Genetic Programming for Ocean Microbial Ecology and Biodiversity

John R. Moisan
NASA/GSFC
Earth Science Division
Code 610.W
John.R.Moisan@nasa.gov
5 May, 2016

GORDON AND BETTY
MOORE
FOUNDATION

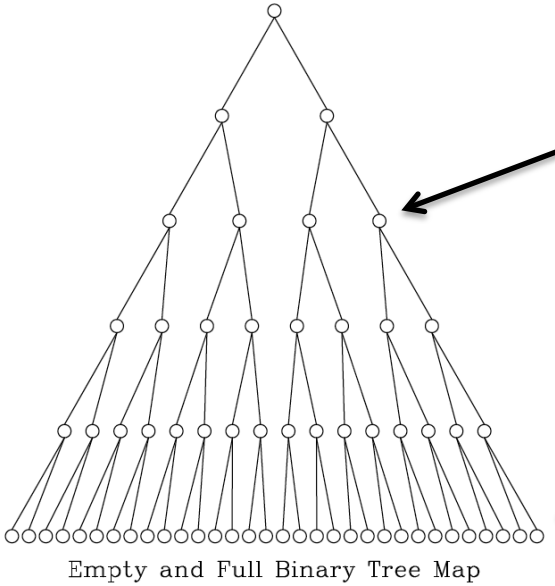
Outline

- Overview of Genetic Programming architecture
- Evolving satellite algorithms
- GP evolution of ecosystem models
- Twin experiment solutions
- Future Directions: Moving to evolving trait-based models.

Genetic Programming Overview

1) Generate initial random population of models/equations
2) Calculate fitness of all individual models/equations
3) Randomly select based on fitness for:
   - Asexual reproduction
   - Sexual reproduction (i.e. Tournament Selection)
4) Carry out genetic mutation (these are rare events ~<1%)
5) Introduce 'recruited' individuals [**NEW MODIFICATION**]
6) Run combination of Genetic Algorithm/lmdif* routine to optimize the model parameters. (*Golfing Analog)
7) Calculate fitness of new individuals
8) Test for completion criteria (low SSE value met?)
9) No: go to 3; Yes: end program

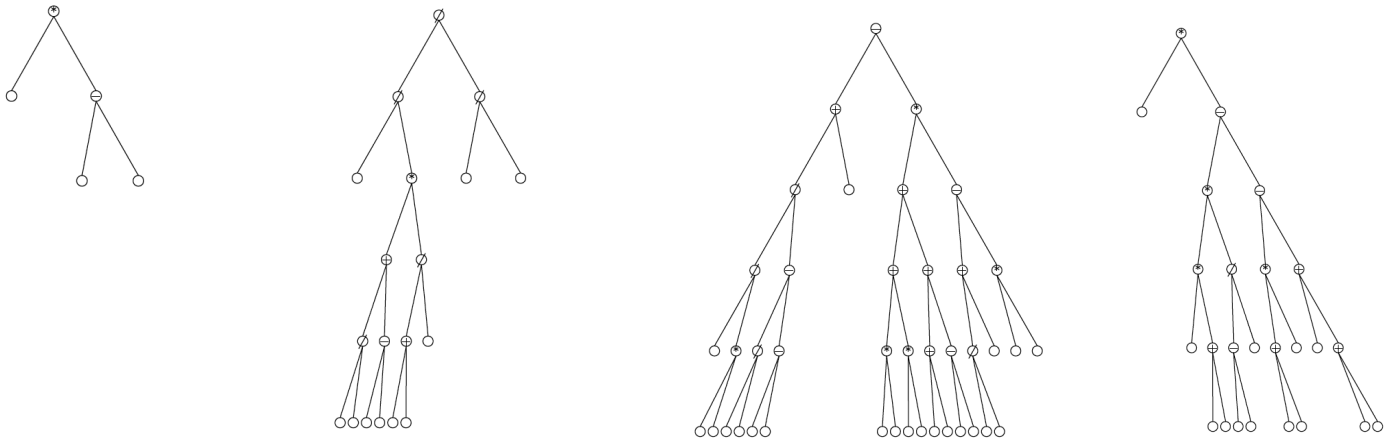
Binary Tree Architecture



Nodes are operators (+, -, /, *, boolean, etc)

Terminals are variables or forcing parameters

Examples of randomly generated initial tree/lisp structures

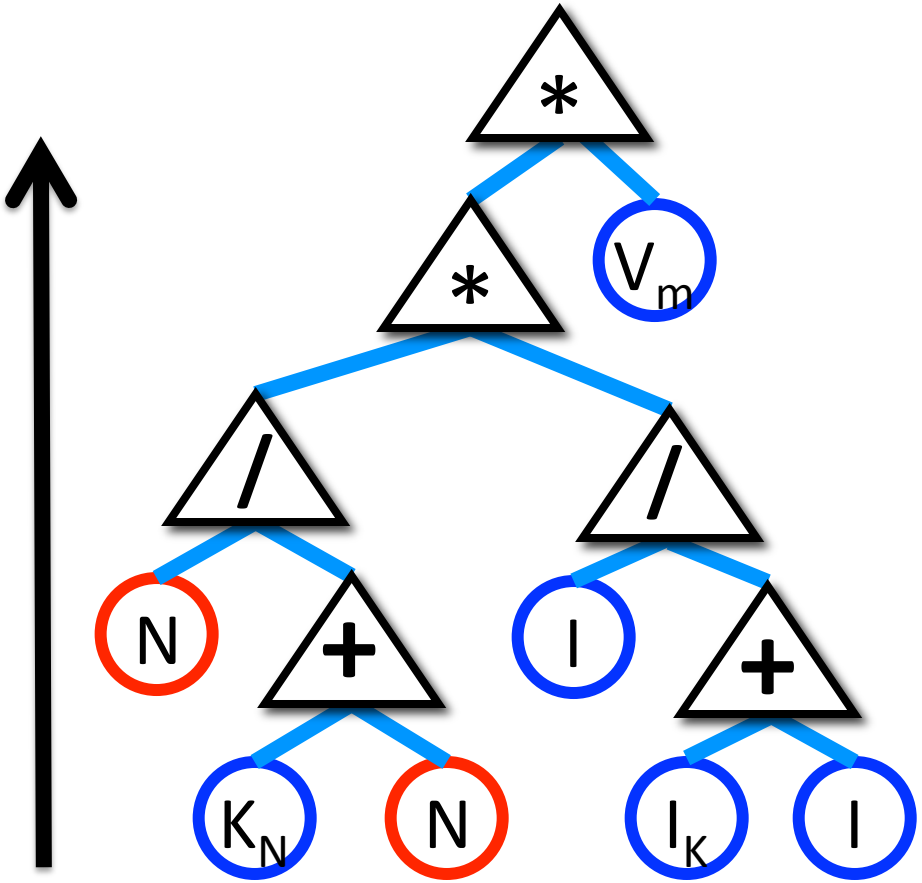


$$\text{Growth} = V_m \left[\frac{N}{K_N + N} \right] \left[\frac{I}{I_K + I} \right]$$

LISP prefix representation:

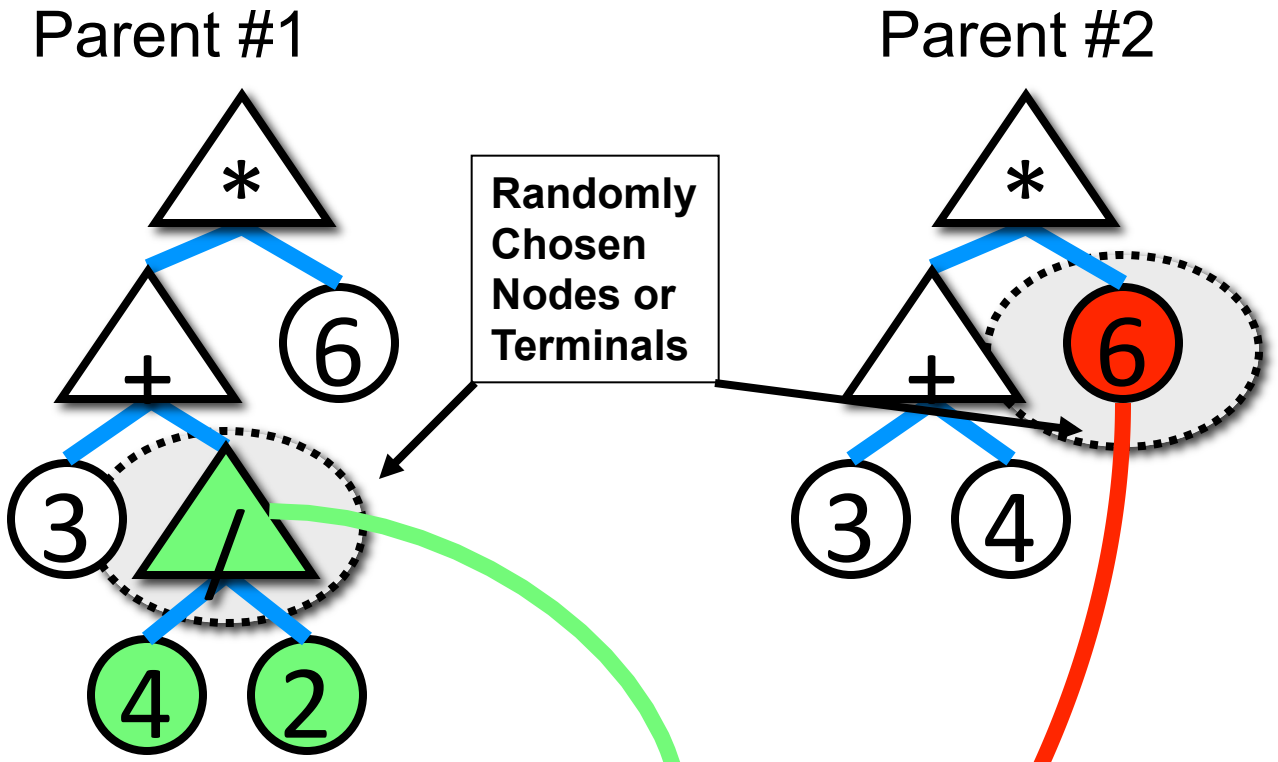
(* V_m (* (/ N (+ (K_N N))) (/ I (+ (I_K I)))))

Calculation of Tree value moves from the bottom up.

