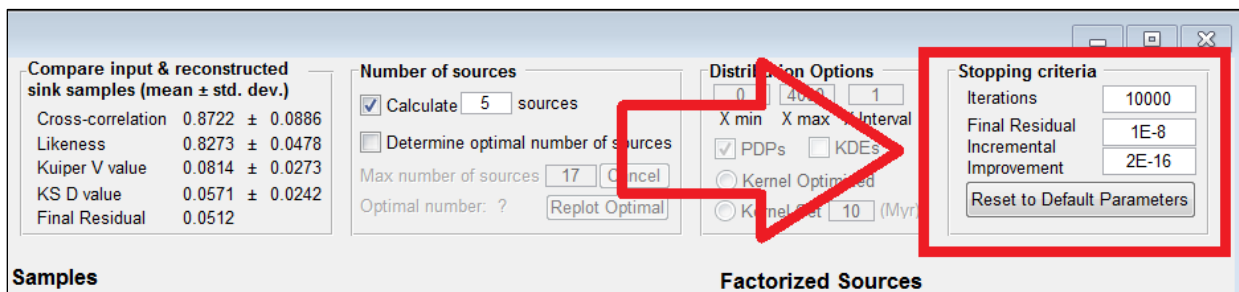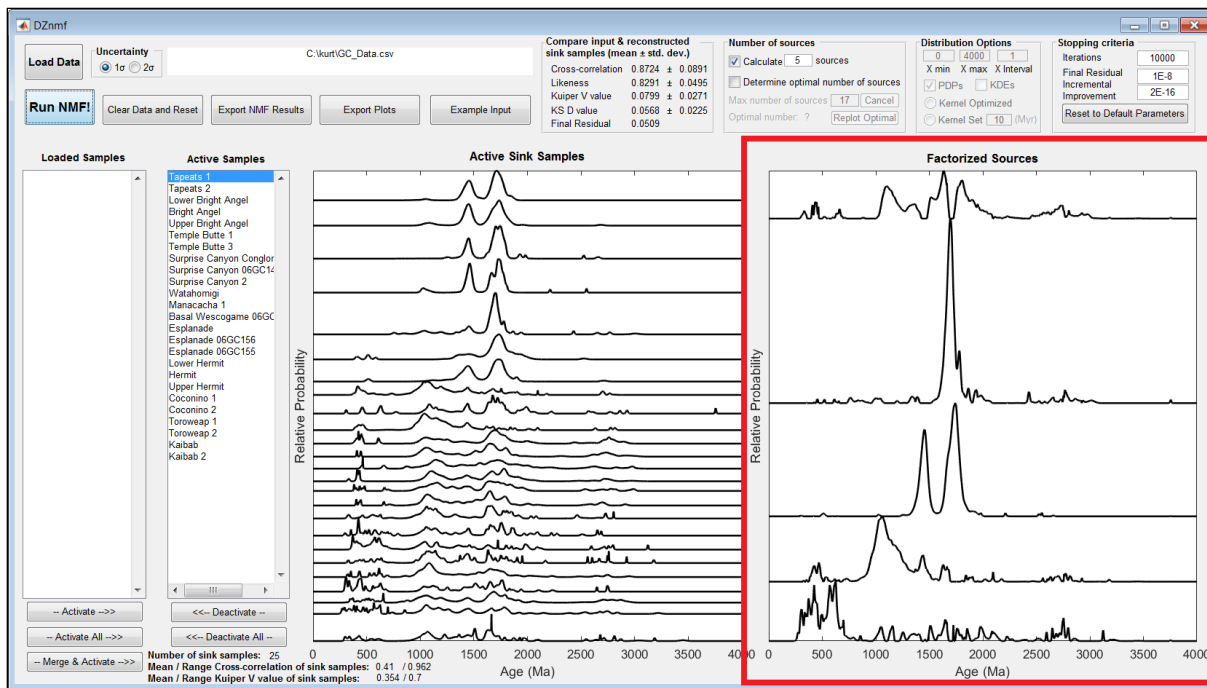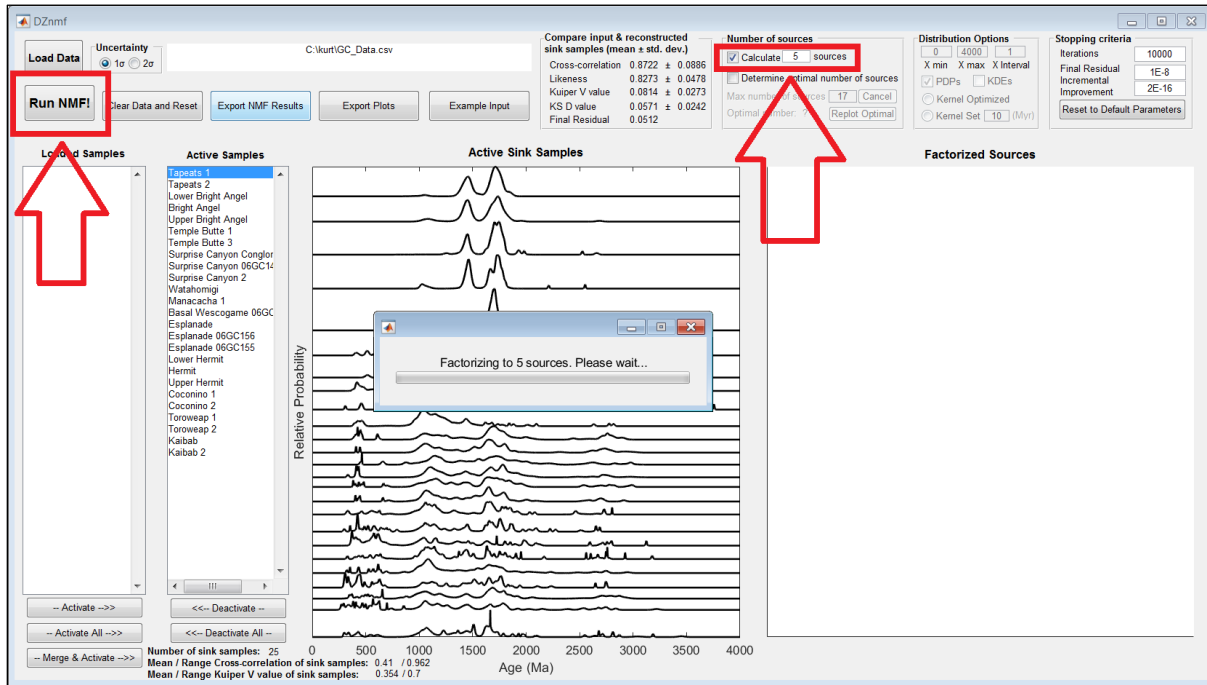

