Optimizing Genetic Algorithm Parameters for Atmospheric Carbon Monoxide Modeling

by Meera Duggal, Dorit Hammerling

Peer Reviewed

In the Southern Hemisphere the primary source of Carbon Monoxide (CO) variability is large burn events (Holloway, Levy II, & Kasibhatla, 2000; Edwards et al., 2006). Thus, atmospheric CO can be used as a proxy for fires (Buchholz et al., 2018). This introduces many benefits as it would give governments time to take the necessary measures to prepare. Fires are directly related to dryness of vegetation and lack of water. Thus, they can be partially summarized with something called climate indices. Climate indices can measure the variability in the climate such as changes in sea surface temperature and wind. With this knowledge a predictive CO model was developed (Buchholz et al., 2018). This model uses climate indices to predict future CO emissions which are directly linked to large burn events that will occur in the southern hemisphere. We use four different climate indices in our model: the El Niño Southern Oscillation in the central tropical Pacific region 3.4, Indian Ocean Dipole/ Dipole Mode Index (IOD/DMI), Tropical South Atlantic Index (TSA), and the Antartic Oscillation Index (AAO). The multiple linear regression model that was developed is as follows,

$$CO(t) = \mu + \sum_{k} a_k \cdot \chi_k(t - \tau_k) + \sum_{i,j} b_{i,j} \cdot (t - \tau_i) \cdot \chi_j(t - \tau_j)$$

Equation 1

$CO(t)$ is the CO anomaly at time t for a specific region in the Southern Hemispher, χ are the climate indices, τ is the lag value for each index in months, μ is the constant mean displacement, a_k and b_{ij} are coefficients, and the subscript k represents the climate indices with a numeric value.

A package called *regClimateChem* was created in order to perform different variable selection techniques on the model, shown in Equation 1 (Daniels, Hammerling, & Buchholz, 2000). Variable selection techniques are designed to select the best predictors in a model. The different variable selection techniques that were implemented in *regClimateChem* are the exhaustive method, stepwise regression, and the genetic algorithm. Both the exhaustive method and the genetic algorithm were implemented via the *glmulti* package in R (Venables & Ripley, 2002).

We needed a way to compare the different models outputted from these different variable selection techniques so we chose the Bayesian Information Criterion (BIC). The BIC was chosen because it is good to use when assessing predictive models and it penalizes the number of terms in your selected model. Thus, the BIC gives models that do not overfit the data and are good for prediction. The lower the BIC the better the model.

The exhaustive method goes through all possible models and outputs the best one, according to the specified criterion (in our case the BIC). Stepwise regression is an iterative process that begins either with a full model or an empty one. With each step it looks at either adding or removing a predictor depending on if it will improve the model. The algorithm ends once it has found the best model. The exhaustive method always finds the best model but often times has a long runtime, whereas stepwise regression has a low runtime and often times fails to find the best model. Thus, the genetic algorithm was implemented in *regClimateChem* as a compromise between computational expense and model accuracy. An optimization study was performed on the genetic algorithm in hopes that we could reduce the runtime while still producing good models.

The genetic algorithm is a stochastic algorithm and therefore it is based on probability. The algorithm begins by randomly selecting a population of models to begin the iterative process. These models are then randomly modified using three different model modification techniques to create a new generation. The population size is preserved with each new generation that is produced. New generations are produced, continuously improving the models, until a stopping criterion is met. The genetic algorithm is designed to produce good models by using the best models from the previous generation to produce the next generation.

In the genetic algorithm a parent model represents a model in the current generation and a child model is a model in the next generation. We will now describe the different model modification techniques that are used to create the next generation.

One of the model modification techniques is called **asexual reproduction**. With this process a single term in each model can mutate at a given rate called the mutation rate. If a given term mutates and it was included in the model then it will not be included and vice versa. The second model modification technique is called **sexual reproduction**. In sexual reproduction there are two parent models that come together to make a child model that is included in the next generation. The last model modification technique that is included is **immigration**. When a model is created using immigration each term has a 50% chance of being included or not. Two parameters determine which models go through which model modification technique. These are called the immigration rate (*imm*) and the sexual reproduction rate (*sexrate*). The immigration rate is the ratio of models that go through immigration to the number of models that go through asexual reproduction. The sexual reproduction rate is the ratio of models that go through sexual reproduction to the number of models that go through asexual reproduction.

