Modeling Global Temperature Changes using Genetic Programming – A Case Study

Karolina Stanisławska*, Krzysztof Krawiec*, Zbigniew W. Kundzewicz+

*Institute of Computing Science,
Poznan University of Technology, Poznań, Poland

+Institute for Agricultural and Forest Environment, Polish Academy of Sciences, Poznań, Poland
and Potsdam Institute for Climate Impact Research, Potsdam, Germany

KAEiOG, September 21, 2011

Outline

The topic

Data-driven discovery of plausible models that link global temperature with natural and anthropogenic forcings (drivers).

The objectives

- To obtain models for
  - forecasting,
  - explanation (hindcasting).
- To verify usefulness of genetic programming (GP) for that task.

Features of the climate system

Climate as a complex system

Involves a large number of highly interconnected components that influence each other in a complex manner (e.g., nonlinear, nonmonotonous).

Known external drivers controlling the Earth's climate:

- Solar activity,
- The distance between the Sun and the Earth
  - also: slowly varying Earth's orbital patterns,
- Volcanic eruptions,
- Properties of the atmosphere (greenhouse gases, dust and aerosols),
- Properties of the Earth's surface
  - albedo of the surface,
  - availability of water on and under the land surface.

Features of the climate system

Several modes of oscillation (inertia) in the Ocean-Atmosphere system:

- El Niño-Southern Oscillation (ENSO),
- North Atlantic Oscillation (NAO),
- Atlantic Multidecadal Oscillation (AMO),
- Pacific Decadal Oscillation (PDO), etc.

Internal feedbacks, e.g:

- Warming => decrease of ice and snow areas => decreasing albedo => less heat reflected into space => further warming
- Thawing of permafrost => emission of methane => further warming

Unknowns?

Contemporary climate modeling

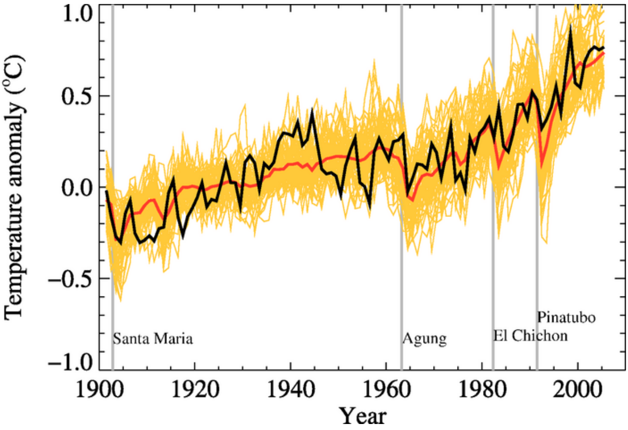
Features:

- Derived from fundamental physical laws,
- Subject to physical approximations,
- Subject to extra approximation due to spatiotemporal discretization.
- Typical size:
  - One to a few degrees in longitude and latitude,
  - 10 to 20 vertical layers in the atmosphere,
  - 30 or more layers in the oceans,
  - $> 10^6$ grid points.
  - Gigantic computational effort.

Drawbacks:

- Some physical processes occur at smaller (sub-grid) scales and cannot be properly modeled.
  - Require integration over larger scale (so-called parameterization)
- Extensive tuning required.

Limitations of contemporary climate models



- Black: observed global mean near-surface temperatures.
- Yellow: 14 different climate models.
- Red: The mean of all models.
- Vertical grey lines: major volcanic eruptions.

(By permission from IPCC, see (Randall et al. 2007)).

Limitations

Problems:

- limited computer power (even if vast),
- limited scientific understanding,
- lack of availability of detailed observations of some physical processes.

The consequence: Climate change information is highly uncertain.

- “known unknowns” and “unknown unknowns” (Trenberth 2010).

Climate models are not yet up to “prime time”, particularly in some application areas.

The approach

The idea

Data-driven approach to model climate phenomena, employed to distill free-form natural laws from experimental data.

Inspiration: Recent advances in Genetic Programming (GP):

- GP can automatically find and correct bugs in commercially-released software (Arcuri & Yao 2008, Forrest 2010).
- GP can be used to 'automate' science, helping the researchers to find the hidden complex models of the observed phenomena (Schmidt & Lipson 2009).

The approach

Discovering the multiple inputs-single output (MISO) dependency between:

- global mean temperature (dependent variable)
- and several climate factors (independent variables)

expressed as monthly data series.

Technically:

- An evolutionary algorithm (genetic programming, GP) evolves a population of programs (expressions),
- Each program is a specific model of dependency between independent variables and the dependent variable.
- Models represented as expression trees.

