Paper 2462-2018

Using SAS® to Fit AmeriFlux Data to Ecosystem Seasonality Models

Tracy Song-Brink, North Carolina State University

ABSTRACT

In ecosystem science research, we have several models to define season transitions in ecosystem gross productivity (GEP) and respiration (ER). These models were built with ecosystem data collected years ago. Thanks to the AmeriFlux ecosystem data community, now we have access to ecosystem data from more than 110 sites located across the Americas, compared with 15 sites in 1997. The purpose of this project is to fit the large volume of data that was not available to previous research to our existing models for model evaluation. We used the NLIN procedure for model fitting for each variable of one year at one specific flux data source. For each model fitting process, we used SAS macros to perform Grubbs' test for outlier detection and removal, and the GPLOT procedure for data visualization. SAS® macros were written to automate the process of all input files, variables, and data years. Data from 132 input files with an average size of 4000 observations and 70 variables were processed, and two models were evaluated in this project.

INTRODUCTION

BACKGROUND

The seasonal cycle of plant community photosynthesis is one of the most important biotic oscillations. We have previously built two models to describe the seasonal transitions and dynamic characteristics. This modeling framework was built upon data collected from seven eddy covariance flux sites.

Model 1:

The seasonal cycle of plant community photosynthesis is described by the temporal variation of the canopy photosynthetic capacity (CPC). The CPC is defined as the maximal gross photosynthetic rate at the canopy level when the environmental conditions (e.g. light, moisture, and temperature) are non-limiting for the time of a year under consideration.

The CPC forms the boundary line that can be adequately represented by the following composite function:

$$A(t) = y_0 + \frac{a_1}{[1+\exp(-\frac{t-t_{01}}{b_1})]^{c_1}} - \frac{a_2}{[1+\exp(-\frac{t-t_{02}}{b_2})]^{c_1}}$$

where $A(t)$ is the CPC in day t; y_0, a_1, a_2, b_1, b_2, c_1, c_2, t_{01}, and t_{02} are empirical parameters to be estimated.

Model 2:

In this model, we use daily flux totals for both ecosystem gross productivity (GEP) and respiration (ER) to present seasonal transitions.

$$y = y_0 + \beta_1 \left[1 - e^{-\left(\frac{x - x_0 + \beta_2 \ln(2)^{\frac{1}{\beta_3}}}{\beta_2} \right)^{\beta_3}} \right]$$

where y is the daily integral of the flux of interest, y_0 is the base-value of y during the dormant season, x is day-of-year (DOY) for the first half of the year, and days until the end of the year for the second half of the year, $x_0$is the DOY at half maximum y (fitted), β_1 is the difference between peak and base y , β_2 is the difference between 75th and 25th percentiles of the time from base to peak y, and β_3 is a shape parameter. Parameters y_0, x_0, β_1 , $\beta_2$and β_3 are estimated in model fitting.

PURPOSE OF THIS PROJECT

The previous research used data from seven eddy covariance flux sites. With the collective effort from AmeriFlux community, now we have access to data from more than 110 sites. The purpose of this project is to fit all the data that are currently available to our models for evaluation and, if possible, to improve the current framework.

METHODS

TOOLS USED

We used SAS® software for the programing. We wrote several macros to control the program flow, read in data files in batch, process grouped data, and perform Grubb's test iterations for outlier detection and removal. We used NLIN procedures for model fitting, GPLOT procedures for visualization and REG procedures to generate metrics for model fitting evaluation. MEANS and FREQ procedures were also used for general statistical analysis.

AMERIFLUX DATA

We used AmeriFlux data that are downloadable from AmeriFlux site (http://ameriflux.lbl.gov).

AmeriFlux is a network of PI-managed sites measuring ecosystem CO2, water, and energy fluxes in North, Central and South America. It was established to connect research on field sites representing major climate and ecological biomes, including tundra, grasslands, savanna, crops, and conifer, deciduous, and tropical forests. The network was launched in 1996. The network grew from about 15 sites in 1997 to more than 110 active sites registered today.

AmeriFlux is now one of the DOE Office of Biological and Environmental Research's (BER) best-known and most highly regarded brands in climate and ecological research. AmeriFlux datasets, and the understanding derived from them, provide crucial linkages between terrestrial ecosystem processes and climate-relevant responses at landscape, regional, and continental scales.

In this project, we had 132 input files in .csv format. Input file size varies. An average input file has 4000 observations and 70 variables.

PROGRAM FLOW DESIGN

We used a macro (P0) to control the overall batch process flow. P0 generates a list of all files in the input file directory and processes through all the files in an alphabetic order. For each file, six SAS programs are executed to fit two models and calculate metrics for evaluation. When we fit one model, we process all variables of interest, and fit the model for the variable for all data years one year at a time. Model fitting is an iterative process based on Grubb's Test result to remove outliers detected.

All NLIN results including parameter estimates and standard errors and GPLOT results are saved in datasets and PDF files for metrics calculation and evaluation.