Figure 10. *Stopping criteria.*

7. RUN NMF!

Once the data are loaded and samples are activated the NMF may be run. The default setting is a specified number source samples to reconstruct (factorize) from the active sink samples. The user has the option to select how many sources to factorize (Figure 10). Press the **RUN NMF!** pushbutton to run NMF.

Figure 11. *Factorization of five sources. Upper: Select five sources to from **Number of sources** panel and push **Run NMF!** pushbutton. Lower: Five factorized sources*

Instead of specifying the number of sources to factorize, an alternative approach is to specify the maximum number of sources (N), and determine the optimal number of sources to factorize. Here, the program will individually factorize to 2 through N sources, and plot the optimal number of sources based on calculating the final residual between input and reconstructed sink samples over the range of ranks (2 to N); see section 3.2 of Saylor et al. (in review, EPSL) for a detailed description. The user has the option to cancel the run (Figure 12).
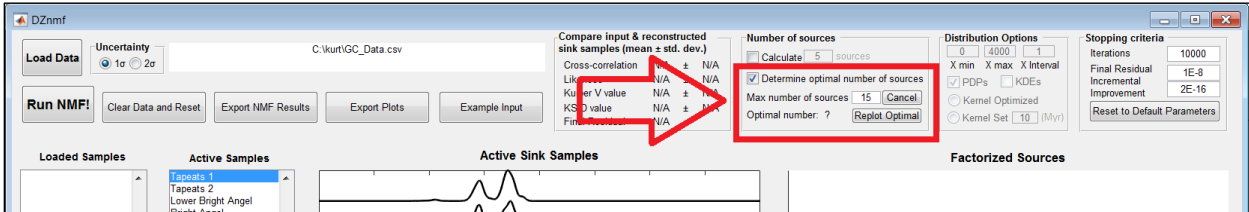
Figure 12. *Check box for determining optimal number of sources. Edit text box specifies the maximum number of sources to factorize.*
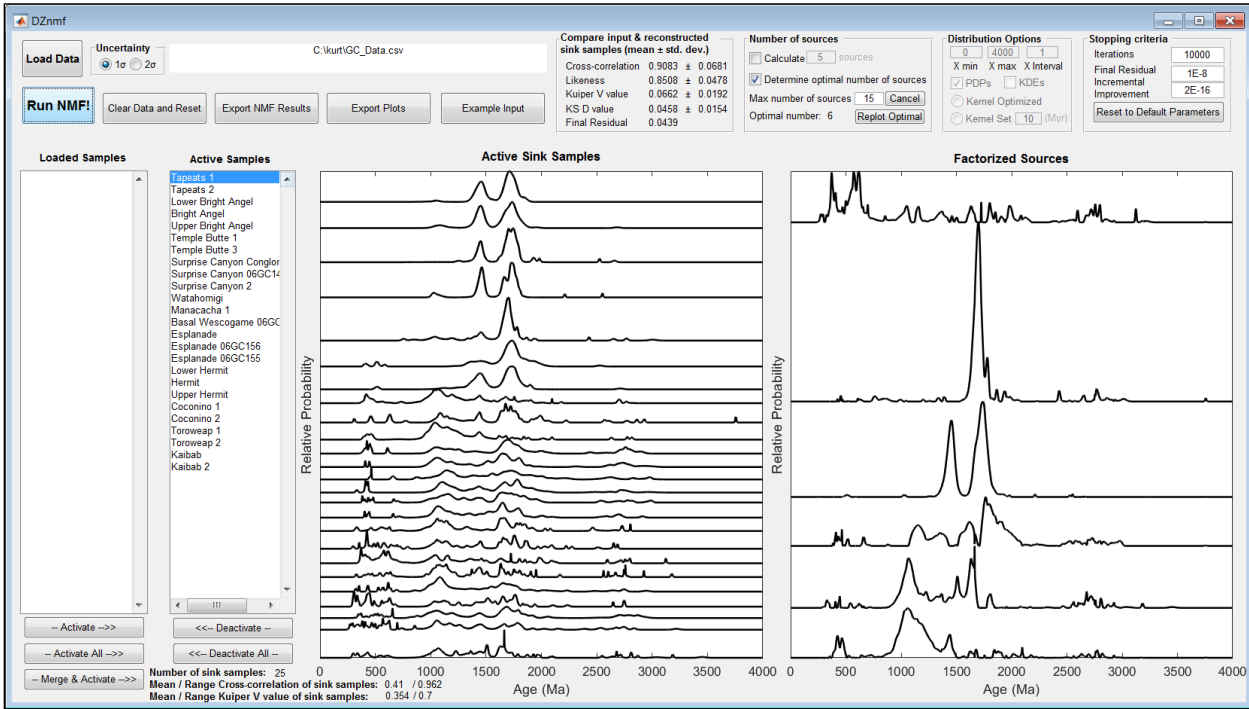

Figure 13. *Six factorized sources determined through optimization routine.*
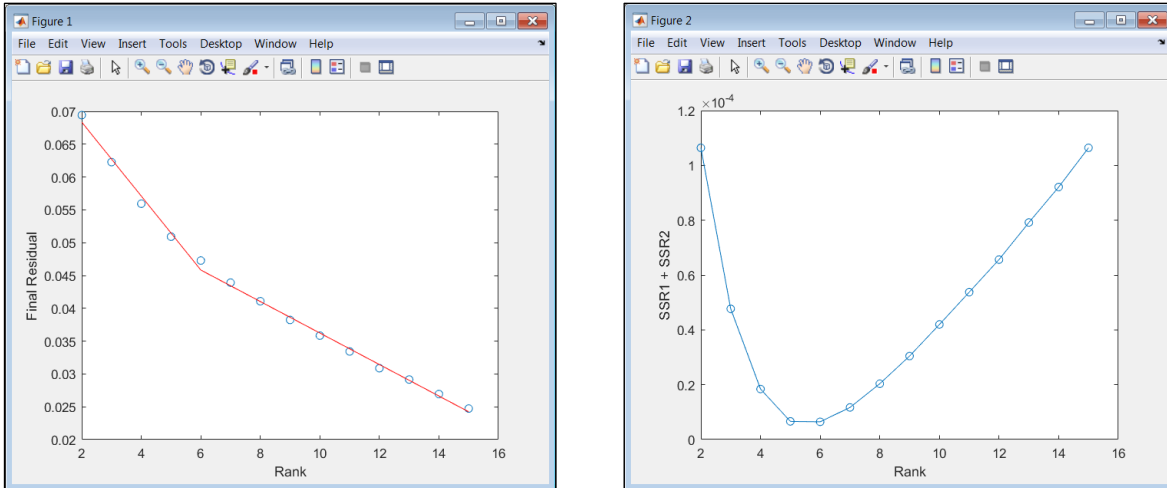
Figure 14. *Plots generated for determining the optimal number of sources to factorize. Left: Segmented linear regression plot. The break in slope is the optimal number of sources. Right: Summed sum of squared residuals. The nadir is the optimal number of sources.*