The genetic algorithm also uses a stopping criterion to determine when it should stop producing new generations. This stopping criterion is checked every 20 generations. This is composed of three parameters: *deltaM*, *deltaB*, and a consecutive rate (*conseq*). *DeltaM* is the target change in the mean information criterion. *DeltaB* is the target change in the best information criterion. The consecutive rate is the number of iterations that the genetic algorithm will go through once both *deltaM* and *deltaB* have been met. A flow of the genetic algorithm, as described above, can be seen in Figure 1.

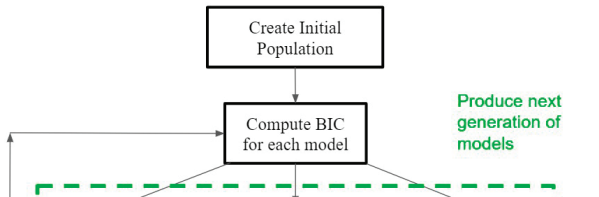


Figure 1: A flow of the genetic algorithm

Our optimization study consisted of varying different parameters in the genetic algorithm via *glmulti* to find the parameter combination that minimizes runtime but maintains model accuracy. Runtime refers to the total time it takes for the genetic algorithm to reach the stopping criterion. Model accuracy refers to the ability of the genetic algorithm to reproduce the models produced by the exhaustive method as the exhaustive method always produces the best models. In order to quantify model accuracy we look at something called the proportion of differences. This is the proportion of models produced from the genetic algorithm that differ from the exhaustive search over the total number of models. The *glmulti* parameters that were varied in this study are seen in the following table.

Table 1: Parameters varied for the study. The parameter value, default, and the range of values we studied can be seen.

Parameter	Default	Values Studied
Population Size	100	5,20,40,60,80,100
Mutation Rate	0.001	10^{-5},0.001,0.2
Sexual Reproduction Rate	0.1	0.001,0.1,0.7
Immigration Rate	0.3	0.001,0.3,0.7
Consecutive Iterations	5	1,2,3,4,5

We started by optimizing the parameters for a four covariate model (a model with the four climate indices, no interaction terms). In this study one parameter was varied one at a time while the others were held constant. An example of our results can be seen with the population size.

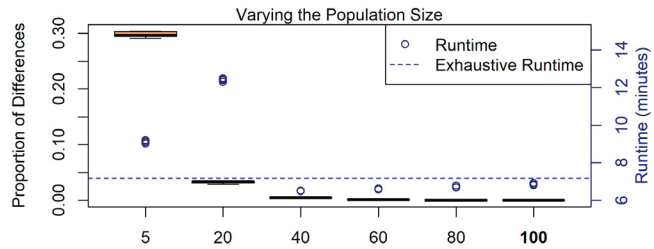


Figure 2: Default value for the population size is seen bolded on the x-axis. Proportion of differences between the genetic algorithm and exhaustive search is represented as boxplots and is based on the left y-axis. Right y-axis shows the runtime of the genetic algorithm which correspond to the blue circles.

In Figure 2, the boxplots represent the proportion of differences, which was calculated with each individual population size value. These boxplots correspond to the left y-axis. The blue dots represents the runtime, which is how long it took for the genetic algorithm to converge with that specific population size value. The blue dots correspond to the right y-axis. Since we are looking to minimize both runtime and proportion of differences we would choose a population size of 40 to be our optimal value. Plots like this were made and analyzed for the four other parameters studied.

After studying each parameter on its own we chose the optimal values for the four covariate case to be:

```
Population Size: 40
Mutation Rate: 0.2
Sexual Reproduction Rate: 0.1
Immigration Rate: 0.001
Consecutive Rate: 2
```

We now compare the proportion of differences and runtime from the default *glmulti* parameters and the optimized values that we chose.

