A Matlab© toolbox for calculating spring indices from daily meteorological data

Toby R. Ault, Raul Zurita-Milla, and Mark D. Schwartz

Contents

1 Setting up ML-SI

The functions, data, and demos of `ML-SI` are included in a tarball online at: `http://ecrl.eas.cornell.edu/ml_si_v5.0.0.tar` .

Begin by downloading and untarring this package in some preferred location:

```
>tar -xvzf ml_si_v5.0.0.tar
>cd ML_SI
```

Next, run the routine `setup_ml_si.m`. This file (described below) can be edited to test various capabilities. For now, you should see something like:

```
SUCCESS!!! ml_si should be ready to run. The following options were tested:
(x) Successfully checked to make sure MDS verification data is available.

The following commands can be used to generate SI from
GHCN station "USC00405187":
 load ../data/select6.mat
 tmin(:,:,i)=conv_temp(USC00405187.TMIN.data,'C','F');
 tmax(:,:,i)=conv_temp(USC00405187.TMAX.data,'C','F');
 lat(i)=USC00405187.lat;
 [LFMTX,BLMTX,LSTFRZAllSites,LFpredAllSites,BLpredAllSites]=...
                          calc_si_ml_v1(tmin,tmax,lat);
```

You should also see a window pop up with plots from the "select 6" sites used for testing and validating the code (Figure 1). Data for this figure originate from the GHNCD, and their metadata is described in Table 1. Output from the `ml_si` routines is shown in green (Leaf Index) and red (Bloom Index). Output from the Fortran `SI-x` code (Schwartz et al., 2013) is overlaid in black and gray for the leaf and bloom indices, respectively. The very small mismatch during a few years reflects different treatments of missing data during preprocessing: running the `ml_si` routines on the *exact same* input as the Fortran code produces identical output.

1.1 Using "setup_ml_si.m"

The setup script `setup_ml_si.m` has several options allowing the user to test different "add-ons"to the core `SI_ML` toolkit. These are found in lines 7-51 along with in-line documentation for each component. Briefly, `setup_ml_si.m` allows the user to change the default directories where data is stored, as well as test the functionality of routines to get data remotely (using `wget`, Mac/Linux only), import the North

American Lilac data (Cayan et al., 2001; Schwartz and Reiter, 2000), and import GHCND metadata and data.

Table 1: Metadata for the "Select 6" stations used to validate code. Time series of the bloom and index values for these stations are shown in Figure 1.

Station ID	Station Name	State	Lat	Lon	Elev
USC00114442	JACKSONVILLE 2E	IL	39.7353	-90.2153	185.9
USC00118740	URBANA	IL	40.0839	-88.2403	219.8
USC00213290	GRAND MEADOW	MN	43.7047	-92.5644	411.5
USC00234705	LAMAR 2W	MO	37.4933	-94.3147	303.6
USC00315771	MONROE 2 SE	NC	34.9797	-80.5233	167.6
USC00405187	LEWISBURG EXP STN	TN	35.4139	-86.8086	239.9

1.2 List of files

The list below describes all directories and files included in the `ml_si` toolbox:

Directories	
`ML_SI`	Base directory containing all sub-directories.
`ML_SI/scripts/`	Directory of all scripts and demos used in the main text, as well as a few additional utilities to download and import data from various sources.
`ML_SI/si_funcs/`	Driver functions comprising the `ML_SI` toolkit. These are the only functions needed to calculate the spring indices (SI) described in (Schwartz et al., 2006) and extended SI (SI-x) to include lower-latitude regions in (Schwartz et al., 2013).
`ML_SI/data/`	Folder with all data needed for demos and code verification.
`ML_SI/data/` `mds_verification_data/`	Folder with verification data from original Fortran programs provided by M. D. Schwartz.
`ML_SI/data/`	

`ghcnd_all/`	Empty directory where GHCND could be stored. This dataset would need to be Downloaded from the NOAA ftp site individually: `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/` Or alternatively as one zipped file: `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd_all.tar.gz`
`ML_SI/figs/`	Default location for figure files to be stored as output from scripts.
`ML_SI/docs/`	Folder with this user's guide and C&G manuscript.
Functions	
`calc_si_ml`	Main driver function used to calculate spring indices.
`leaf`	Function called by `calc_si_ml` to handle most of the actual calculations.
`growdh`	Calculates growing degree hours (GDH) from `Tmin` `Tmax` and `latitude` inputs.
`calc_daylen`	Returns day lengths (in hours) from `calendar_day` and `latitude` inputs.
`calc_soldec`	Placeholder function to produce `solar_declination` values for a given `latitude` and `calendar_day`. Needed for `calc_daylen`.
`synval`	Function to calculate the occurrence of "high energy synoptic events" as described by Schwartz (1985).
`read_ghcnd_dly_file`	Reads `.dly` file, which is the format of the GHCN data supplied by the NOAA GHCN archive.
`PrevFig`	Helper function to resize figures so that they will appear on screen approximately in the way they will when printed to an `.eps` file.

