Optimizing Genetic Algorithm Parameters for Atmospheric Carbon Monoxide

Meera Duggal, William Daniels, Dorit Hammerling

Motivation

- The main source of carbon monoxide (CO) in the southern hemisphere are large burn events
- This makes CO an useful proxy for fires
- Fires are influenced by the change in atmosphere and oceans (measured by climate indices)
- Thus, the atmospheric CO is modeled using climate indices as predictor variables
- These models can help countries prepare for large burn events

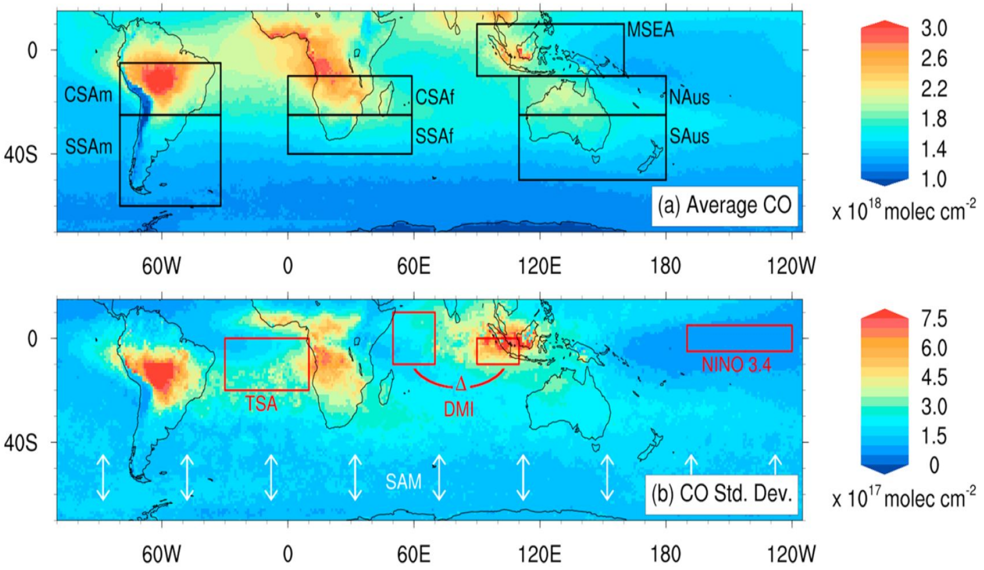
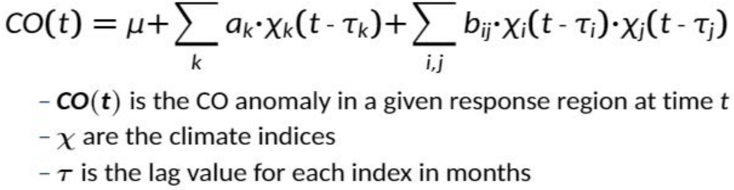


Figure 1: In plots (a) the 7 different regions are displayed along with the average CO. In plots (b) the 4 different climate indices are shown along with the standard deviation of CO.

Statistical Modeling

- We use multiple linear regression to model CO



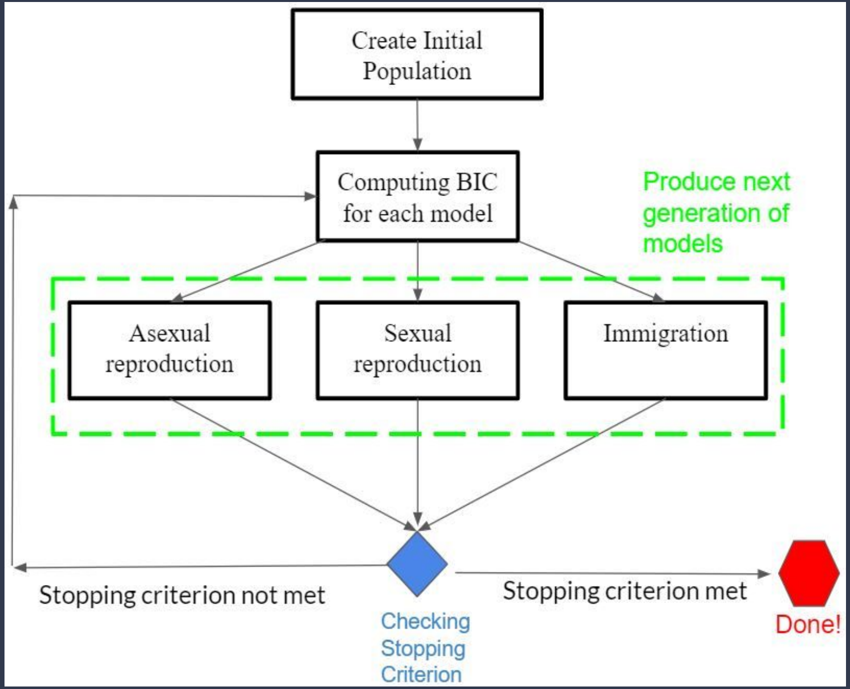
$$CO(t) = \mu + \sum_{k} a_k \cdot \chi_k(t - \tau_k) + \sum_{i,j} b_{ij} \cdot \chi_i(t - \tau_i) \cdot \chi_j(t - \tau_j)$$