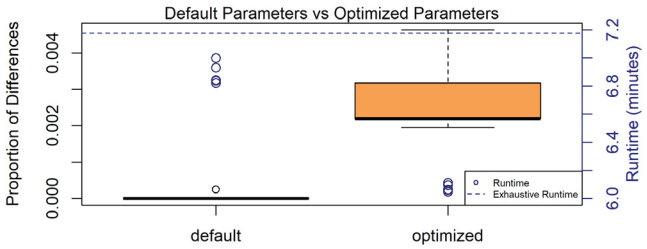


Figure 3: Default parameters compared with the optimized parameters for the four covariate case. Proportion of Differences is represented with boxplots and runtime is represented with triangles. As before, the boxplots (proportion of differences) correspond to the left yaxis and the blue dots (runtime) correspond to the right y-axis.

Looking at the boxplots in Figure 3 we can see that the model accuracy is still very good because the proportion of differences only increased 0.4% with the optimized parameters, compared to the default. Looking at the blue dots we see that the optimized parameters decrease the runtime compared to both the default parameters and the exhaustive method (seen with a dashed line). Thus, we were able to successfully optimize the genetic algorithm for the four covariate case. The runtime savings for the four covariate case was small because of the size of the model. Thus, we studied the five covariate case (which has an additional term in the model) in hopes of having increased savings.

The five covariate case was run on a high performance computing system (HPC) at NCAR due to the long runtimes. The parameters were initially varied individually to find an optimal range of study. Once a range was established for each parameter we varied two parameters at a time. Upon looking at the results of varying the sexual reproduction rate individually it appeared that it didn't impact the runtime or proportion of differences significantly so it was omitted from the study. For the five covariate case we found the optimal parameters to be the following:

```
Population Size: 1000
Mutation Rate: 0.01
Sexual Reproduction Rate:
Immigration Rate: 0.001
Consecutive Value: 5
```

These findings may vary on a personal computer. We now compare the optimized parameter values against the default parameter values for the five covariate case.

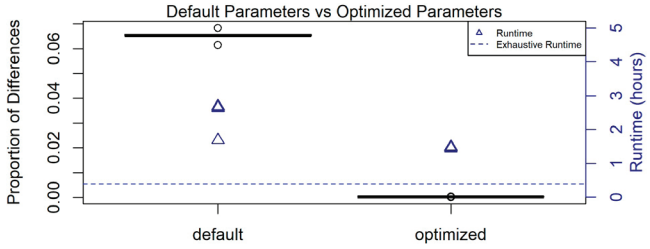
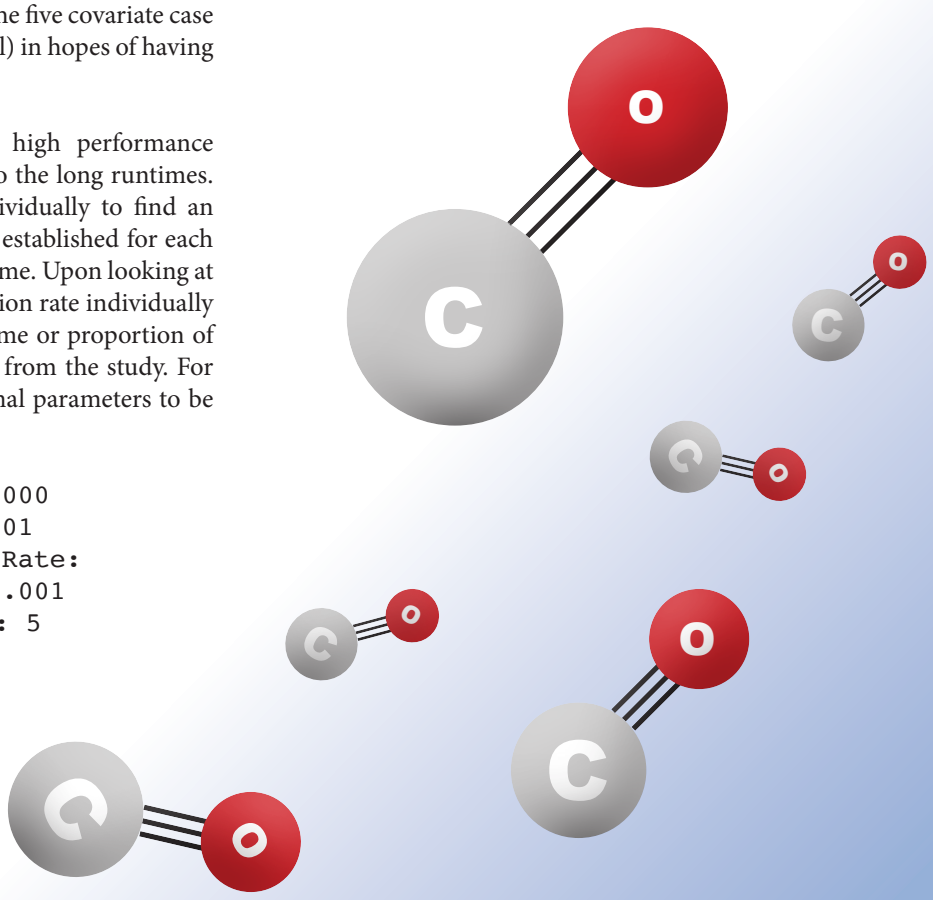