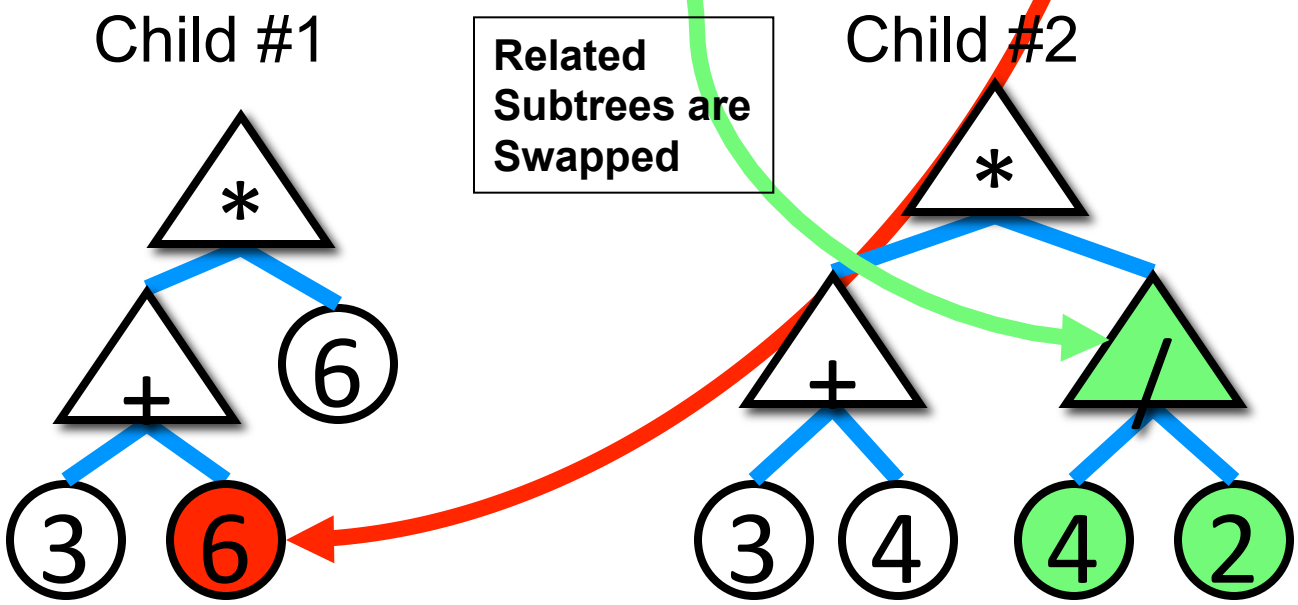


Cross methodology

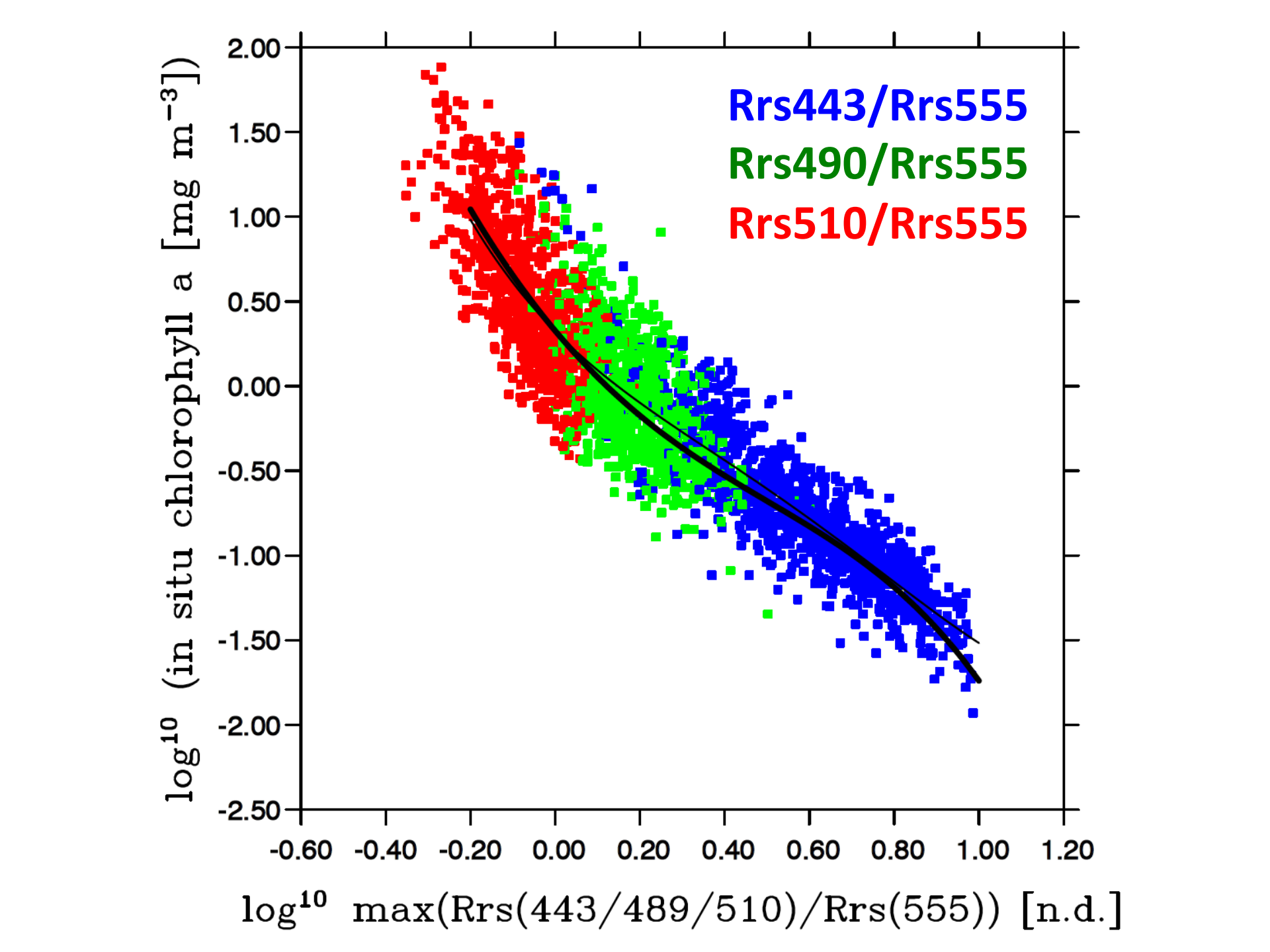
Before Cross

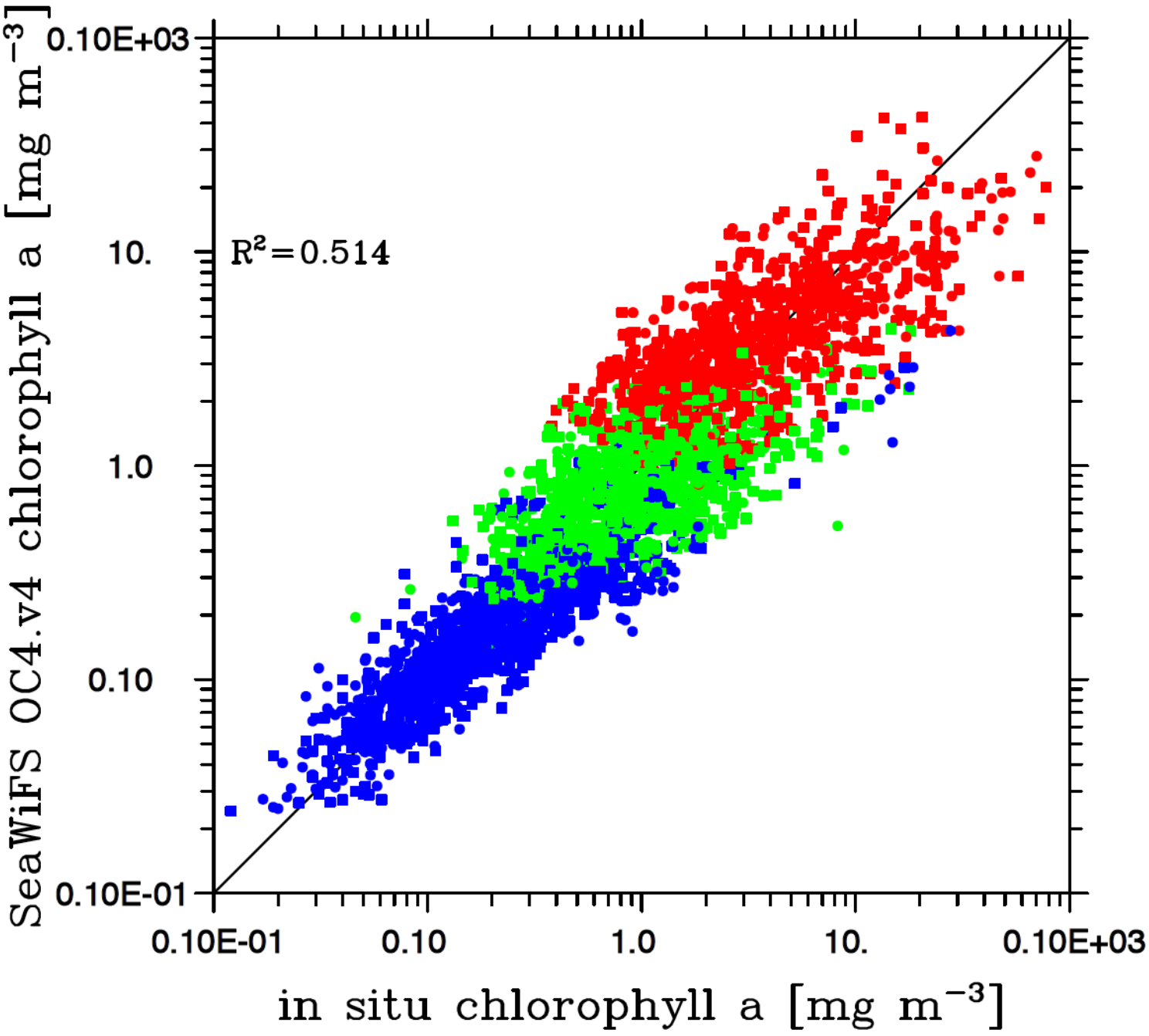


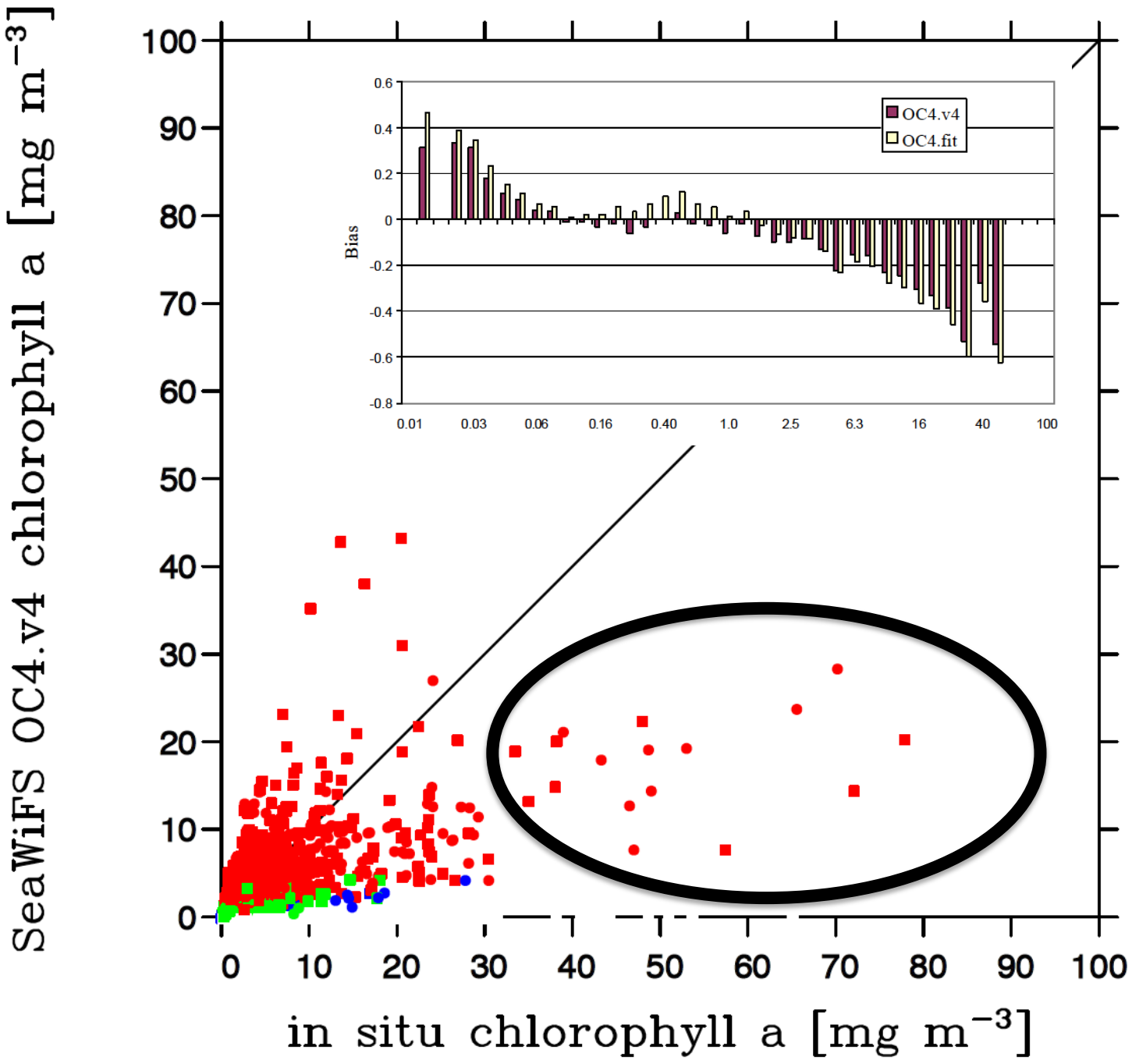
After Cross

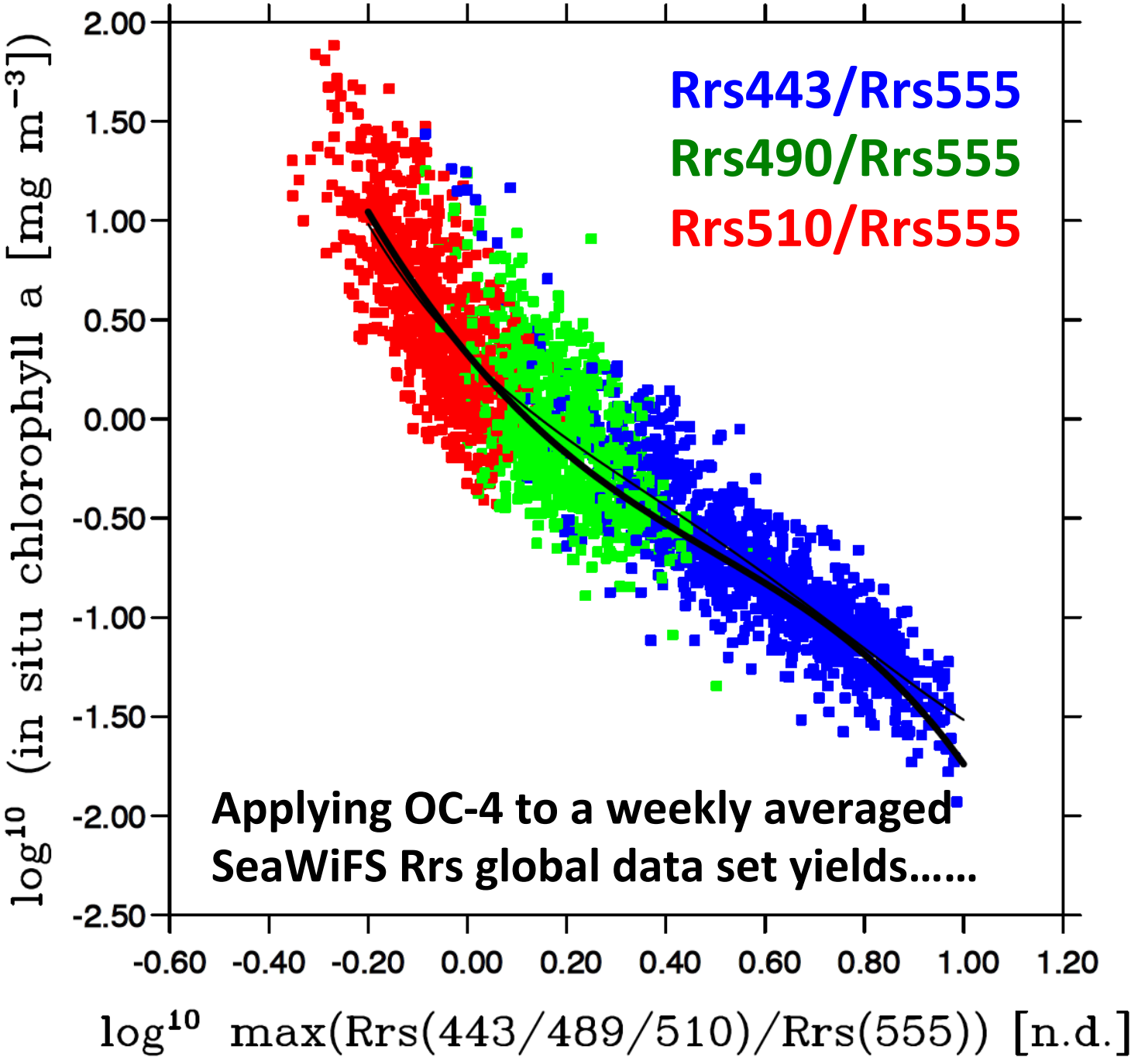


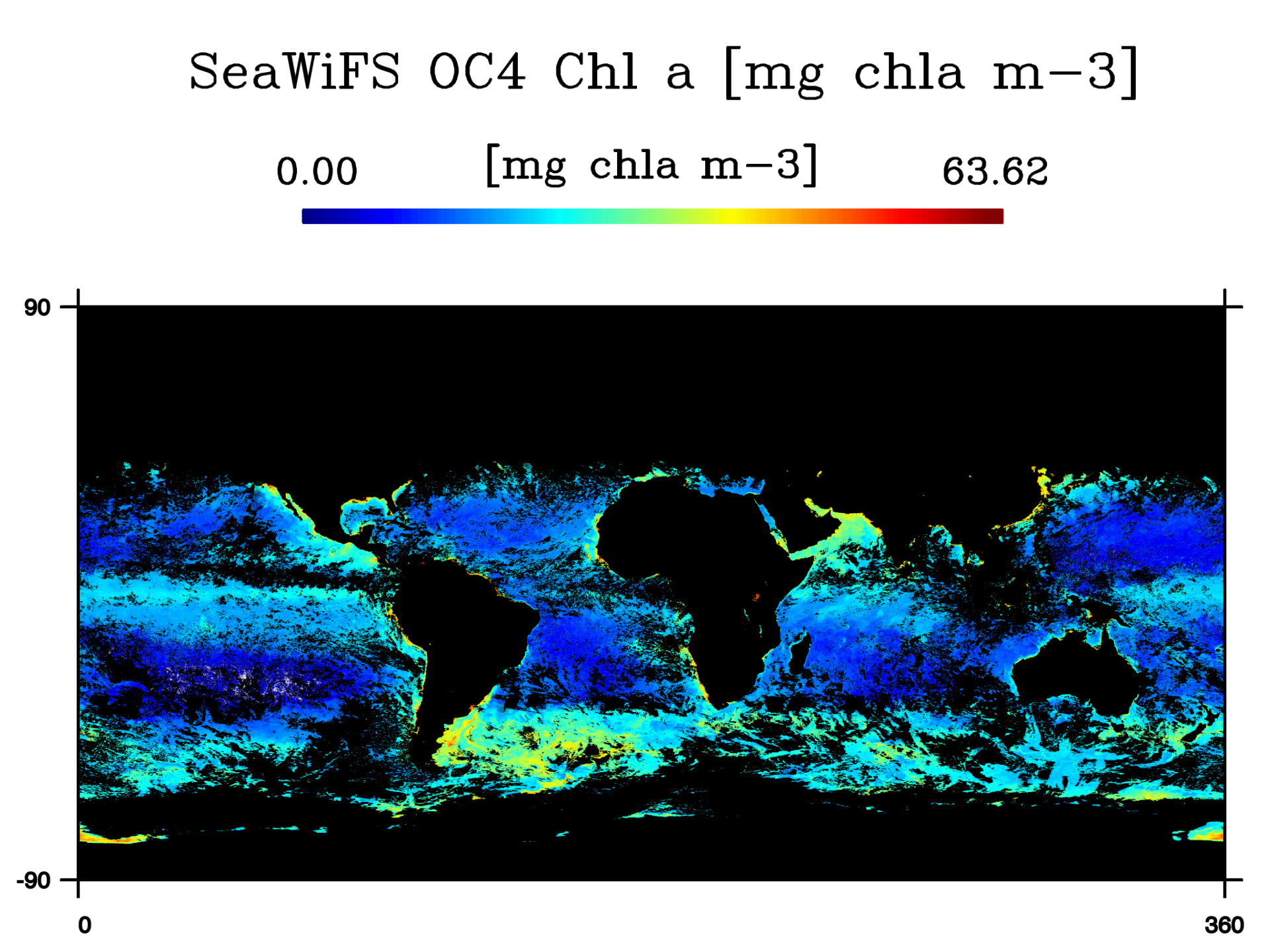
Use of Genetic Programming to Evolve Satellite Algorithms

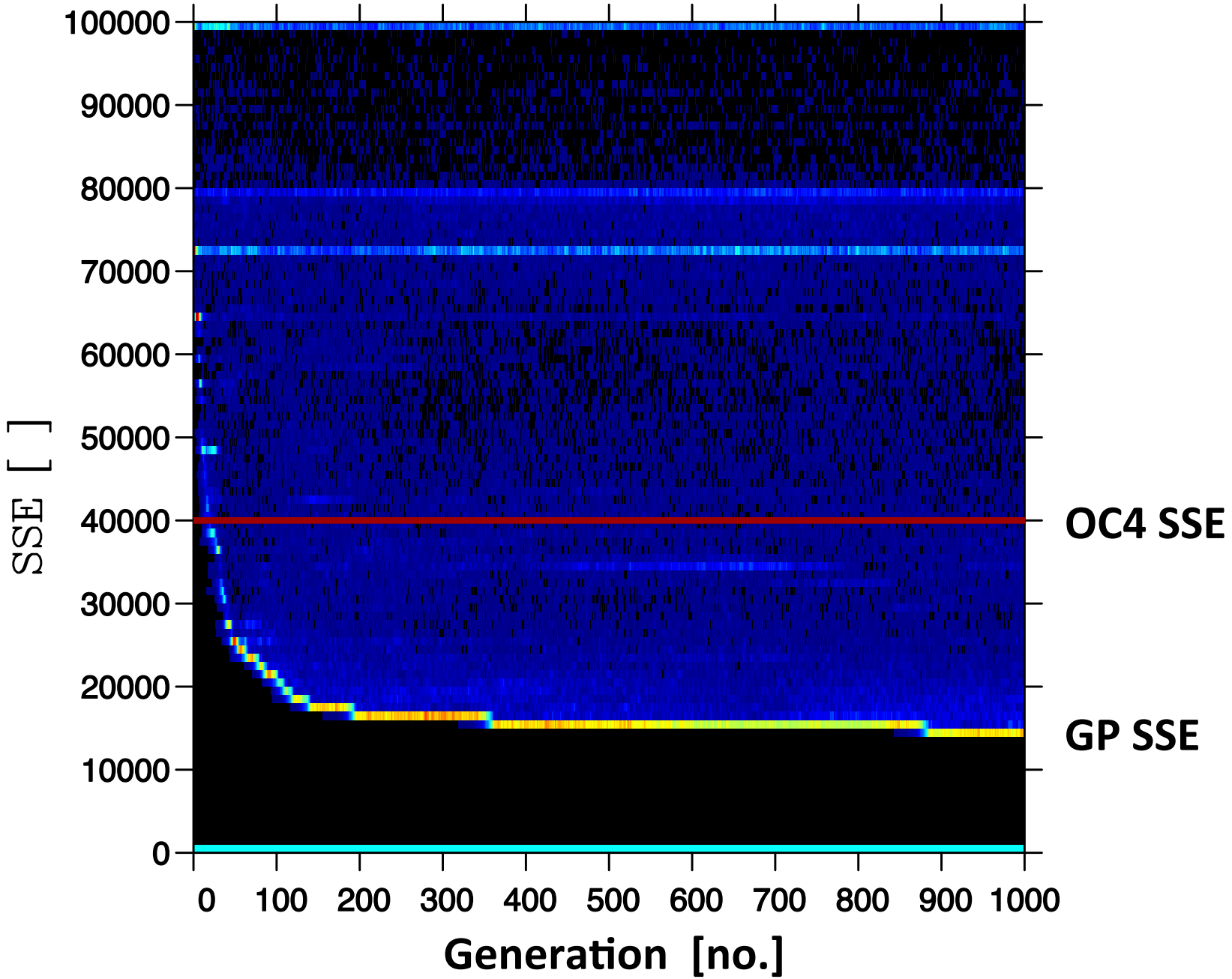

