Genetic Algorithm Optimization Study for Atmospheric Carbon Monoxide Models

Meera Duggal[1], William Daniels[1], Dorit Hammerling[1,2]

Colorado School of Mines[1], National Center for Atmospheric Research[2]

Motivation

- In the Southern Hemisphere, the main source of atmospheric carbon monoxide (CO) are large burn events
- Therefore, CO can be used as a proxy for fires
- Predictive models for atmospheric CO concentrations can help countries prepare for large burn events

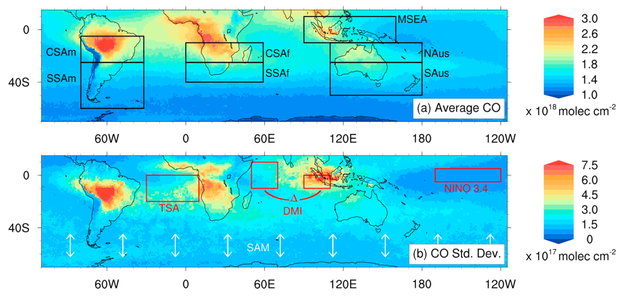

Figure 1: In plots (a) the 7 different regions are displayed along with the average CO In plots (b) the 4 different climate indices are shown along with the standard deviation of CO.

Introduction

- We use multiple linear regression to model atmospheric CO

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

  - $CO(t)$ 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*: Finds the best model, but has a long runtime
  - *Genetic*: A middle ground between stepwise and exhaustive
  - *Stepwise*: Very fast, but finds the best model least often
- We performed an optimization study to optimize both model accuracy and runtime in the genetic algorithm

The Genetic Algorithm

- A stochastic variable selection technique
- Maintains a population of models throughout the algorithm
- Three different model modification techniques are used to produce new generations
- A stopping criterion is checked after each new generation is produced

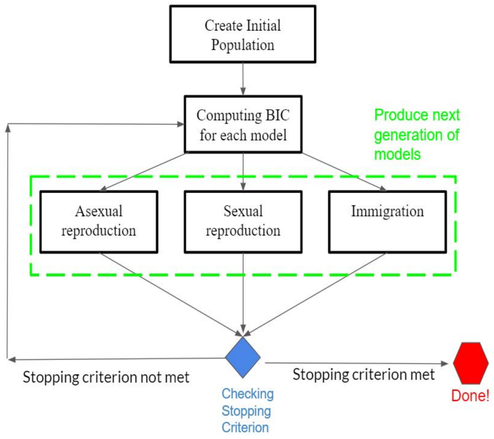

Figure 2: Flow of the Genetic Algorithm

Genetic Algorithm Optimization Study

The genetic algorithm is implemented in *regClimateChem* via the *glmulti* package. It contains hyperparameters. We have chosen to study the following hyperparameters, varying them one at a time:

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

Table 1: Values tested for each hyperparameter

Optimization Results

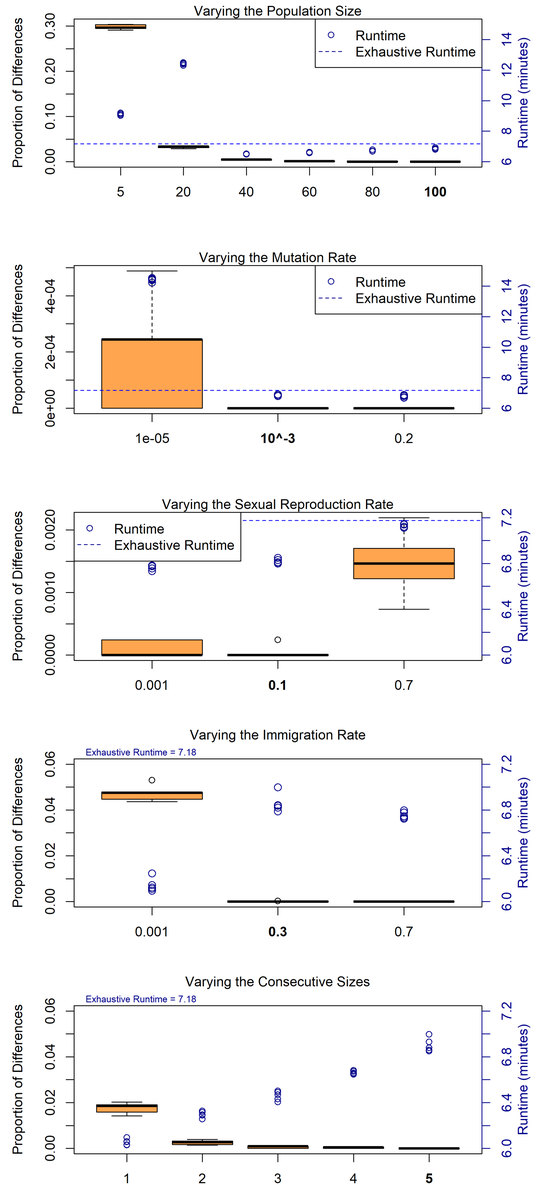

Figure 3: The x axis shows the different hyperparameter values, with the default value in bold. There are two y axes, runtime and proportion of differences The proportion of differences is the percentage of models that differ from the exhaustive models, over the total models.

Discussion of Results

- The optimized hyperparameters are:
  - **population size = 40**
  - **mutation rate = 0.2**
  - **sexual reproduction rate = 0.001**
  - **immigration rate = 0.001**
  - **consecutive = 2**
- The hyperparameters in red are more sensitive to change compared to the other hyperparameters
- Comparison of results between the default and all optimized hyperparameters:

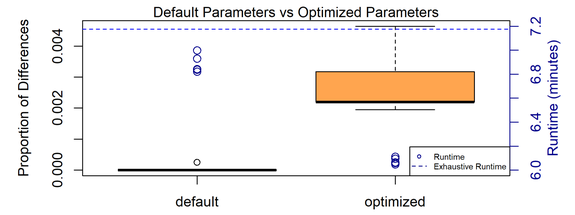

Figure 4: Comparison of using default hyperparameter values and optimized hyperparameter values.

- When using all optimized values the run time decreased an average of 11.8%
- Compared to the default hyperparameters the model quality doesn't decrease much

Future Work

- Testing models with more covariates
  - Testing these findings on larger models which will potentially make a larger impact on runtime
- Varying multiple hyperparameters at a time
  - Finding the overall optimal solution by accounting for interactions between hyperparameters

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.