8. EXPORT PLOTS

All plots may be exported by pressing the **Export Plots** pushbutton. These figures pop up in new windows and may be saved as vector images to be modified in a drafting program such as Adobe Illustrator. To save a figure as a vector image select file → Save as (Figure 15). Choose either .eps or .emf format (Figure 15).


11
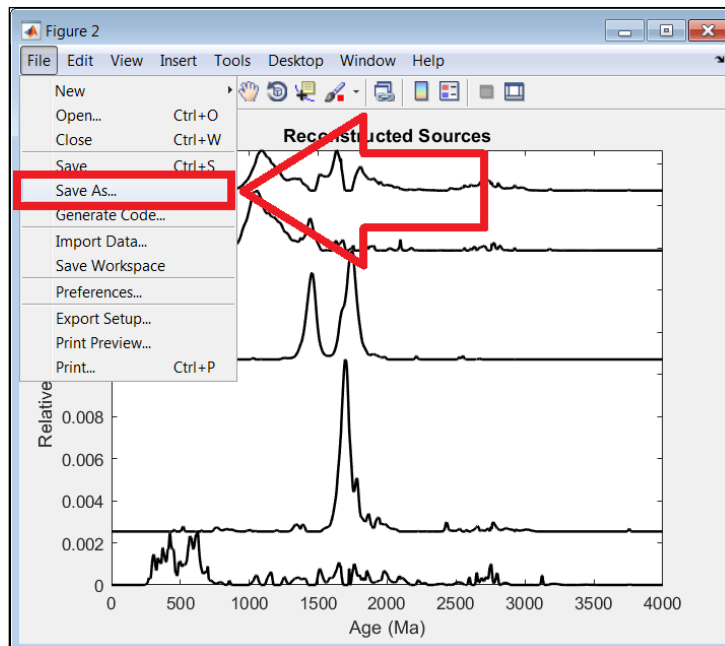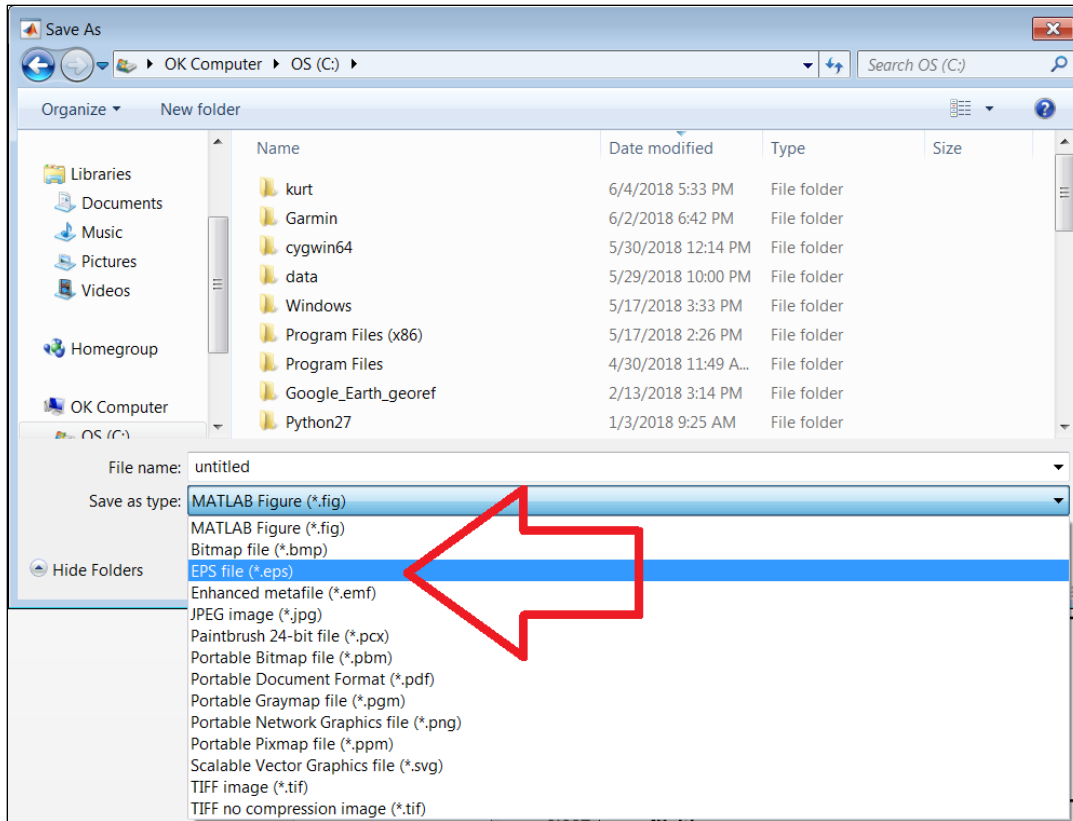
Figure 15. *Upper: Select Save As and choose .eps format (lower) to save as a vector image. Alternatively, .emf or .tif will save vector images that can be easily modified in a drafting program such as Adobe Illustrator.*
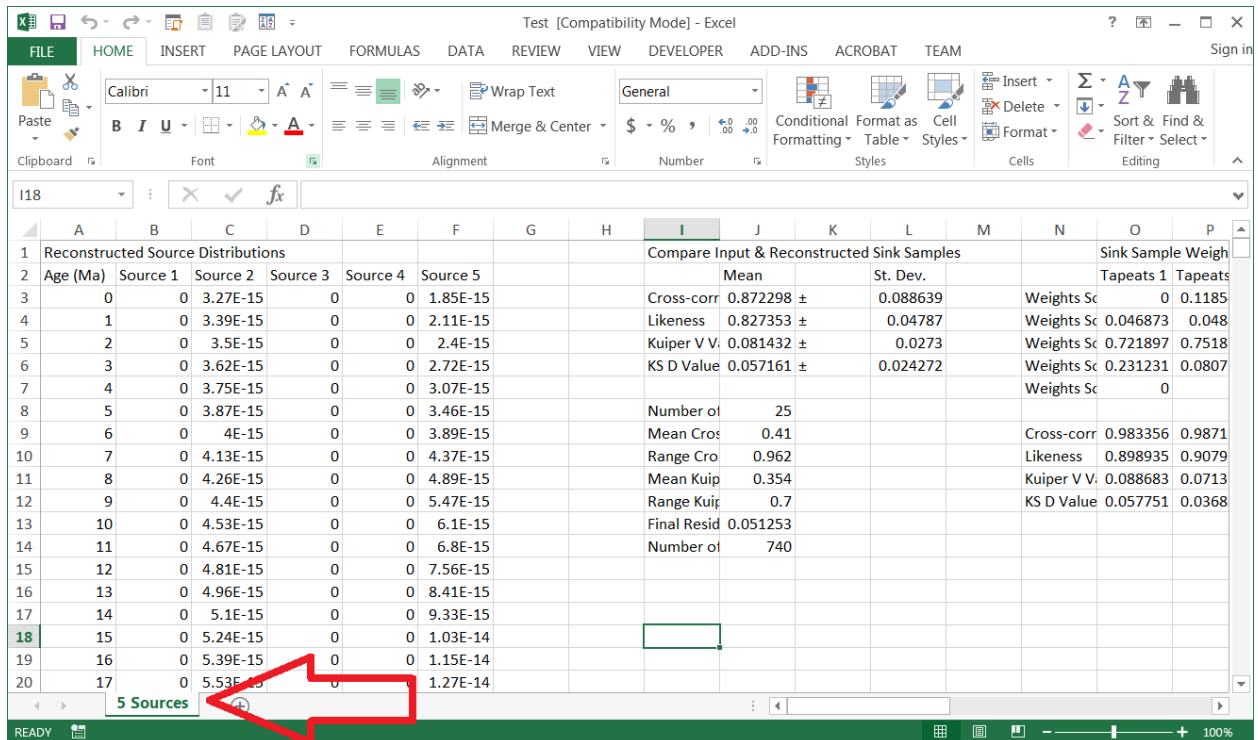
9. EXPORT RESULTS