Rrs443/Rrs555
Rrs490/Rrs555
Rrs510/Rrs555
log10 (in situ chlorophyll a [mg m−3])
log10 max(Rrs(443/489/510)/Rrs(555)) [n.d.]
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
-2.00
-2.50
-0.60
-0.40
-0.20
0.00
0.20
0.40
0.60
0.80
1.00
1.20

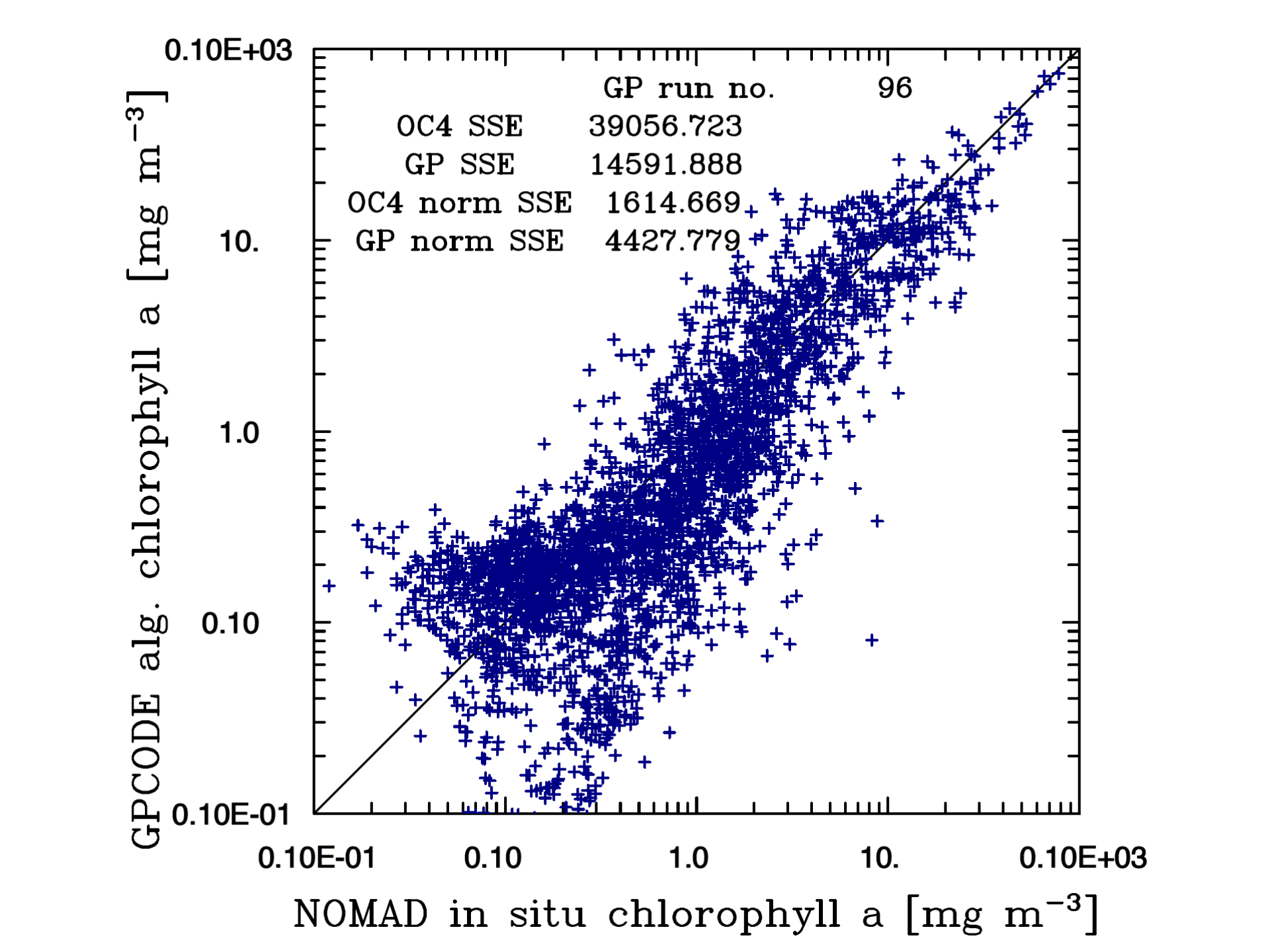

SeaWiFS OC4.v4 chlorophyll a [mg m^{-3}]
in situ chlorophyll a [mg m^{-3}]
R^2=0.514
0.10E-01
0.10
1.0
10.
0.10E+03