– $\mathbf{CO}(\boldsymbol{t})$ is the CO anomaly in a given response region at time t
– χ are the climate indices
– τ is the lag value for each index in months

- The R package ***regClimateChem*** provides three variable selection algorithms
  - Exhaustive: always finds best possible model, but most computationally expensive
  - Stepwise: often fails to find the best possible model, but computationally inexpensive
  - Genetic: implemented as a possible middle-ground between exhaustive and stepwise

- We are performing an optimization study on the genetic algorithm to find the best balance between runtime and accuracy

The Genetic Algorithm

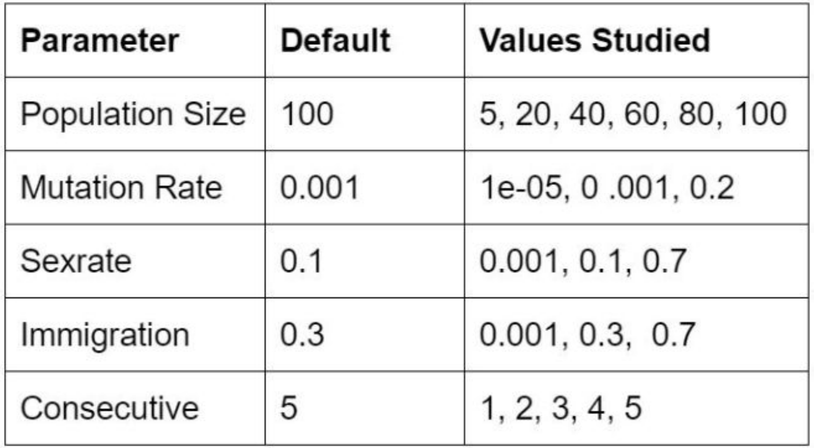


- The genetic algorithm is implemented in ***regClimateChem*** via the ***glmulti*** package
- A stochastic variable selection technique, it is based on probability and will potentially produce different results each time you run it
- The algorithm converges to the best model by continuing to modify a population of models
- There are various parameters in glmulti that affect how this modification process occurs

Genetic Algorithm Optimization Study

The four covariate case

The genetic algorithm contains many different parameters. We chose the following parameters for our study.

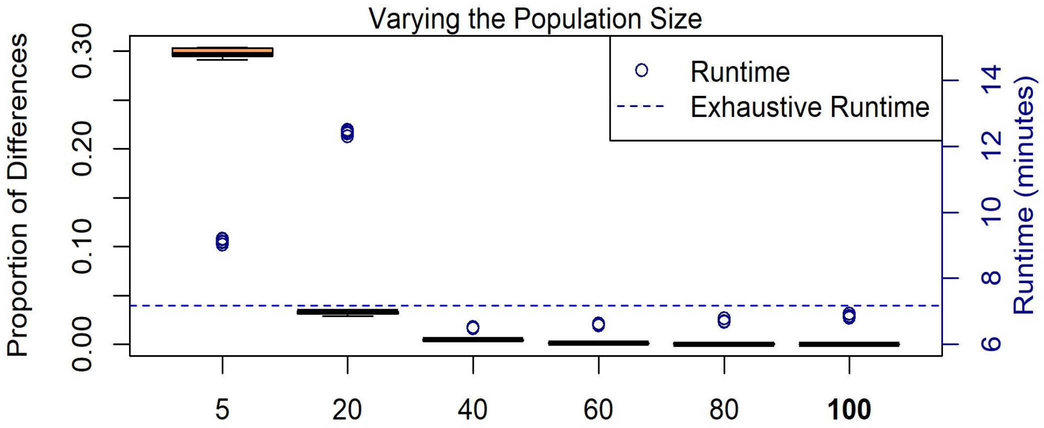


Parameter	**Default**	**Values Studied**
Population Size	100	5, 20, 40, 60, 80, 100
Mutation Rate	0.001	1e-05, 0 .001, 0.2
Sexrate	0.1	0.001, 0.1, 0.7
Immigration	0.3	0.001, 0.3, 0.7
Consecutive	5	1, 2, 3, 4, 5