To export the NMF results press the **Export Results** pushbutton. This will prompt a browser window for the user to name the file and select its destination (Figure 16). Results will be saved as a .xls file (Figure 17). If the optimization routine was used, all results from 2 to the maximum number of specified sources will be saved in individual sheets.
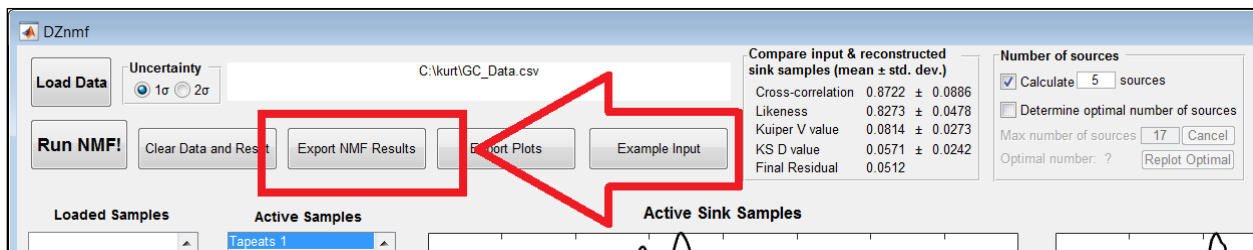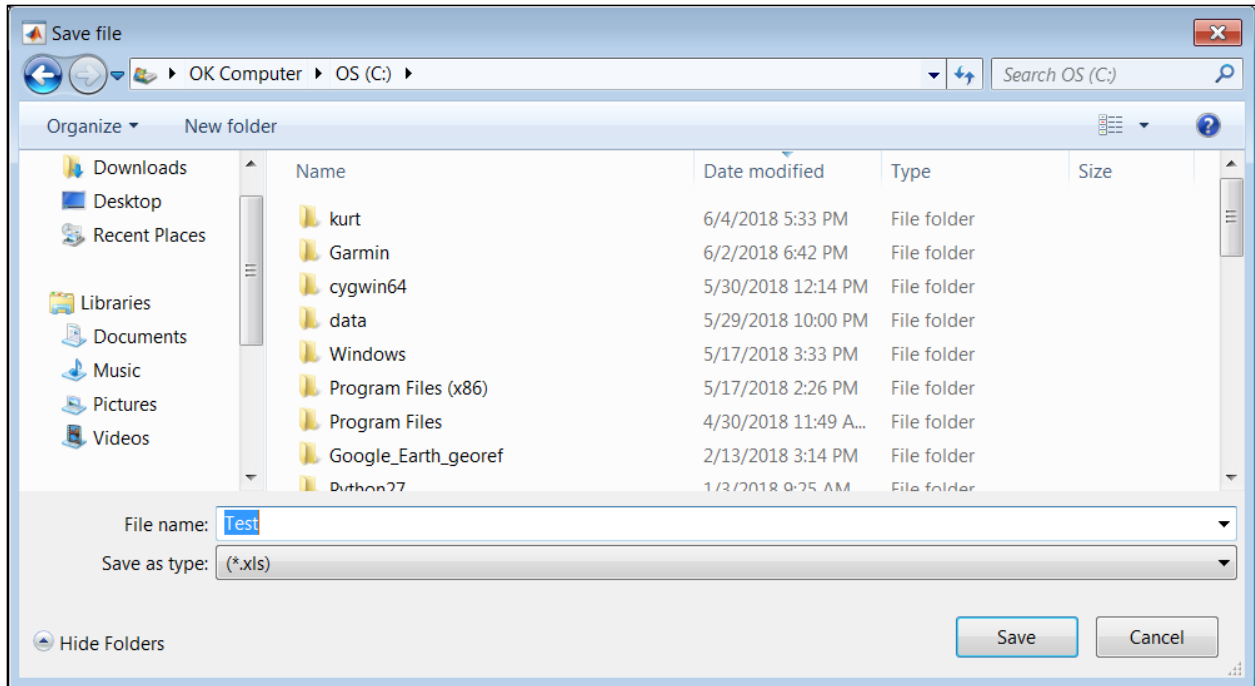
Figure 16. Upper: Save results by pressing **Export NMF Results** pushbutton. Lower: Browser window to name the results and specify the file location.

Figure 17. Example export saved as .xls file. Specifying the number of sources saves results into a single sheet (red arrow). If the optimization routine is used then results for each factorization (e.g., 2, 3, 4 sources) will be saved into separate sheets within the same file.

10. NONUNIQUE SOLUTIONS

The NMF model may result in nonunique solutions. This typically happens if there are a large number of sources, distributions may be factorized into individual components of those distributions (typically due to sediment recycling), or if the active samples are very similar. Hence we recommend that once the optimal number of sources is found, the user rerun NMF at that rank. In this case, if the NMF model does not converge on a solution within the specific number of iterations it will ask the user if they would like to repeat the calculation starting from a different random starting point in order to find the best solution.