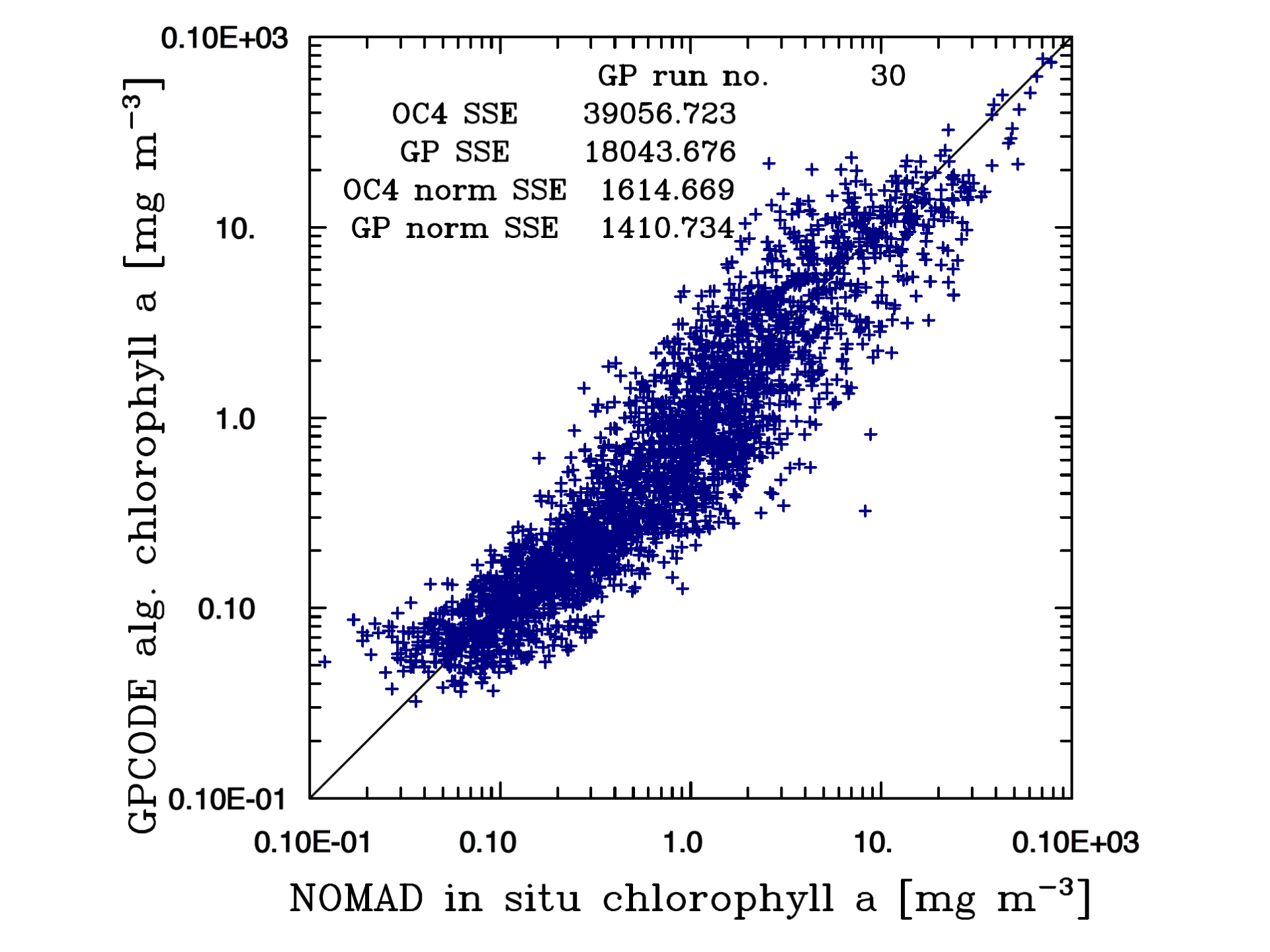

SeaWiFS OC4.v4 chlorophyll a [mg m^{-3}]
in situ chlorophyll a [mg m^{-3}]
0
10
20
30
40
50
60
70
80
90
100
Bias
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0.01
0.03
0.06
0.16
0.40
1.0
2.5
6.3
16
40
100
OC4.v4
OC4.fit

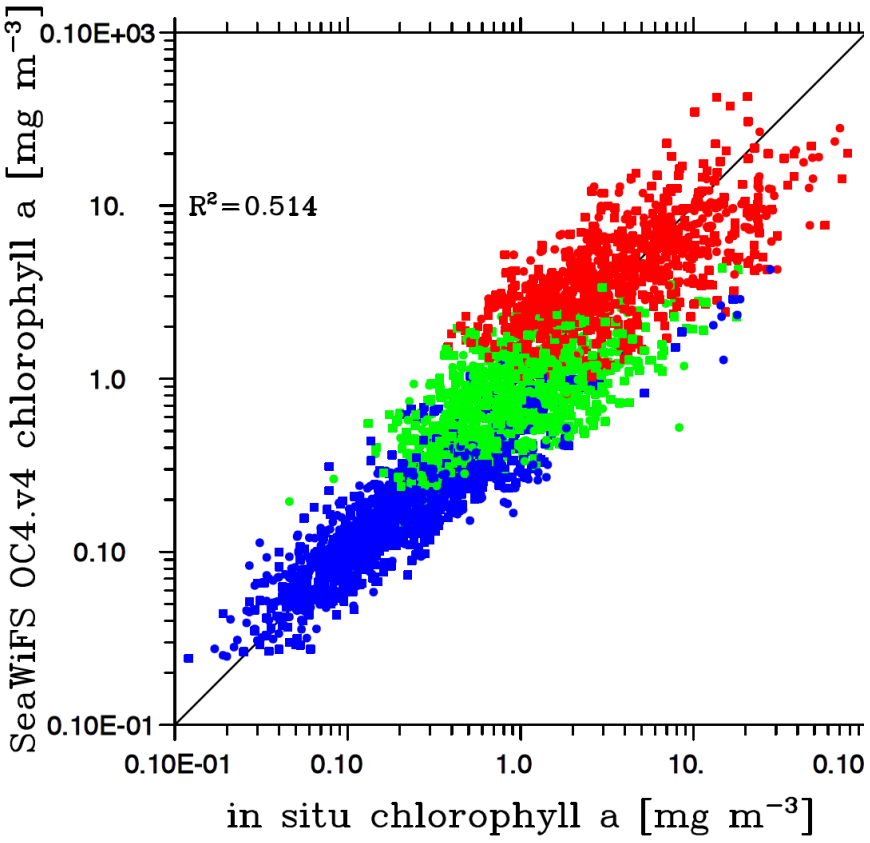

Rrs443/Rrs555
Rrs490/Rrs555
Rrs510/Rrs555
Applying OC-4 to a weekly averaged SeaWiFS Rrs global data set yields……
log10 (in situ chlorophyll a [mg m−3])
log10 max(Rrs(443/489/510)/Rrs(555)) [n.d.]
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
-2.00
-2.50
-0.60
-0.40
-0.20
0.00
0.20
0.40
0.60
0.80
1.00
1.20

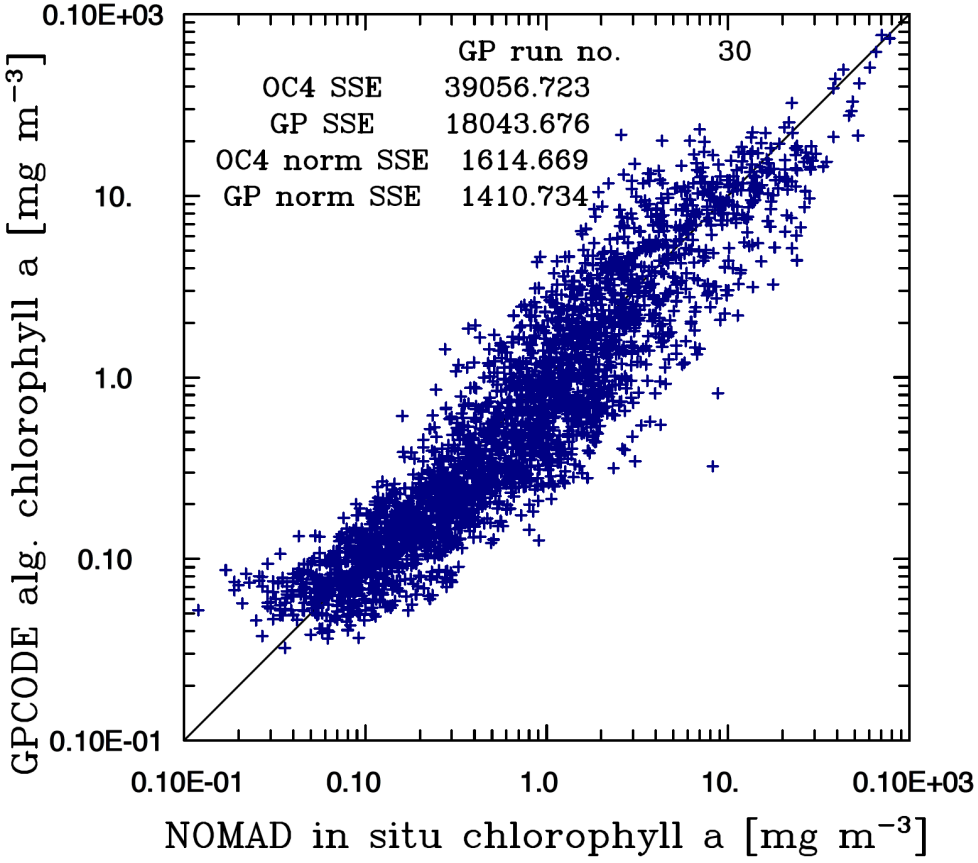

SeaWiFS OC4 Chl a [mg chla m−3]
0.00
[mg chla m−3]
63.62
90
-90
0
360


Evolution of chlorophyll algorithm with GPCODE
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
SSE [ ]
OC4 SSE
GP SSE
0
100
200
300
400
500
600
700
800
900
1000
Generation [no.]


GP run no. 96
OC4 SSE 39056.723
GP SSE 14591.888
OC4 norm SSE 1614.669
GP norm SSE 4427.779
0.10E+03
10.
1.0
0.10
0.10E-01
GPCODE alg. chlorophyll a [mg m^{-3}]
0.10E-01
0.10
1.0
10.
0.10E+03
NOMAD in situ chlorophyll a [mg m^{-3}]

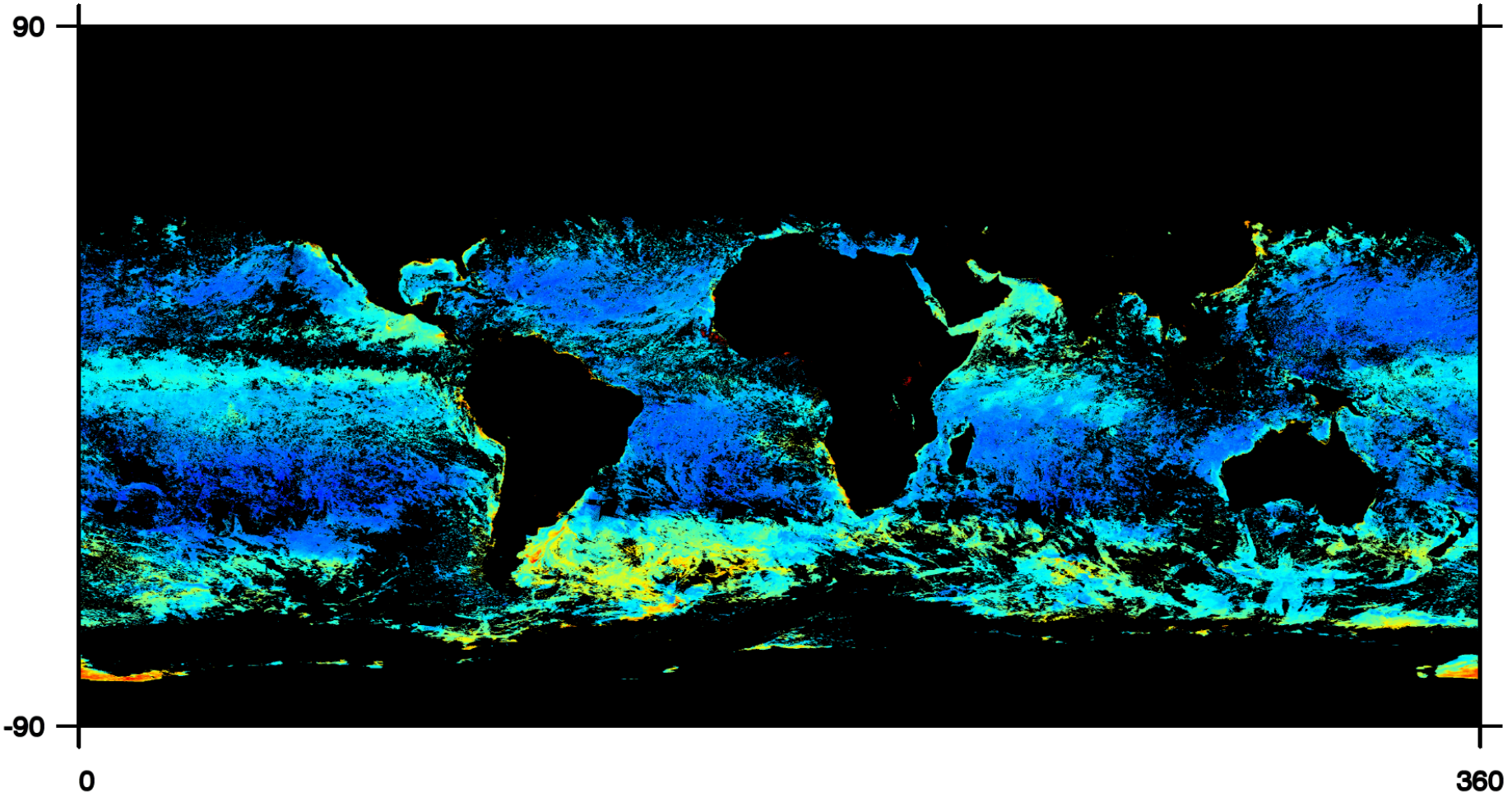

GP run no. 30
OC4 SSE 39056.723
GP SSE 18043.676
OC4 norm SSE 1614.669
GP norm SSE 1410.734
0.10E+03
10.
1.0
0.10
0.10E-01
GPCODE alg. chlorophyll a [mg m^{-3}]
0.10E-01
0.10
1.0
10.
0.10E+03
NOMAD in situ chlorophyll a [mg m^{-3}]