The dependent variable

University of East Anglia global mean temperature (UEA):

- Aggregates the temperature over $5° \times 5°$ grid boxes over land (air temperature) and oceans (sea surface temperature, SST),
- Relative to the mean from 1961-1990
- Starts in 1850

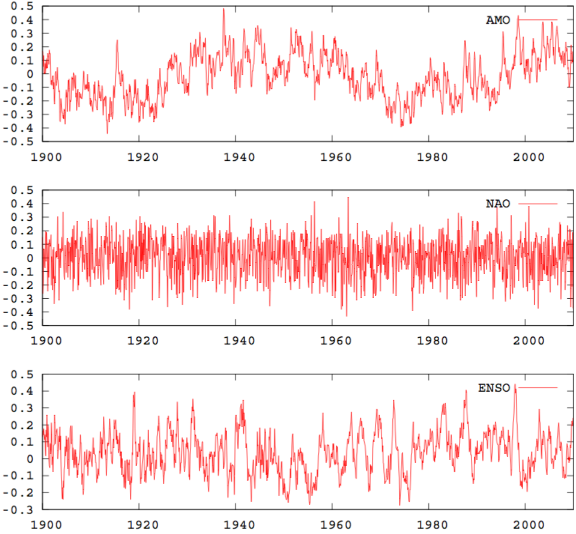
The independent variables

- Sun Spots Number (SSN, since 1749).
- Atlantic Multidecadal Oscillation (AMO, since 1856): The mean sea surface temperature (SST) of North Atlantic (latitude 0°-70°N, detrended to remove the influence of global warming).
- North Atlantic Oscillation (NAO, since 1865): An index calculated from the measurements of air pressure at two locations: Ponta Delgada, Azores, and Stykkisholmur/Reykjavik in Iceland.
- El Niño/Southern Oscillation (ENSO, since 1845): Temperature fluctuations expressed by the average SST anomaly of the region 20°N-20°S minus 90°N-20°N and 20°S-90°S.
- Concentrations of greenhouse gases:
  - CO_2
  - N_2O,
  - CH_4
- Volcanic Explosivity Index (VEI, since 1851): An index marking major volcanic explosions.

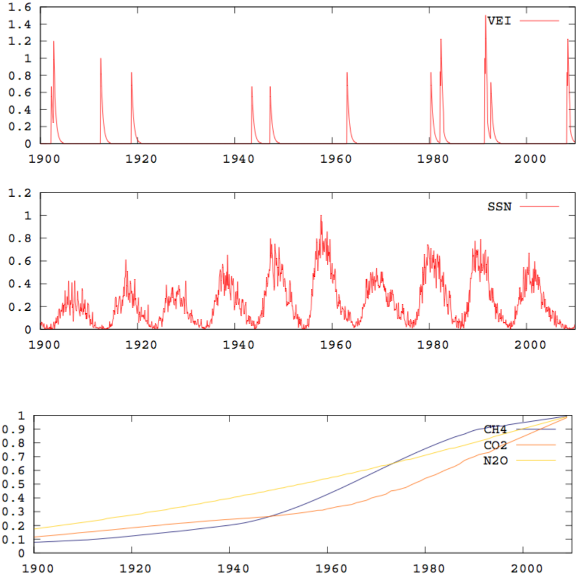
The data

- The considered time period: 1900-2009 ($110 \times 12 = 1320$ data points)
  - training period: 1900-1999 (1200 data points)
  - testing period: 2000-2009 (120 data points).
- Preprocessing:
  - normalization
    - zero-preserving normalization for bipolar variables (AMO, NAO, ENSO)
  - VEI models the decreasing impact of eruption over time

The independent variables



The independent variables



Evaluation (fitness calculation)

- One-step ahead forecasting (training period only):
  - At the time step (month) t, the model forecasts the temperature at time step $t + 1$ based on historical data ($\leq t$).
    - No access to historical temperature.
  - Errors aggregated by mean absolute error (MAE).
- The terminal nodes in expression trees return either
  - the current value of an independent variable (at time step t),
  - an aggregate of historical values (e.g., weighted averages of historical values).

Definition of GP terminals

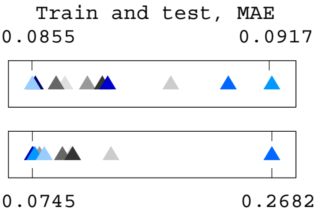
Terminal name	Terminal semantics
NAO, AMO, ENSO, SSN, VEI, CO_2, N_2O, CH_4	The value at time point t
AMO_n, NAO_n, $ENSO_n$, VEI_n, SSN_n	The value at time point $(t - n)$, with $n \in [0, 11]$ determined randomly at the moment of node creation.
$AMO_{m,n}$, $NAO_{m,n}$, $ENSO_{m,n}$, $VEI_{m,n}$, $SSN_{m,n}$	The mean value in time period $[t - m, t - n]$ with $m, n \in [1, 12]$, $m < n$ determined randomly at the moment of node creation.
NAOw	An aggregate value of the NAO index for the preceding winter (Dec-Mar of current or previous year)
C	A constant drawn uniformly from interval $[-1, 1]$ at the moment of node creation.