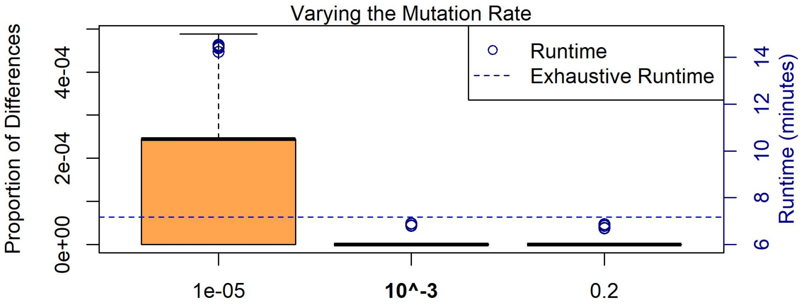
Four covariate case

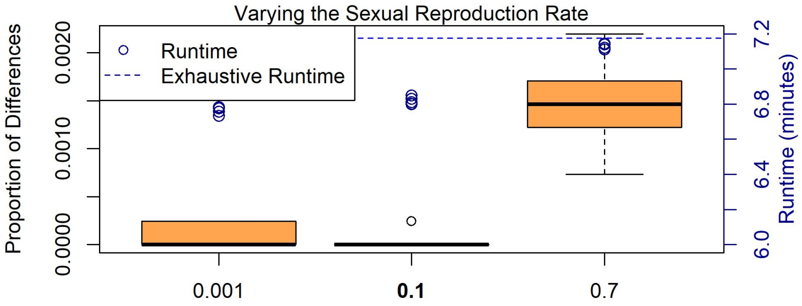
- The x axis shows the different parameter values, with the default value in bold

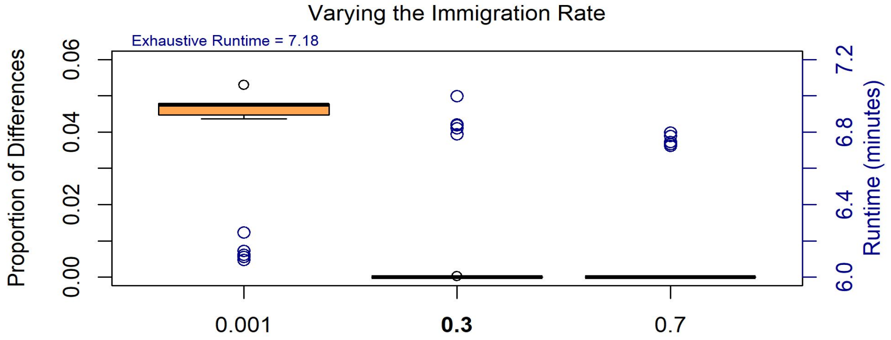
- There are two y axes, runtime and proportion of differences

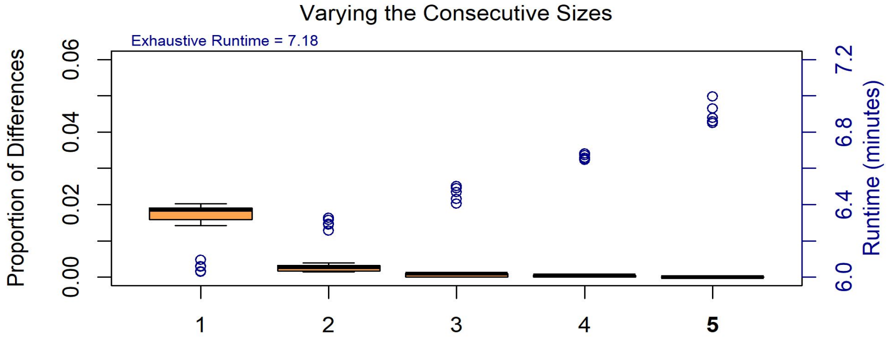


$$\text{Proportion of Differences: } \frac{\#\text{ of models different from exhaustive}}{\text{total }\#\text{ of models}}$$