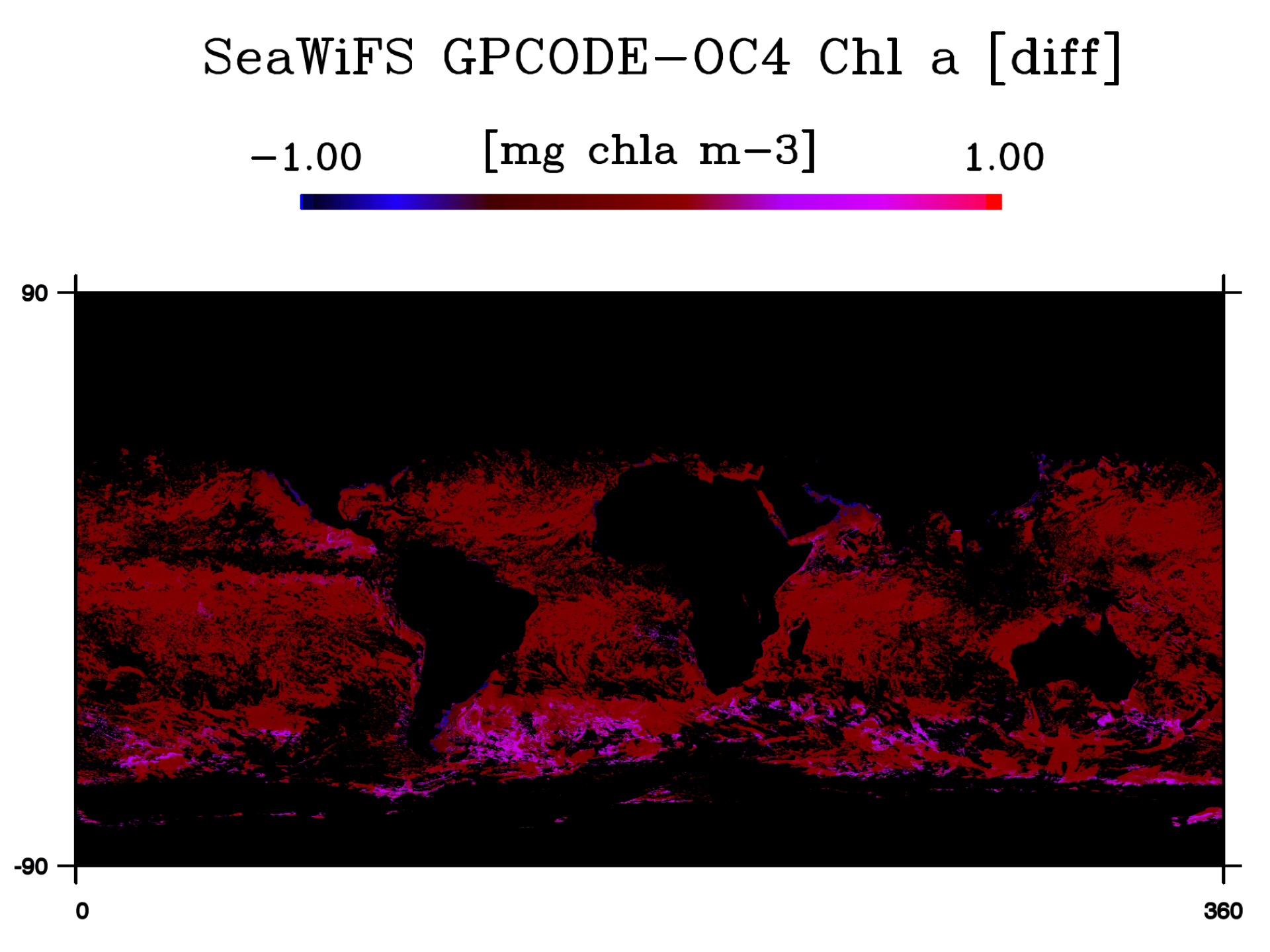

SeaWiFS OC4.v4 chlorophyll a [mg m^{-3}]
in situ chlorophyll a [mg m^{-3}]
R^2=0.514
0.10E+03
10.
1.0
0.10
0.10E-01
0.10E-01
0.10
1.0
10.
0.10

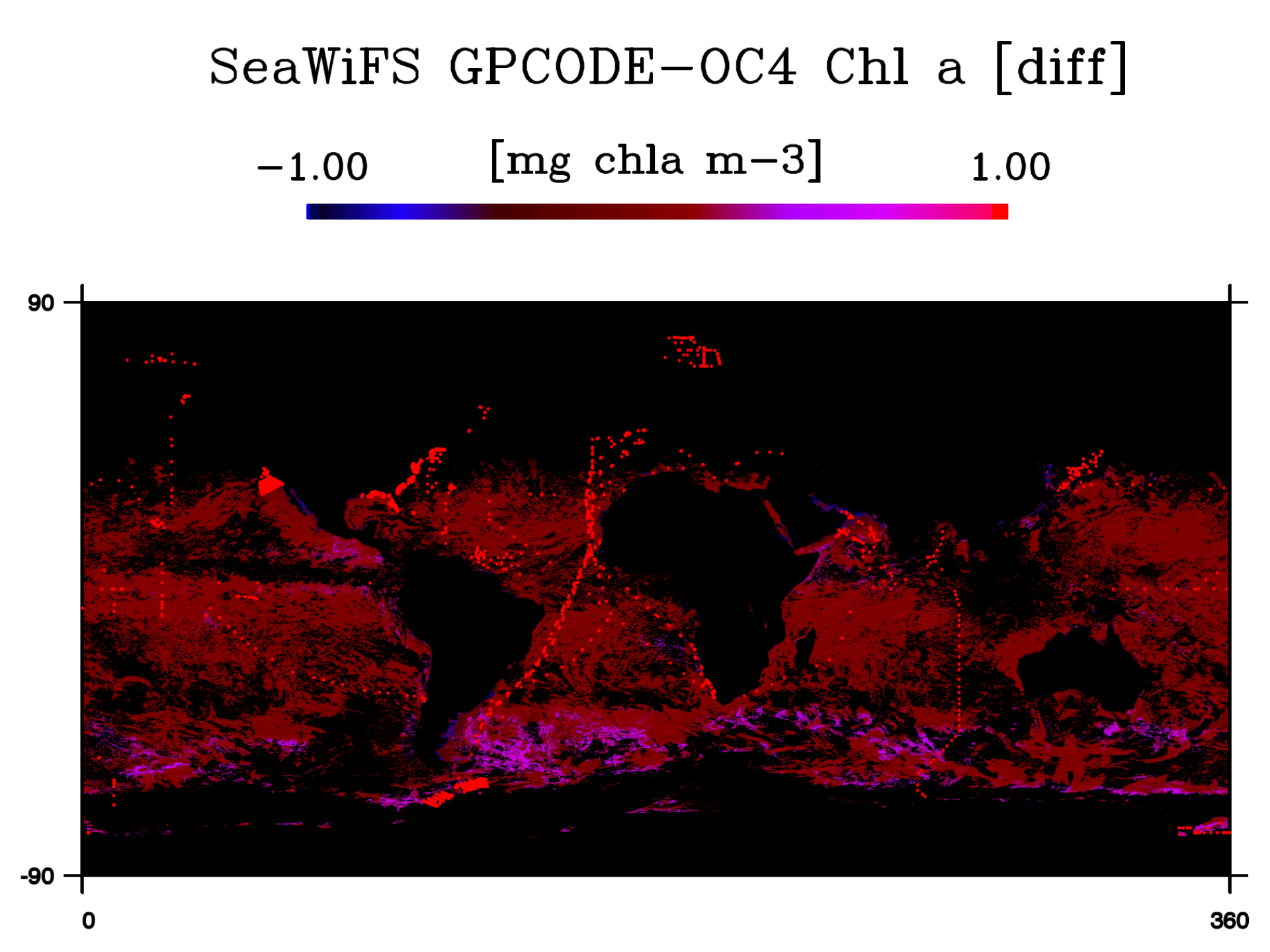

GPCODE alg. chlorophyll a [mg m^{-3}]
NOMAD in situ chlorophyll a [mg m^{-3}]
GP run no. 30
OC4 SSE 39056.723
GP SSE 18043.676
OC4 norm SSE 1614.669
GP norm SSE 1410.734
0.10E+03
10.
1.0
0.10
0.10E-01
0.10E-01
0.10
1.0
10.
0.10E+03

SeaWiFS GPCODE Chl a [mg chla m−3]

0.00 [mg chla m−3] 74.99




SeaWiFS GPCODE-OC4 Chl a [diff]
-1.00
[mg chla m-3]
1.00
90
-90
0
360


SeaWiFS GPCODE−OC4 Chl a [diff]
−1.00
[mg chla m−3]
1.00
90
-90
0
360

Use of Genetic Programming to Evolve Ecosystem Models

Satellite Algorithms: S

Box Models

$$\frac{dB_i}{dt} = S_i$$

One Dimensional (Vertical) Models

$$\frac{\partial B_i}{\partial t} + (w + w_B)\frac{\partial B_i}{\partial z} - \frac{\partial}{\partial z} K_z \frac{\partial B_i}{\partial z} = S_i$$

Three Dimensional Models

$$\frac{dB_i}{dt} + (\vec{v} + \vec{v}_B)\bullet\nabla B_i - \nabla\bullet(\vec{K}\nabla B_i) = S_i$$

Box Models

$$\frac{dB_i}{dt} = S_i \Longrightarrow \left[\frac{dP}{dt} = S_P ; \frac{dN}{dt} = S_N ; \frac{dZ}{dt} = S_Z \right]$$

$$\frac{dN}{dt} = S_N = -\rho_{\max} \frac{N}{K_N + N} P + \mu Z + (1-\beta) g_{\max} \frac{P}{K_P + P} Z + \delta Z^2$$

$$\frac{dP}{dt} = S_P = \rho_{\max} \frac{N}{K_N + N} P - g_{\max} \frac{P}{K_P + P} Z$$

$$\frac{dZ}{dt} = S_Z = \beta g_{\max} \frac{P}{K_P + P} Z - \mu Z - \delta Z^2$$

Box Models

$$\frac{dB_i}{dt} = S_i \Longrightarrow \left[\frac{dP}{dt} = S_P; \frac{dN}{dt} = S_N; \frac{dZ}{dt} = S_Z \right]$$

$$\frac{dN}{dt} = S_N = -\rho_{\max} \frac{N}{K_N + N} P + \mu Z + (1-\beta) g_{\max} \frac{P}{K_P + P} Z + \delta Z^2$$

$$\frac{dP}{dt} = S_P = \rho_{\max} \frac{N}{K_N + N} P - g_{\max} \frac{P}{K_P + P} Z$$

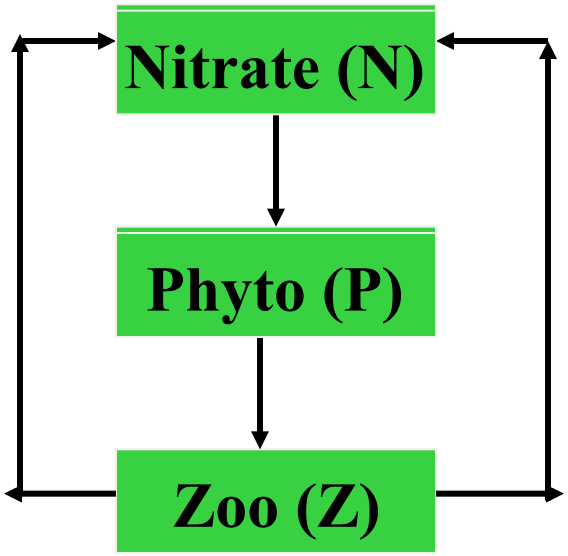
$$\frac{dZ}{dt} = S_Z = \beta g_{\max} \frac{P}{K_P + P} Z - \mu Z - \delta Z^2$$

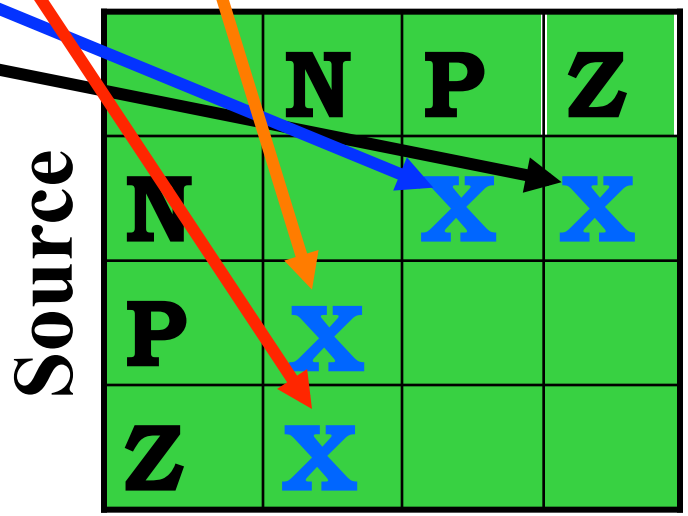
Ecosystem/Food Web Models can be represented as a matrix of terms/equations

$$\frac{dN}{dt} = -\rho_{max}\frac{N}{K_N + N}P + \mu Z + g_{max}\frac{P}{K_P + P}Z - \beta g_{max}\frac{P}{K_P + P}Z$$