`nancorr`	Calculates correlations between two series when missing values (`NaN`) are present. Requires Matlab's© *statistics toolbox*.
`ll2dist`	Calculates distances (in Km) between latitude/longitude pairs (derived from function `m_lldist` in the `m_map` toolkit: `http://www.eos.ubc.ca/~rich/map.html`).

Scripts and Demos

`setup_ml_si.m`	Sets up `ml_si` working directory, directory names, and data directories. Also checks code using the "Select 6" validation sites. There are several options that can be turned on or off to: (a) test the functionality of wget to retrieve station data if is installed (http://www.gnu.org/software/wget/); and (b) test the functions used to import station data. User will need to specify the correct executable for wget (`wget_cmmnd='/usr/local/bin/wget'`). The following routines can be called from within this script: `get_select6`, `mk_ghcnd_metadata`, `import_select6`. **Note:** None of the scripts called by `setup_ml_si` are required for the functions that calculate spring indices to work properly. Settings are saved in a file called `si_paths.mat` in the directory `ML-SI/scripts/`
`mk_ghcnd_metadata.m`	Creates a Matlab© structure from the list of GHCN daily station metadata (e.g., lat, lon, etc...), supplied online at: `ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-inventory.txt`.
`get_select6.m`	Short routine to obtain the "select 6" records of daily Tmin and Tmax from the NOAA ftp repository of GHCN data (available online at: `ftp.ncdc.noaa.gov/pub/data/ghcn/daily/`).

`import_select6.m`	This script will read through a list of stations (`select6_stn_ids`, set in `setup_ml_si`), use `wget` to retrieve them from the NOAA ftp site, and save them to a matlab file (`select6.mat`). This file comes with the `ml_si` distribution, so you shouldn't need to run this script unless you want to.
`import_lilac.m`	Routine to import and save *observational* lilac phenological data collections described in Schwartz and Reiter (2000) and Cayan et al. (2001), and archived online by the USA National Phenology Network (USA-NPN: https://www.usanpn.org/). Sources files for these datasets are also available as part of this distribution (see "Datasets" below). These data include observer based records of first bloom and first leaf dates from thousands of sites across North America for two species of lilac. Data will be saved as a self-describing structure called `lilac` in the file `ML-SI/data/Schwartz-Caprio`
`gdh_demo.m`	An illustration of how the "growing degree hour" calculation is made using the function `growdh`.
`si_demo_1.m`	Example illustrating how each of the predictors comprising the *leaf index* and *bloom index* are generated using the functions in `ML_SI/si_funcs/`.
`si_demo_2.m`	Calculation of SI from a single station. The station ID is specified in the top and could be changed to any other one in the GHCND.
`si_demo_3.m`	Example illustrating how `ml_si` can be used to relate station-based values of the Leaf and Bloom indices to observational phenology. Plotting is handled in the end by `mk_data_vs_index_fig.m`, which allows the program to run in the background (e.g., by calling Matlab© from a shell script and starting the routine).
`mk_data_vs_index_fig.m`	Script to plot the spring indices from `si_demo_3.m` against the nearest phenological observations (provided by the USA-NPN: https://www.usanpn.org/).

si_demo_4.m	Script to plot the trends and means of the data calculated in si_demo_3.m.

Datasets

select6.mat	Data for the "select 6" verification stations used to test ml_si. Data from each station is saved as a *structure*. This file may be generated by first obtaining the GHCN daily data for each site (e.g., by running get_select6.m), then running import_select6.m.
ghcnd_metadata.mat	Metadata for the GHCND network, generated from the file ghcnd-stations.txt using the script mk_ghcnd_metadata.m. GHCN metadata is saved as a *structure* with information about each site.
Schwartz-Caprio.mat	Observational phenological data for three species of lilac. The raw data are available online through the USA-NPN webpages (https://www.usanpn.org/), and are also distributed as part of this package (files SC_SV_Lilac_1961_2008.xls and Lilac_SV_westUSA_1956_2009.xls). These Excel© files were imported using the script import_lilac.m. Data are saved as a *structure* called lilac.
Schwartz-Caprio-withSI.mat	Paired lilac observational data (from Schwartz-Caprio.mat, see above) and spring indices calculated from the nearest GHCN sites. This data is saved as a *structure* called lilac, and is generated using si_demo_3.m.
ghcnd-stations.txt	File of GHCN daily station metadata, obtained from NOAA ftp site (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt).