Evolutionary parameters

- Number of generations: 100,
- Population size: 10000 individuals,
- Probability of crossover: 0.9,
- Probability of mutation: 0.1,
- Maximum tree depth 17,
- Tournament selection 7.

Implementation based on ECJ software package.

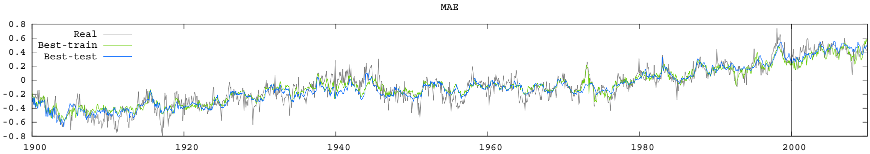
Results



Comparison of MAE error committed by the evolved models on the training set (top) and test set (bottom).

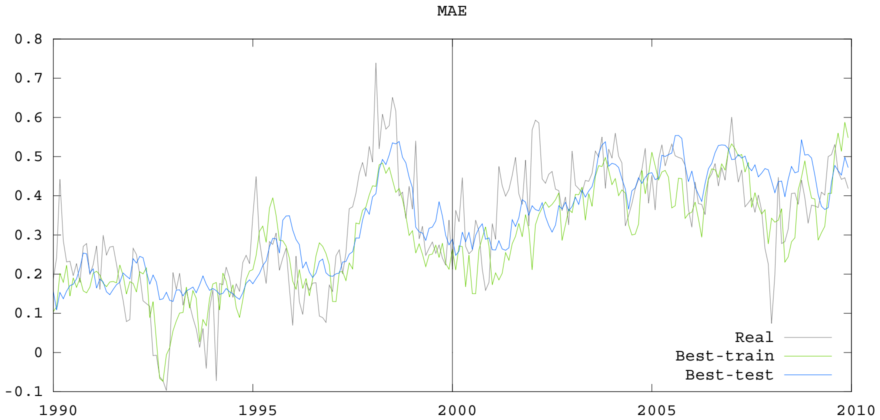
- Each color corresponds to a single model produced by an independent GP run.

Results: The entire period



- Grey: The actual UEA global mean temperature record
- Green: The forecast produced by the best-on-training-set model
- Blue: The forecast produced by the best-on-test-set model

Results: Last 10 years of training period and test



- Grey: The actual UEA global mean temperature record
- Green: The forecast produced by the best-on-training-set model
- Blue: The forecast produced by the best-on-test-set model

Conclusions

- GP is capable of inducing models that mimic the aggregate behavior of selected aspect of the complex climate system,
  - without resorting to historical temperature itself,
  - unbiased by the preferences of human experimenter.
- Future work:
  - Evolving models representing differential equations.
- The data-driven approach allows making interpretations that are potentially useful in climatology.
  - Interpretation?

Thank you.

$$temperature_{UEA} = x_1 \left(e^{N2O+e^{x_2}}\right)^{-1}, \text{ where:}$$

$$x_1 = -e^{x_5} + VarPre1\,(AMO) + CO2\, e^{N2O} + ENSOpre -$$
$$0.15502\, log\,(1.10235)\, e^{N2O} - x_6 - e^{-\left(e^{e^{N2O}}\right)}$$
$$x_2 = -\left(e^{x_3 x_4}\right)$$
$$x_3 = -0.15502\, \frac{e^{-0.18342-0.15502\, e^{AMOpre}}}{-\left(CH4 - e^{-\left(e^{N2O}\right)}\right) + e^{-\left(e^{e^{2\,N2O}}\right)}} + \left(e^{log(VarN16\_10(VEI))}\right)$$
$$x_4 = -0.31004\, \frac{N2O}{e^{N2O} + log\left(CH4 - e^{-\left(e^{N2O}\right)}\right)} + \left(e^{-\left(e^{VarN13\_8(AMO)}\right)}\right)$$
$$x_5 = \frac{e^{-0.18342+NAOpre}}{-\left(e^{AMOpre+\left(e^{N2O}\right)}\right) + e^{-\left(e^{e^{2\,N2O}}\right)}\left(e^{e^{AMOpre+\left(e^{N2O}\right)}} - N2O\right)} + 0.13508$$
$$x_6 =$$
$$log\,(1.10235)\, log\left(CH4 - e^{-\left(e^{N2O}\right)}\right) log\left(-\left(e^{N2O}\right) - 2\,N2O + 2\,ENSOpre\right)$$

http://www.cs.put.poznan.pl/kkrawiec