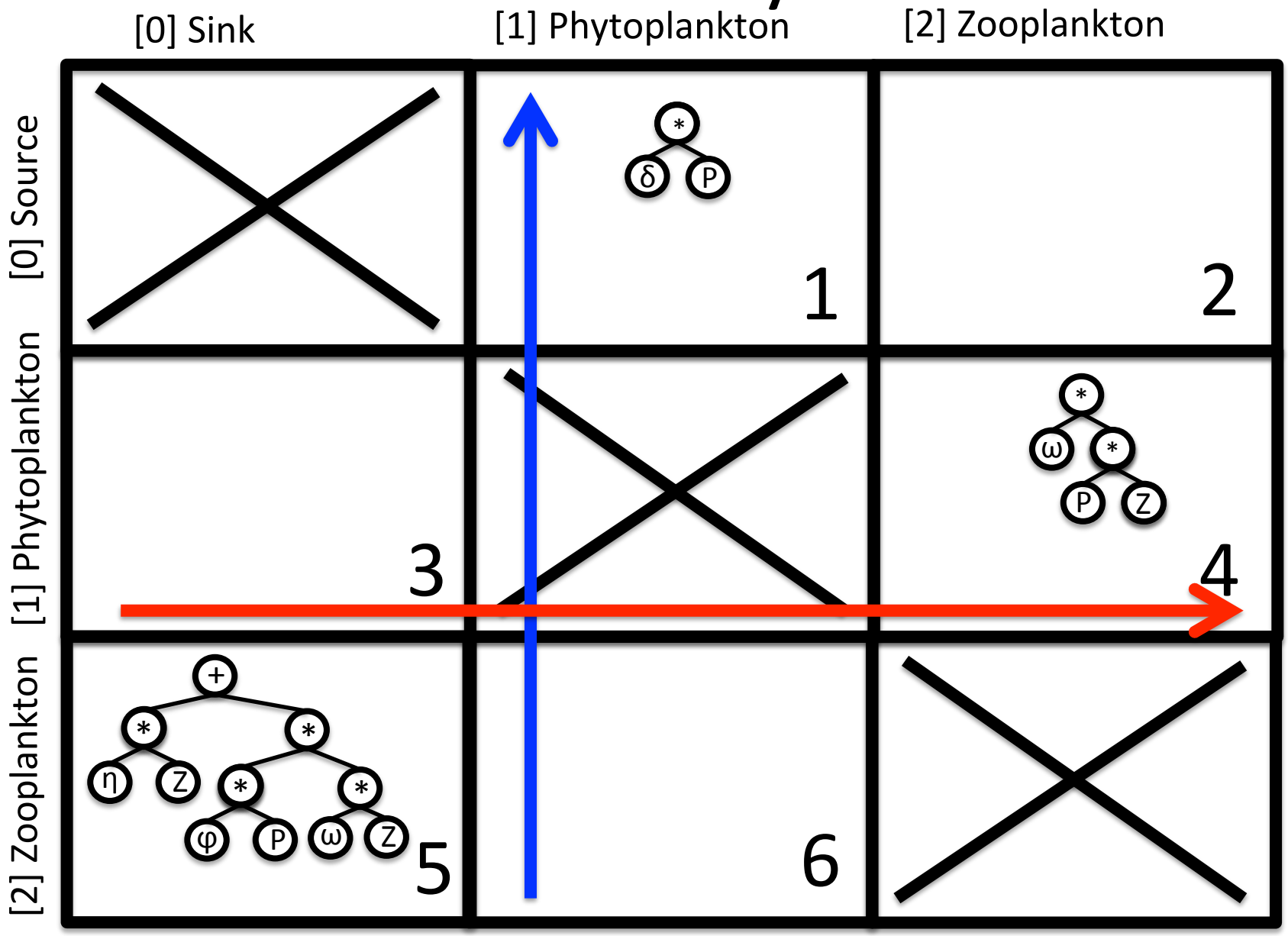
$$\frac{dP}{dt} = \rho_{max}\frac{N}{K_N + N}P - g_{max}\frac{P}{K_P + P}Z$$

$$\frac{dZ}{dt} = \beta g_{max}\frac{P}{K_P + P}Z - \mu Z$$



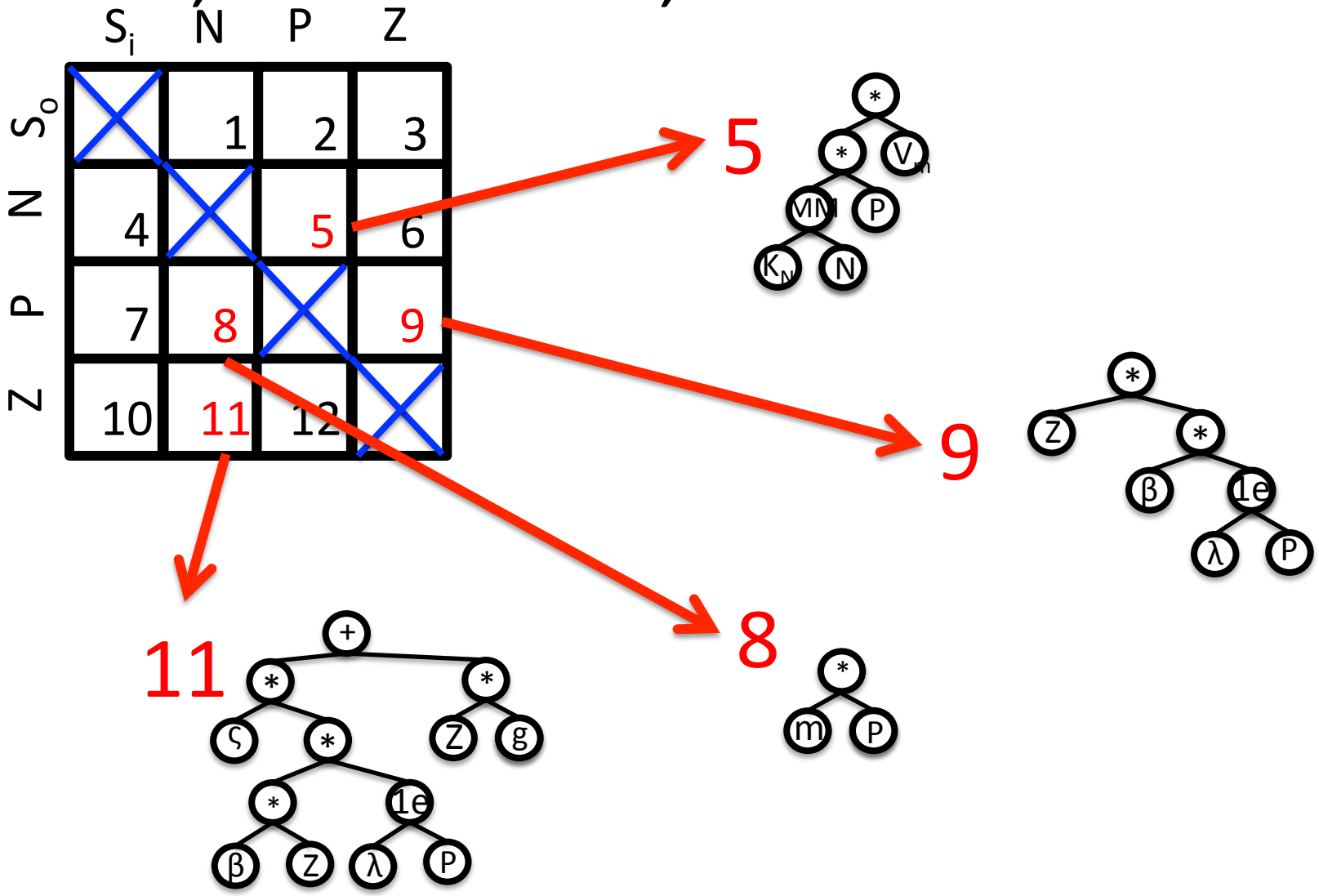


Lotka-Volterra System N=2



Total RHS = Sum of Source Terms – Sum of Sink Terms; Total Possible Equations = $(N)^2-(N)=6$

NPZ, Franks et al., Mar. Bio. 1986 N=3



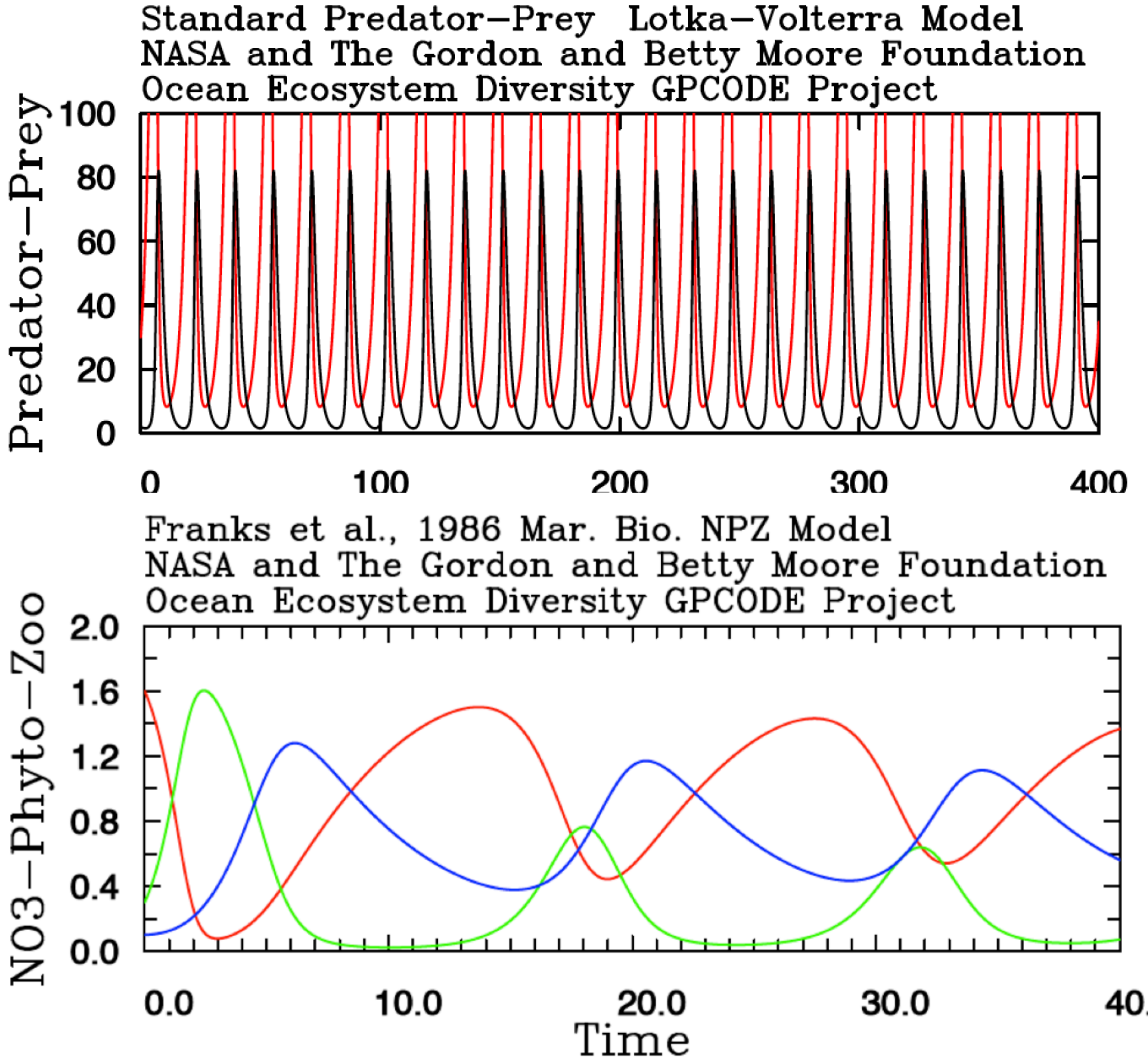
Total RHS = Sum of Source Terms – Sum of Sink Terms;