Figure 4: Default parameters compared with the optimized parameters for the five covariate case. Proportion of Differences is represented with boxplots and runtime is represented with triangles.

We can see from Figure 4 that the optimized parameters, while they decreased the runtime from the default parameters, were unable to make the genetic algorithm faster than the exhaustive method (seen with the dotted line). We find that the optimized five covariate model on an HPC system trends towards the exhaustive search as the optimized population size was 1000, which is close to the total number of models of 1,450. Despite being unsuccessful, our findings were useful in illustrating that on an HPC system, the optimized genetic algorithm trends towards the exhaustive search and that the exhaustive search has a high efficiency.



References

Buchholz, R. R., Hammerling, D., Worden, H. M., Deeter, M. N., Emmons, L. K., Edwards, D. P., & Monks, S. A. (2018). Links Between Carbon Monoxide and Climate Indices for the Southern Hemisphere and Tropical Fire Regions. Journal of Geophysical Research: Atmospheres, 123. doi: 10.1029/2018JD028438

Calcagno, V., & de Mazancourt, C. (2010). glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. Journal of Statistical Software, 34(12), 29. doi: 10.18637/jss.v034.i12

Daniels, W., Hammerling, D., & Buchholz, R. (2000). regclimatechem: An R Package for Data Driven Variable Selection Applied to Atmospheric Carbon Monoxide (Tech. Rep.). Colorado School of Mines.

Deeter, M. N., Edwards, D. P., Francis, G. L., Gille, J. C., Mart´ınezAlonso, S., Worden, H. M., & Sweeney, C. (2017). A Climate-Scale Satellite Record for Carbon Monoxide: the MOPITT Version 7 Product. Atmospheric Measurement Techniques, 10(7), 2533–2555. Retrieved from https://amt.copernicus.org/articles/10/2533/2017/ doi: 10.5194/amt-10-2533-2017

Deeter, M. N., Emmons, L. K., Francis, G. L., Edwards, D. P., Gille, J. C., Warner, J. X., ... Drummond, J. R. (2003). Operational Carbon Monoxide Retrieval Algorithm and Selected Results for the MOPITT Instrument. Journal of Geophysical Research: Atmospheres, 108(D14). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD003186 doi: https://doi.org/10.1029/2002JD003186

Drummond, J. R., Zou, J., Nichitiu, F., Kar, J., Deschambaut, R., & Hackett, J. (2010). A Review of 9-year Performance and Operation of the

MOPITT Instrument. Advances in Space Research, 45(6), 760–774. doi: 10.1016/j.asr.2009.11.019

Edwards, D. P., Emmons, L. K., Gille, J. C., Chu, A., Atti´e, J.-L., Giglio, L., ... Drummond, J. R. (2006). Satellite-Observed Pollution from Southern Hemisphere Biomass Burning. Journal of Geophysical Research: Atmospheres, 111(D14). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006655 doi: https://doi.org/10.1029/2005JD006655

Holloway, T., Levy II, H., & Kasibhatla, P. (2000). Global Distribution of Carbon Monoxide. Journal of Geophysical Research: Atmospheres, 105(D10), 12123-12147. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JD901173 doi: https://doi.org/10.1029/1999JD901173

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (Fourth ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4/ (ISBN 0-387-95457-0)

About the Author

Meera Duggal is a Senior in the Applied Mathematics and Statistics Department. She conducted research under Dr. Hammerling and Willian Daniels, who is a graduate student also working with Dr. Hammerling. She is going on to pursue a Masters in Statistics this fall at U.C. Berkeley.