`SC_SV_Lilac_1961_2008.xls`	Observational lilac phenology for two plant cultivars (*Syringa Chinensis* and *Syringa vulgaris*) from the Eastern US, obtained from the USA-NPN (https://www.usanpn.org/) National coordinating office (NCO). The dataset is described in Schwartz and Reiter (2000).
	Lilac phenology for one plant cultivar (*Syringa vulgaris*) from the Eastern US, also archived online by the USA-NPN (https://www.usanpn.org/) National coordinating office (NCO). The dataset is described in Cayan et al. (2001).
`SI-x_1981_2010_\` `norms25_noWBAN.xls`	List of 733 high-quality GHCND sites along with some metadata and the `SI-x` averages from 1981-2010 from M. D. Schwartz's FORTRAN code. The criterion for selecting sites are documented in Schwartz et al. (2006).

Parameter files
(in `ML_SI/scripts/`)

`si_paths.mat`	File that is *created* by `setup_ml_si.m`. Contents include path and variable names that will be used by the various demos. This file won't exist until `setup_ml_si.m` has been run.
`axis_settings.mat`	Some axis settings to customize figures.
`custom_linecolors.mat`	Customized line colors for figures.

Documents

`si-ml_v3.pdf`	C&G manuscript.
`users_guide.pdf`	This document.

2 Running and modifying the demo scripts

2.1 gdh_demo.m

A simple line plot of hourly temperatures, modeled from T_{min} and T_{max} values, is generated for an idealized case. The parameters of this idealized case are specified on lines 4-7:

```
MN=20; % Tmin value (Deg F)
MX=60; % Tmax value (Deg F)
DOY=75; % Day of year (75=March 17th)
lat=45; % Latitude
```

Modifying any of these parameters will change the profile of 24 hour temperature profile (generated from Tmin/Tmax), and hence the value for GDH.

2.2 si_demo_1.m

This demo plots the accumulation of raw T_{min} and T_{max} for a single year from a single station (USC00114442, in this case), along with the derived predictor variables for both the leaf and bloom Indices. For the leaf index, these predictors are growing degree hours (GDH), number of "synoptic events" (SYNOP), accumulation of GDH in the 2-3 previous days (DDE2) and accumulations of GDH in the 5-7 days prior (DD57). The station and year for which these diagrams can be changed on lines 11 and 12:

```
stn_nums=[114442]; % Station #
yrq=47; % index of year (47 = 1896 AD)
```

2.3 si_demo_2.m

Script to calculate SI from all years for a single station. The station choice can be made on line 9:

```
stn_nums=[405187];
```

Or it could be imported from the GHCND. However, a different set of pre-processing steps (lines 11-22) would be needed to extract the latitude and time for this site, as well as convert the T_{max} and T_{min} values to Fahrenheit.

2.4 si_demo_3.m

Example illustrating how `ml_si` can be used to relate station-based values of the Leaf and Bloom indices to observational phenology. In this case, the phenology dataset overlaps with the data used to develop the model, so this the good correlation shown in the main figure is in some sense built in. The motivation for providing this example, however, is just to illustrate how to import a large number of GHCN station data then plot them against their nearest observational neighbors. The plotting is handled by `mk_data_vs_index_fig.m`, so that if `si_demo_3.m` has already been run, it is easy to regenerate the plot using different settings. For instance, one could change the variable `maxdist` variable (default value is 10km) to generate a scatter plot with more distant GHCND sites plotted against phenological observations.

Note: This program takes about an hour to run with 733 sites.

2.5 si_demo_4.m

This program plots the means and trends of the leaf and bloom indices calculated in `si_demo_3.m`.

3 Downloading Data

In addition to the datasets described in Table 1 that come with the `ml_si` distribution, running some of the demos will require GHCN data to be downloaded and stored in `ML_SI/data/ghcnd_all/` by default (the path to the GHCN data can be modified in `setup_ml_si.m`). The entire data set can be downloaded as a zipped tarball (`ghcnd_all.tar.gz`) from the NOAA ftp site[1]: `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/`.