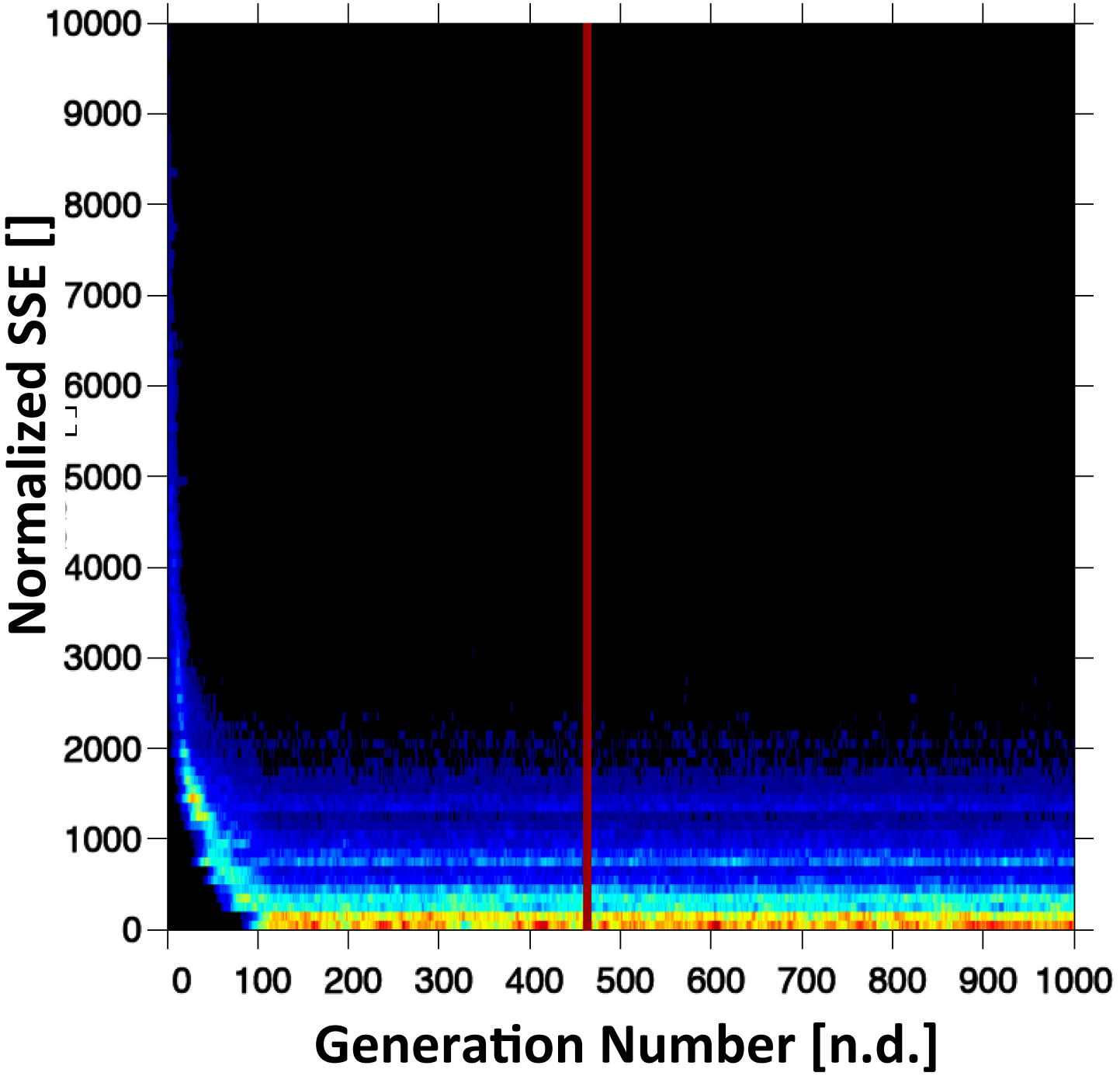
Total Possible Equations = $(N+1)^2-(N+1)=12$; N=3

Ecosystem Model Twin Experiments

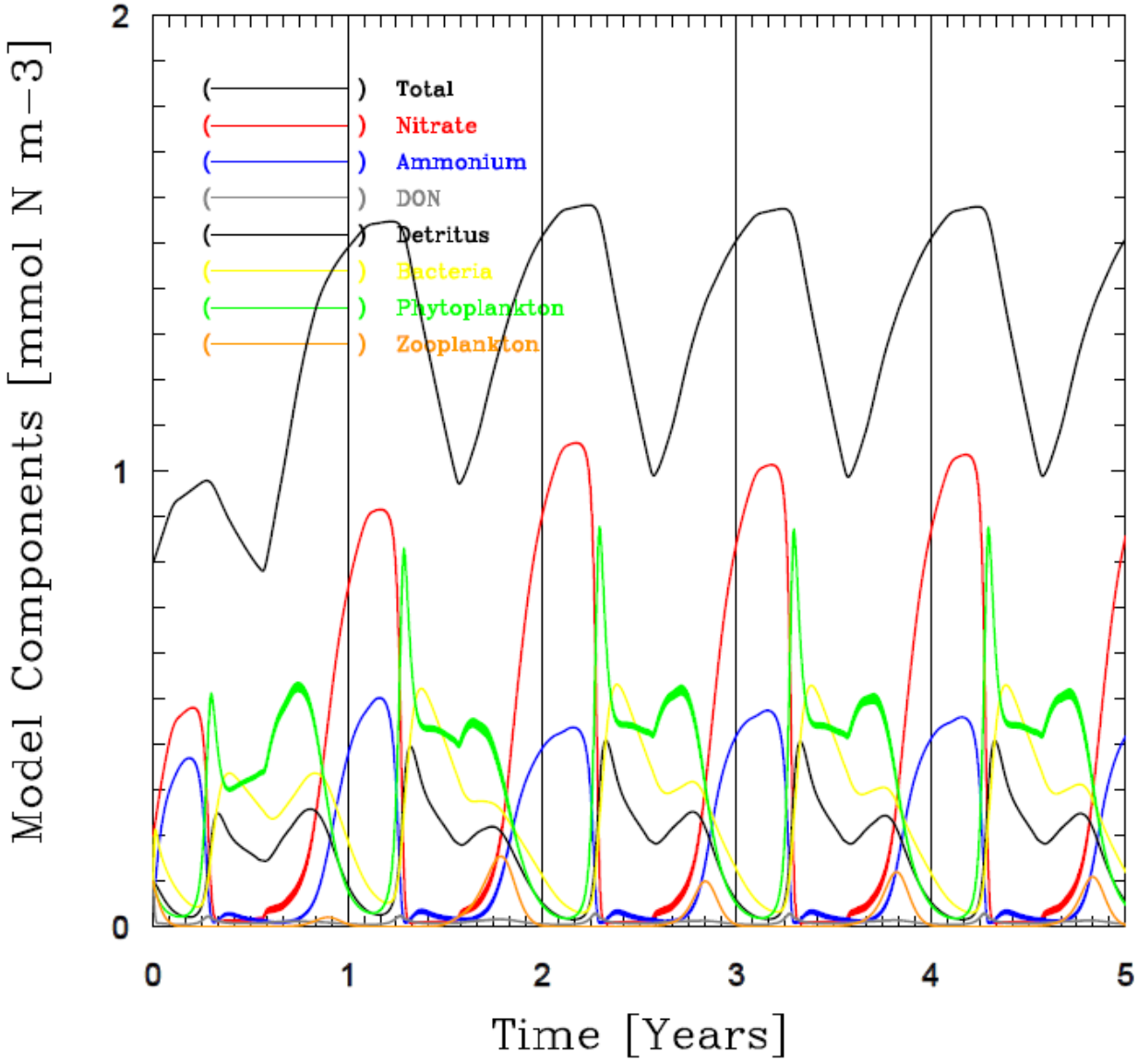
- Lotka-Volterra Equation Set (2 system ODE)
  - Closed system model dependent on I.C.
  - Completed
- NPZ model of Franks et al. 1986 Mar. Biol.
  - Open system box model, dependent on forcing.
  - Completed
- Fasham et al., 1990 Mixed-Layer Model
  - Mixed-layer model, dependent on forcing.
  - Completed.
- Huisman & Weissing, 1999 [chaos model]
  - Multi-species box model with chaotic solutions.
  - In Progress.
- ECCO Ecosystem Model
  - Global randomized ecosystem model with varying geo-located solutions.
  - In Progress.

GPCODE Box Model Twin Experiments

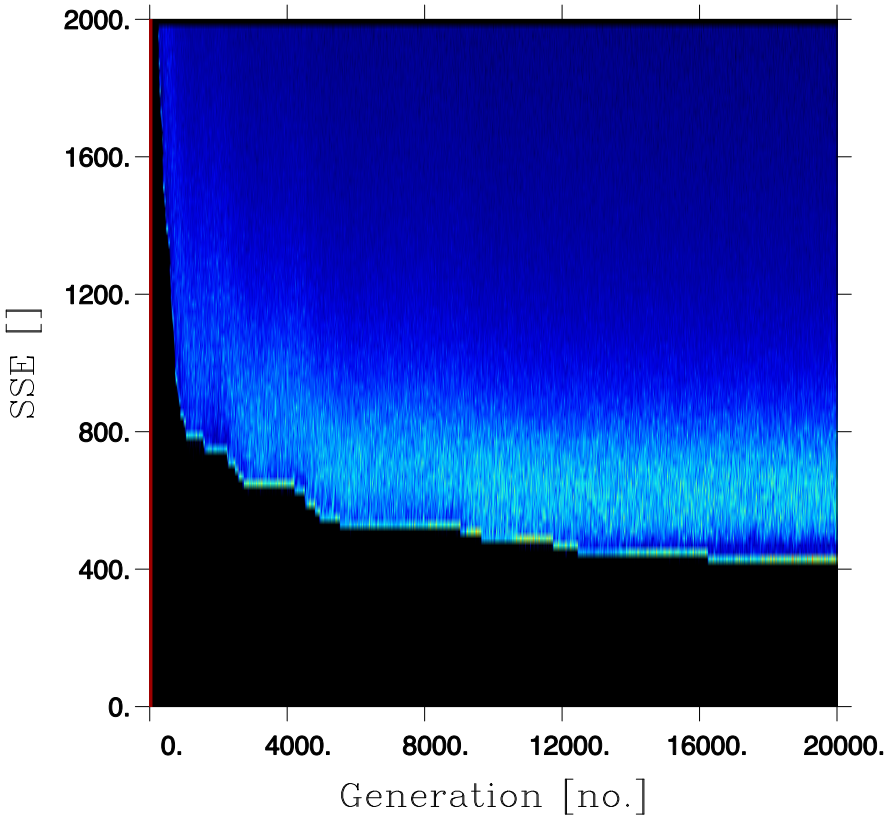



Normalized SSE []
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
600
700
800
900
1000
Generation Number [n.d.]

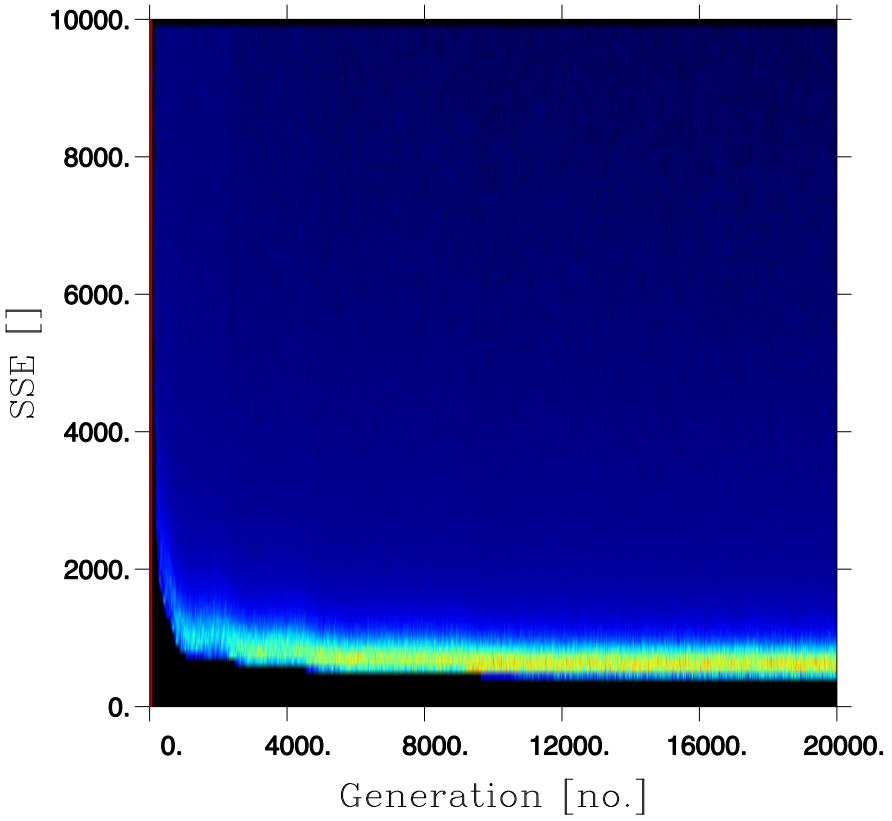
Fasham et al., 1990 Mixed-Layer Ecosystem Model (BATS Simulation)



Progression of Fasham et al. 1990 Model Twin Experiment Solution



Poor diversity causes too quick a solution and results in a poor model/outcome.