Varying the Mutation Rate


Varying the Sexual Reproduction Rate

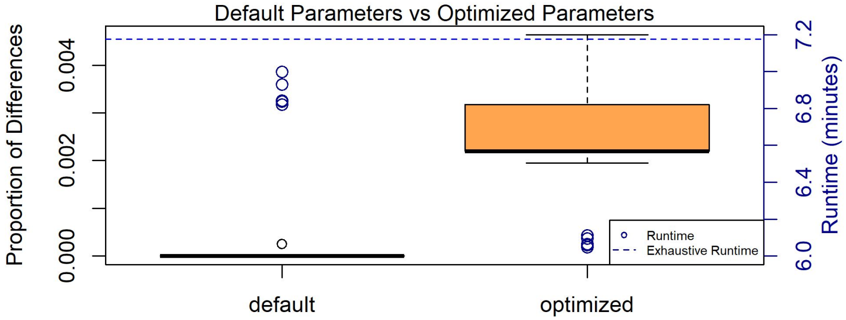

Varying the Immigration Rate


Varying the Consecutive Sizes

Results: four covariate case

- The optimized parameters are
    - Population size = 40
    - Mutation Rate = 0.2
    - Sexual Reproduction Rate = 0.001
    - Immigration Rate = 0.001
    - Consecutive = 2

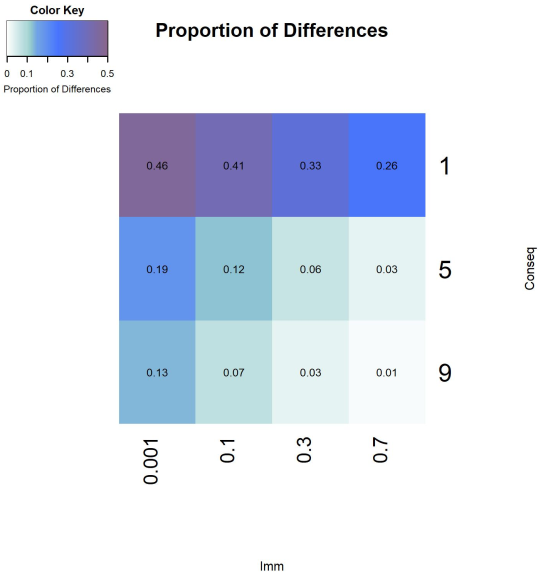
*the parameters in red are more sensitive to change



- When using all optimized values the run time decreases an average of 11.8%
- The proportion of models different from the exhaustive search only differed by 0.28%

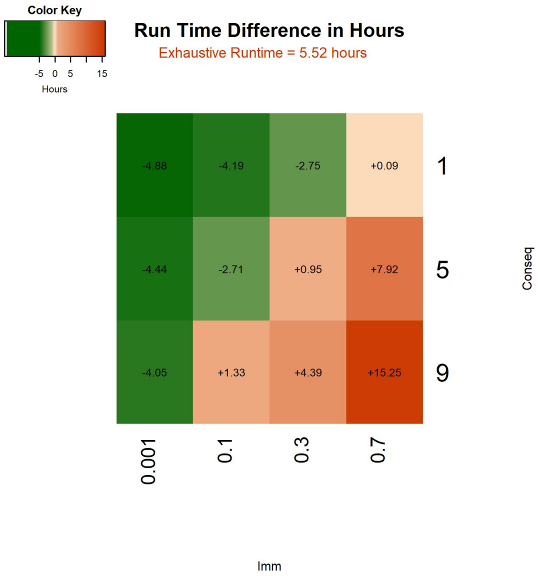
Testing the Five Covariate Case

- We tested the same parameters on five covariate models
- Varied parameters one at a time to estimate optimal range, then tested two parameters concurrently on those ranges



Proportion of Differences

Conseq \ Imm	0.001	0.1	0.3	0.7
1	0.46	0.41	0.33	0.26
5	0.19	0.12	0.06	0.03
9	0.13	0.07	0.03	0.01