4 History of Code

The original code was created to implement a model of Syringa chinensis Red Rothomagensis (cloned lilac) first leaf dates in 1984-1985 (Schwartz, 1985). The model/code underwent several improvements, including adding "synoptic capstone events" in the late 1980s (Schwartz and Marotz, 1986, 1988). In 1990, the first spring index was developed for the first leaf event, by averaging output dates from the

[1]Tested 11-4-2013

cloned lilac model, plus similar models for two cloned honeysuckles, Lonicera tatarica (Arnold Red) and L. korolkowii (Zabeli) (Schwartz and Karl, 1990; Schwartz, 1993). At this point the models all used a January 1st start date, and had no consideration of chilling accumulation.

In 1995-1996, the models underwent major revisions (Schwartz et al., 1997): 1) chilling hour requirements (based on accumulations from 1 October of the previous year) was implemented prior to initiation of the first leaf models, and 2) a first bloom model was added (again, the average of outputs of the three plants bloom models), thus rendering the plural "Spring Indices" appropriate (though the abbreviation, "SI" remained the same).

Up until this point the code was only built to handle input from locations in the eastern and central USA. In 1998, the code was made more flexible, so any location in the northern hemisphere could be used, provided it had daily maximum-minimum shelter-height air temperature data available, and the latitude was provided.

Between 1998 and 2011 the code underwent only a few minimal changes to optimize performance. However, in 2011 the code was modified to "turn off" the chilling requirements, and return to starting calculations for the first leaf model on January 1st. This new version was termed "extended Spring Indices" or SI-x, as it allowed output to be produced farther south into the sub-tropics than the earlier model version, now termed "Spring Indices Original" or SI-o (Schwartz et al., 2013). Between 2011 and May 2013, additional modifications in the code were made to optimize performance. These modifications may introduce changes in individual station-year outputs, but are not significantly different overall from earlier SI-x output overall (i.e., differences average zero and are uncorrelated).

The Fortran code served as the basis for our Matlab©implementation of SI-x. We translated the routines as literally as possible, making this distribution look similar in structure and syntax to the original Fortran. This choice, as opposed to a complete "top down" re-write in Matlab©, preserves the basic heritage of the routine and makes going between the two platforms relatively straightforward. Contact M.D. Schwartz directly for copies of the original Fortran.

5 Additional Notes

By default, we have used a base temperature of 31°F (-0.5556°C) to calculate growing degree hours. Moreover, the routines are currently set to only compute the **X** predictor variables to model leaf/bloom events out through day 240 (August 27th or 28th on leap/non-leap years). While fundamentally arbitrary, this makes the routines somewhat more efficient when one runs into years/regions that remain too cool to ever trigger meaningful leaf-out dates. This threshold could be modified, however, if there were good ecological/physical reasons to do so by modifying the value of daystop in `calc_si.m` and daymax in `leaf.m`.

References

Cayan, D. R., S. A. Kammerdiener, M. D. Dettinger, J. M. Caprio, and D. H. Peterson, 2001: Changes in the onset of spring in the western United States. *Bulletin of the American Meteorological Society*, **82 (3)**, 399–415.

Schwartz, M., 1985: The advance of phenological spring across Eastern and Central North America. Ph.D. thesis, University of Kansas.

Schwartz, M. and G. Marotz, 1986: An approach to examining regional atmosphere plant interactions with phenological data. *Journal of Biogeography*, **13 (6)**, 551–560.

Schwartz, M. D., 1993: Assessing the Onset of Spring: A Climatological Perspective. *Physical Geography*, **14 (6)**, 536–550.

Schwartz, M. D., R. Ahas, and A. Aasa, 2006: Onset of spring starting earlier across the Northern Hemisphere. *Global Change Biology*, **12 (2)**, 343–351.

Schwartz, M. D., T. R. Ault, and J. L. Betancourt, 2013: Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *International Journal of Climatology*, doi: 10.1002/joc.3625.

Schwartz, M. D., G. J. Carbone, G. L. Reighard, and W. R. Okie, 1997: A model to predict peach phenology and maturity using meteorological variables. *Hortscience*, **32 (2)**, 213–216.