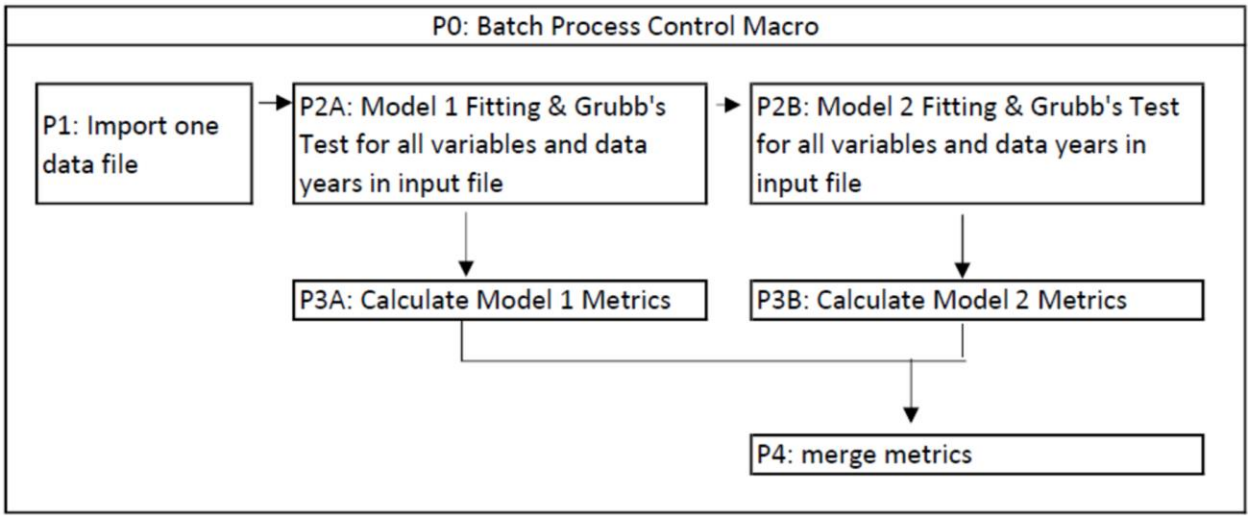


Figure 1. Overall Program Flow

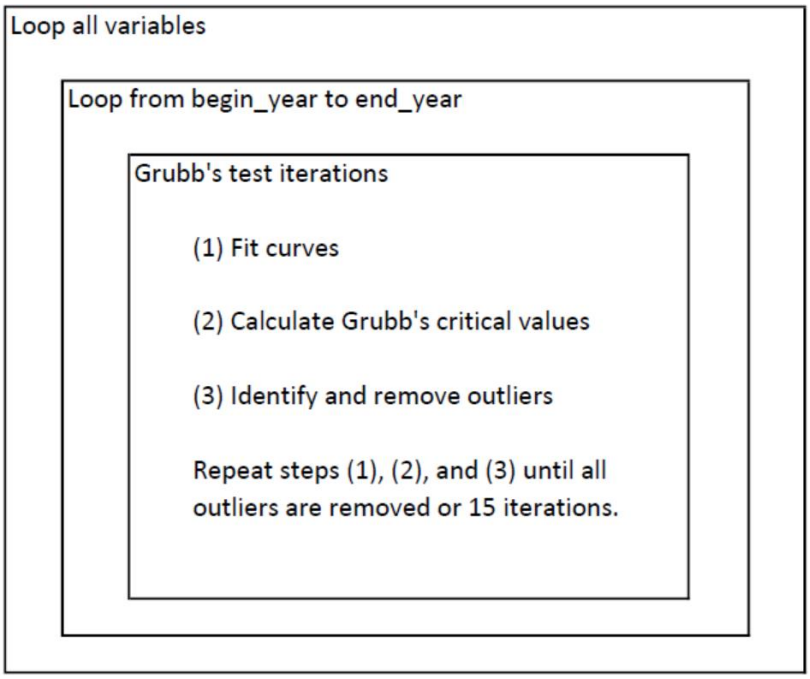
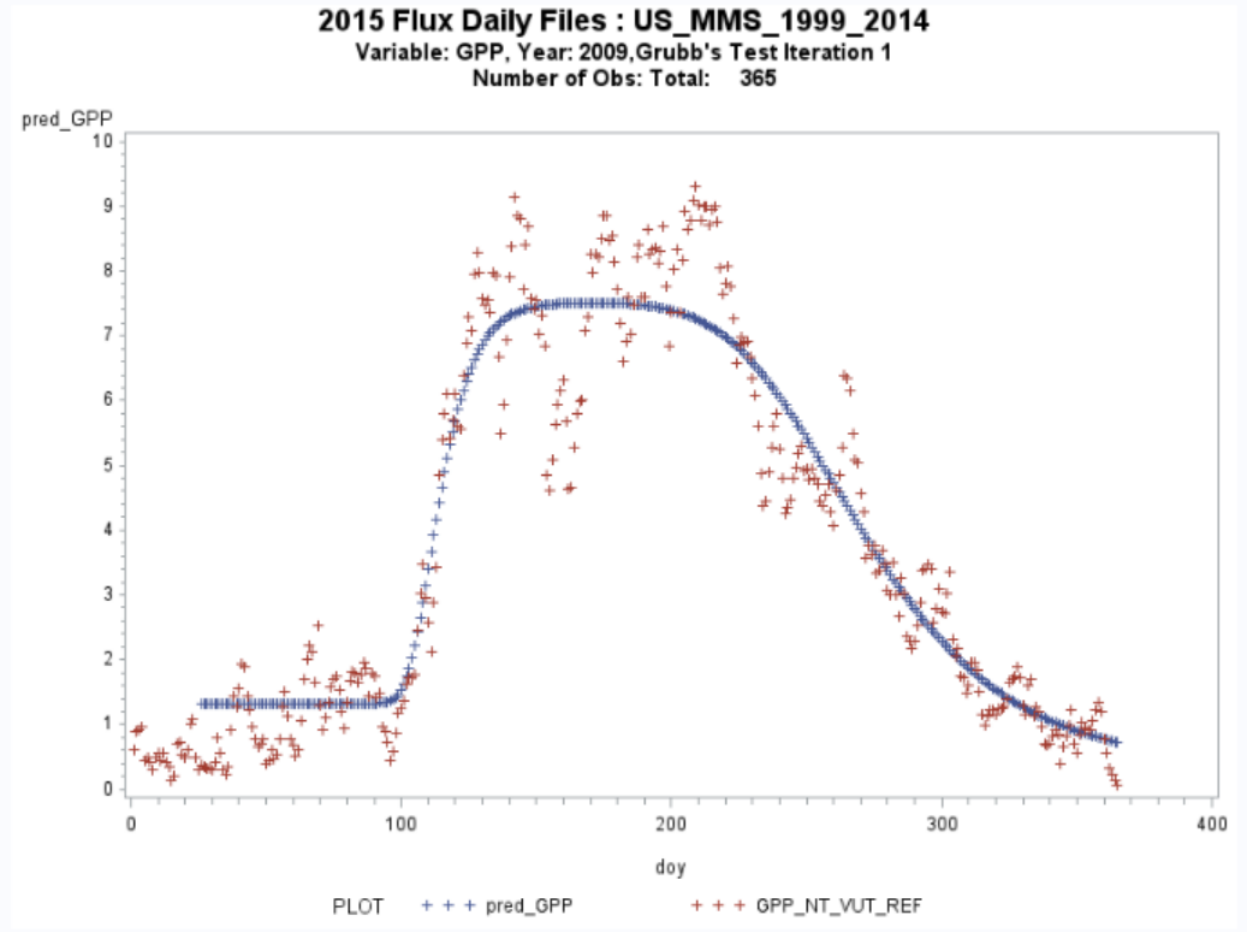