Run Time Difference in Hours

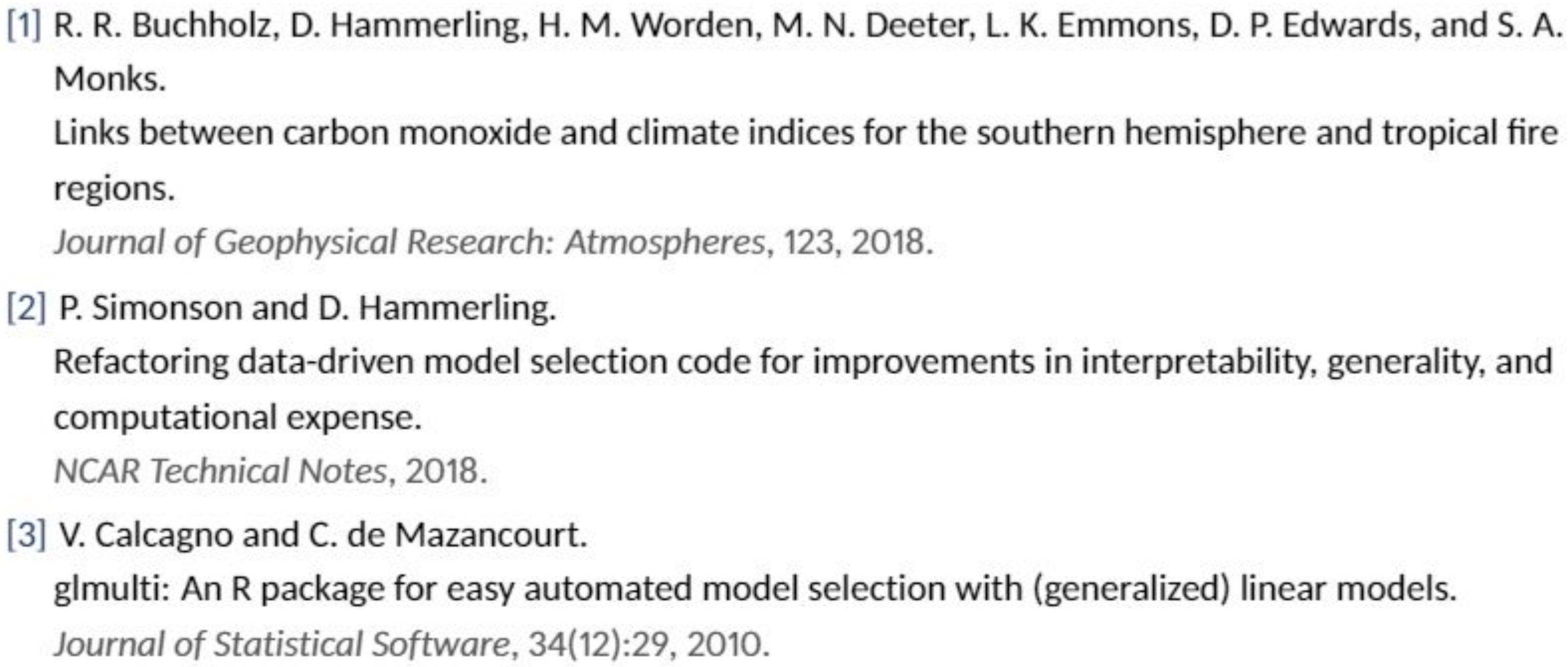
Exhaustive Runtime = 5.52 hours

Conseq \ Imm	0.001	0.1	0.3	0.7
1	-4.88	-4.19	-2.75	+0.09
5	-4.44	-2.71	+0.95	+7.92
9	-4.05	+1.33	+4.39	+15.25

Future Work

- We are still in the process of varying two parameters at a time for the five covariate case
- We moved our study from my personal laptop to the HPC system (Cheyenne) at the National Center for Atmospheric Research due to very long runtimes (20 hours)
- We hope to come up with a parameter combination that makes genetic algorithm fall between stepwise and exhaustive

References



[1] R. R. Buchholz, D. Hammerling, H. M. Worden, M. N. Deeter, L. K. Emmons, D. P. Edwards, and S. A. Monks.
Links between carbon monoxide and climate indices for the southern hemisphere and tropical fire regions.
Journal of Geophysical Research: Atmospheres, 123, 2018.

[2] P. Simonson and D. Hammerling.
Refactoring data-driven model selection code for improvements in interpretability, generality, and computational expense.
NCAR Technical Notes, 2018.

[3] V. Calcagno and C. de Mazancourt.
glmulti: An R package for easy automated model selection with (generalized) linear models.
Journal of Statistical Software, 34(12):29, 2010.