Schwartz, M. D. and T. R. Karl, 1990: Spring Phenology - Natures Experiment to Detect the Effect of Green-up on Surface Maximum Temperatures. *Monthly Weather Review*, **118 (4)**, 883–890.

Schwartz, M. D. and G. A. Marotz, 1988: Synoptic Events and Spring Phenology. *Physical Geography*, **9 (2)**, 151–161.

Schwartz, M. D. and B. E. Reiter, 2000: Changes in North American spring. *International Journal of Climatology*, **20 (8)**, 929–932.

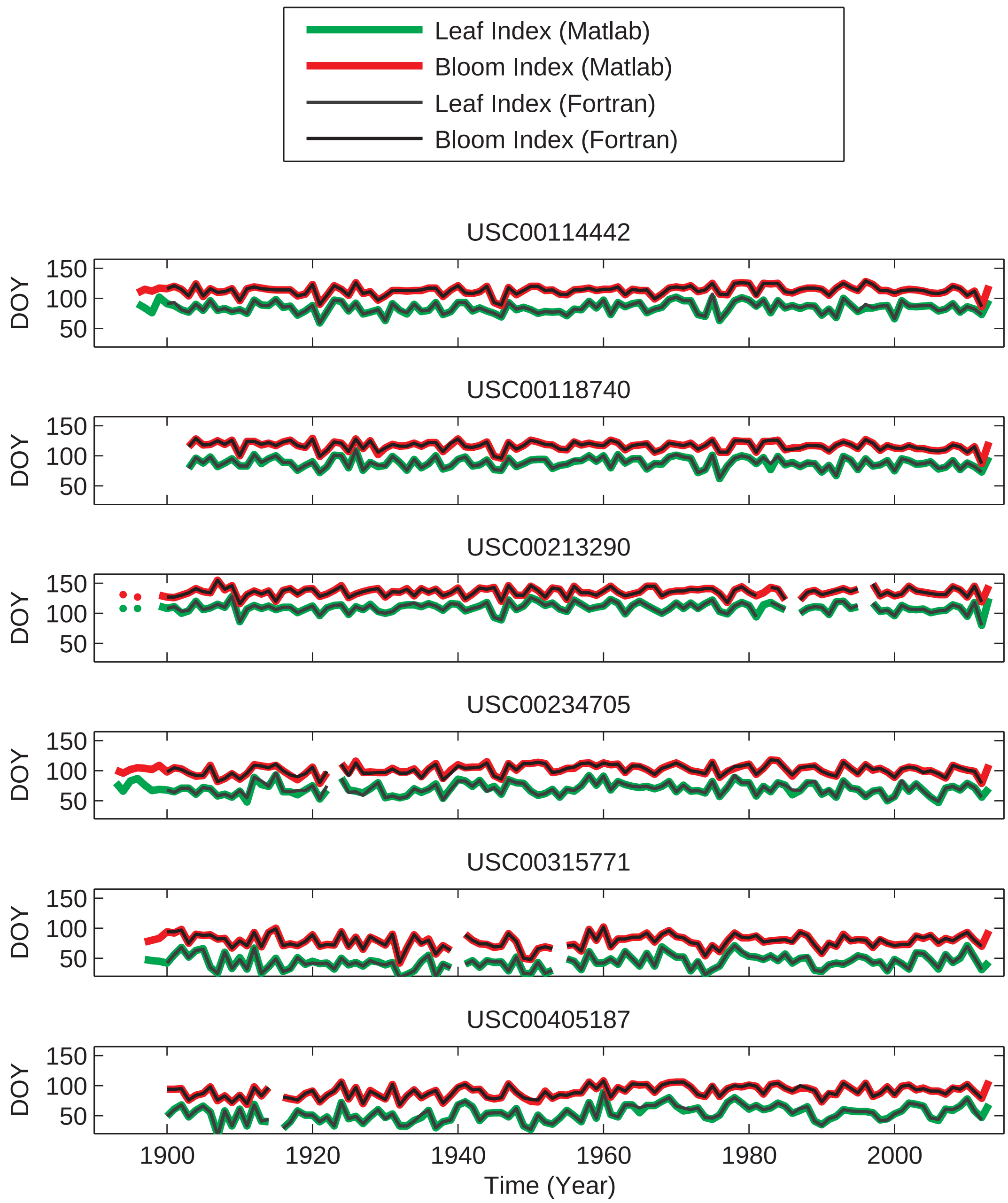

Figure 1: Time series of leaf index values for the “select 6” stations.