Figure 2. Model Fitting Programs (P2A and P2B) Process Flow

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	6	2539.7	423.3	571.76	<.0001
Error	332	245.8	0.7403		
Corrected Total	338	2785.4			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		Label
y0	1.3304	0.1012	1.1314	1.5294	
a1	6.1892	0.1597	5.8750	6.5033	
a2	7.1206	0.3183	6.4945	7.7466	
b1	8.7431	1.1635	6.4542	11.0320	
b2	34.3249	3.2019	28.0263	40.6235	
t01	67.0364	6.1077	55.0217	79.0510	
t02	200.0	0	200.0	200.0	
c1	150.0	0	150.0	150.0	
c2	5.8165	0.8223	4.1990	7.4340	
Bound3	0.0175	0.000054	0.0174	0.0176	150 <= c1
Bound4	0.0246	0.00616	0.0126	0.0367	t02 <= 200

Output 1. Output from an NLIN Procedure for Model 1



Output 2. Output from a GPLOT Procedure

MODEL EVALUATION

A group of metrics were calculated for each model fitted along with standard errors from the NLIN procedure results. Metrics mainly include first and second derivatives of raw data, compared with those of predicted values for spring and fall seasons. We also used the GPLOT results to determine if the model fitting result was valid.

CONCLUSION

In about five days' continuous program execution on a Windows server, all data were successfully processed, and model fitting results were saved and evaluated automatically by our programs. Both models that we fitted showed good performance where more than 50% of the results were valid for assessments. With the quality of the data we obtained from AmeriFlux, this has met our expectation. Comparison of the metrics showed that Model 1 fit the data better.

In institutional research of many areas, researchers are now having much more data collected with emerging technologies. To use newly collected data to improve existing models, we are interested in efficiently processing large amount of data to fit existing models. This project provided an approach to perform repetitive model fitting over large amount data.

REFERENCES

AmeriFlux website, https://ameriflux.lbl.gov/

Noormets, A., Chen, J., Gu, L., Desai, A. 2009*. The Phenology of Gross Ecosystem Productivity and Ecosystem Respiration in Temperate Hardwood and Conifer Chronosequences.* New York, NY: Springer

Gu, L., Post, W. , Baldocchi, D., Black, T., Suyker, A., Verma, A., Vesala, T., Wofsy, S. 2009. *Characterizing the Seasonal Dynamics of Plant Community Photosynthesis Across a Range of Vegetation Types.* New York, NY: Springer