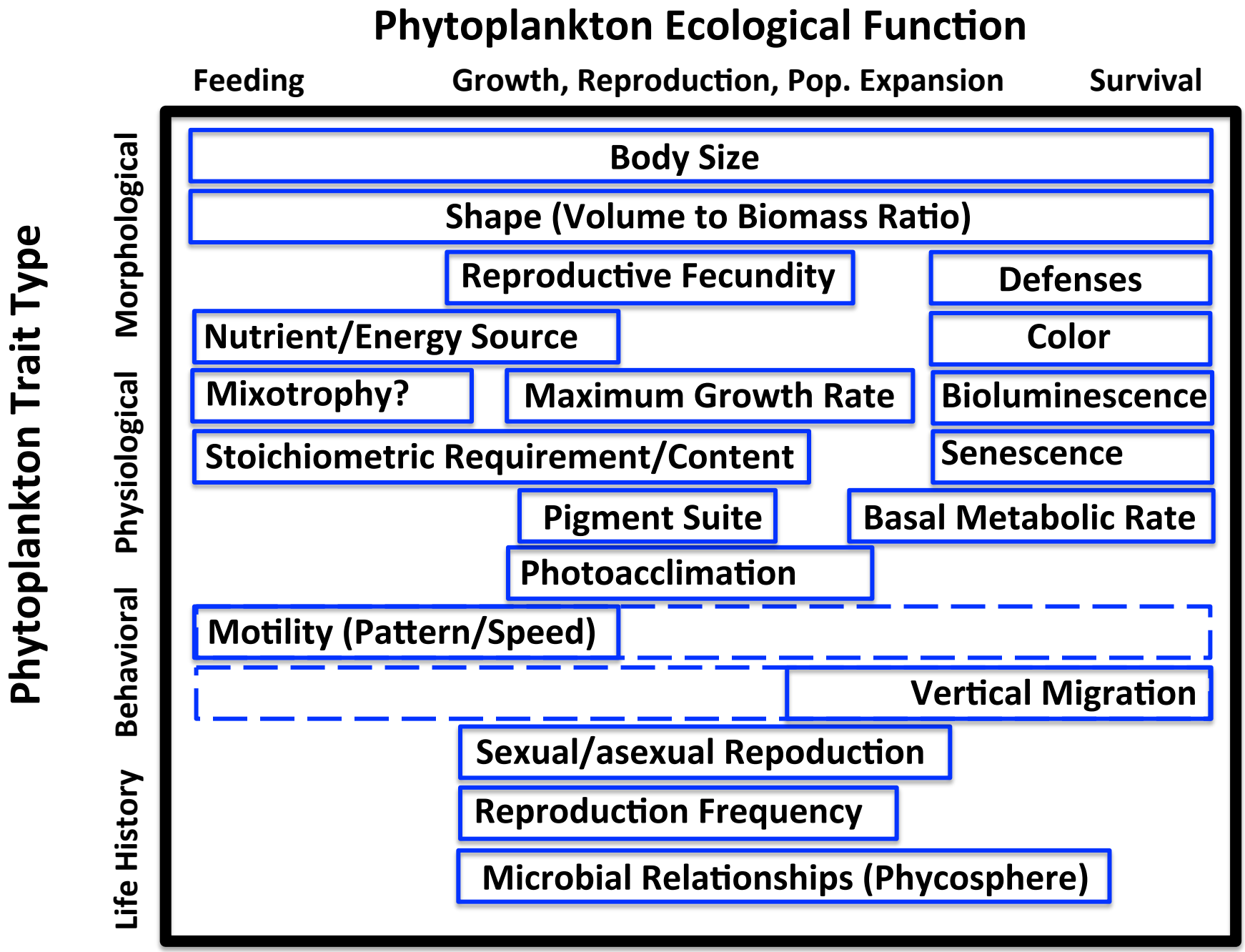


Healthy diversity supports a continued evolution to an optimal solution for a model equation set*.

*Difference in SSE levels between Left and Right panels from longer (t) solutions (N)

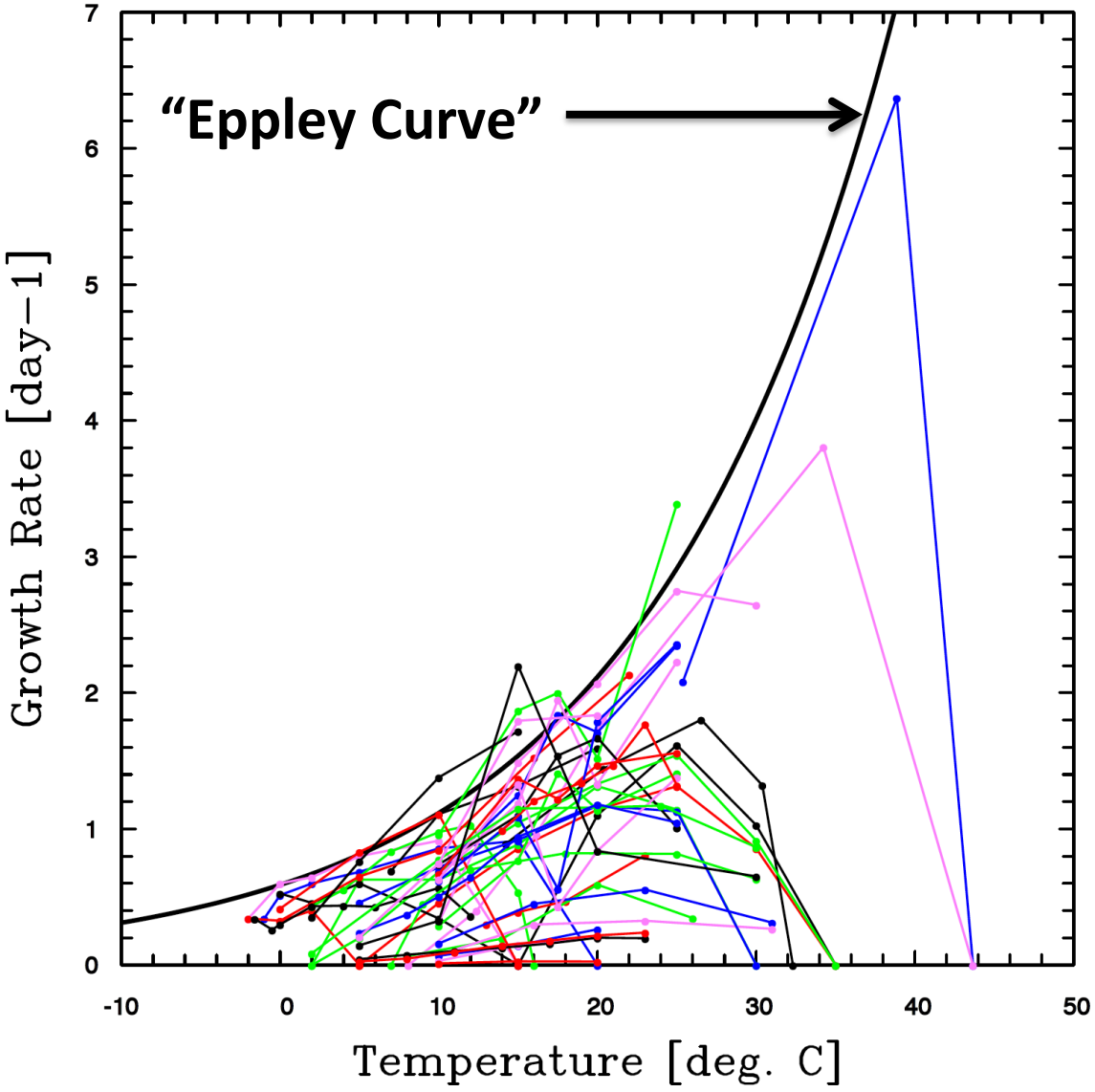
"…the simple sad fact is that not for a single clone of a plankton species can we construct a model describing its activity under a realistic range of environmental conditions. This is before we consider the impact of genetic diversity."[1]

[1]Flynn, K. J., Reply to Horizons Article 'Plankton functional type modelling: running before we can walk' Anderson (2005): II Putting trophic functionality into plankton functional types, J. Plankton Res., 9, 873-875, 2006



Adapted from: Litchman et al., J. Plankton Res. (2013) 35(3): 473–484

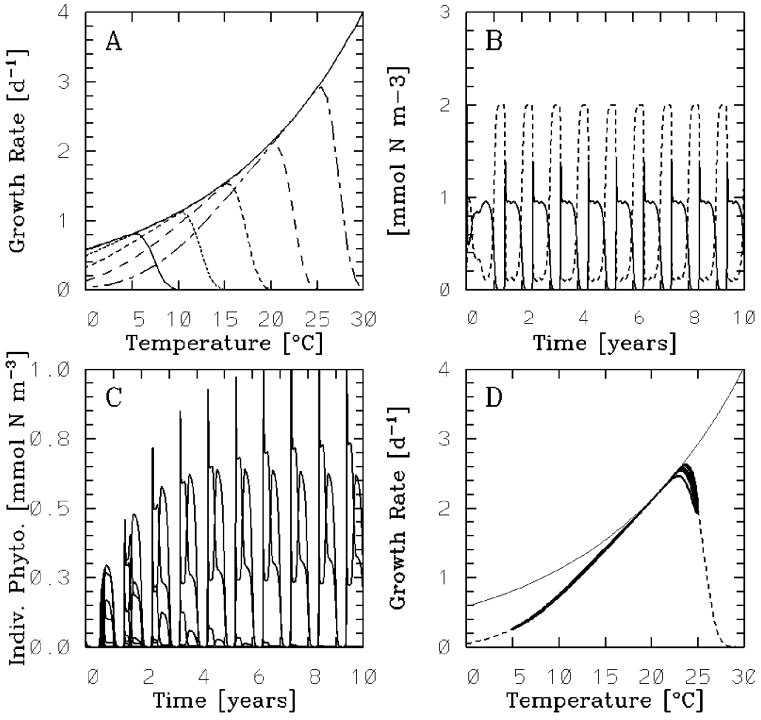
Phytoplankton Maximum Growth Rates



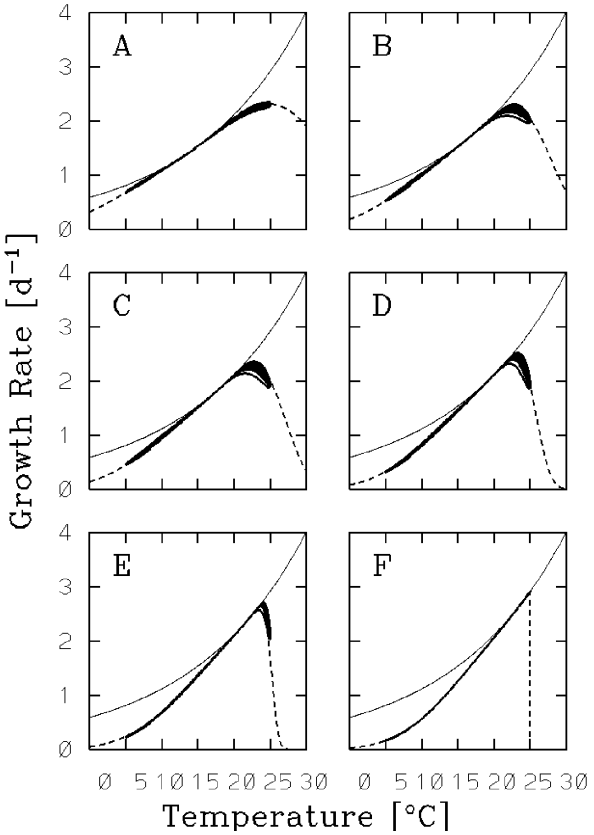
Eppley, R.W., 1972. Temperature and phytoplankton growth in the sea. Fishery Bull. 70, 1063–1085.

“Everything is everywhere, **but** the environment selects*”.

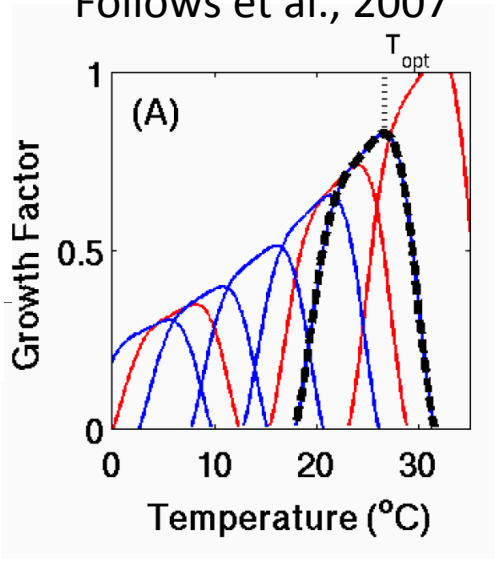
Moisan et al., 2002



Moisan et al., 2002



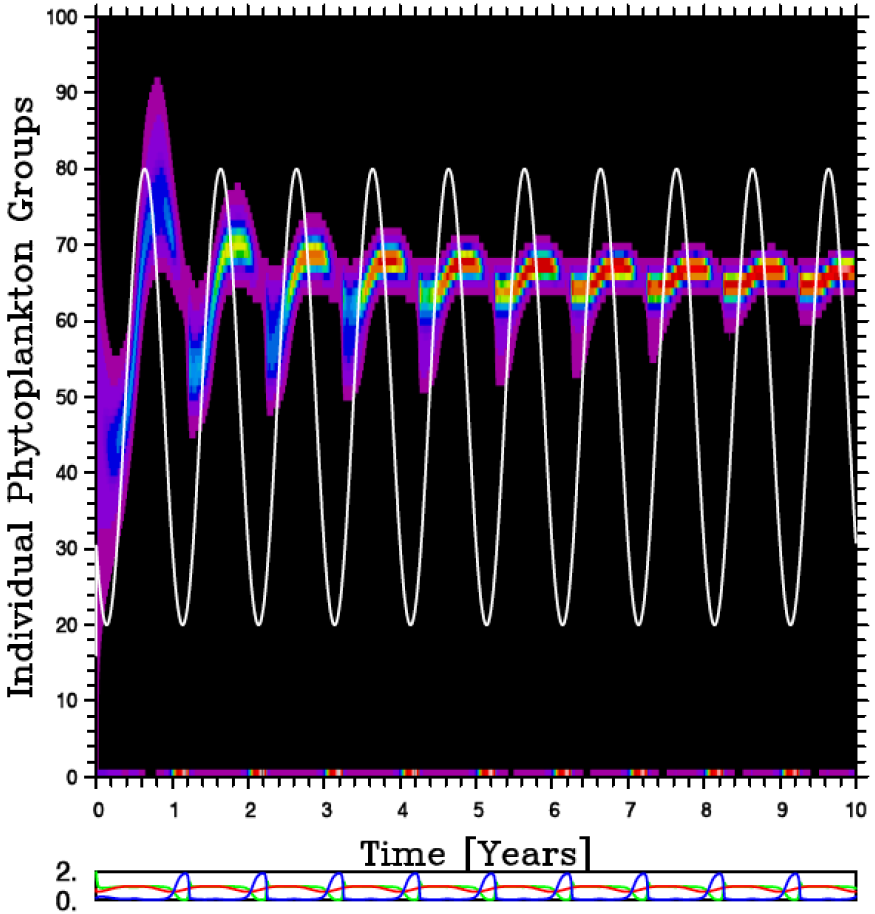
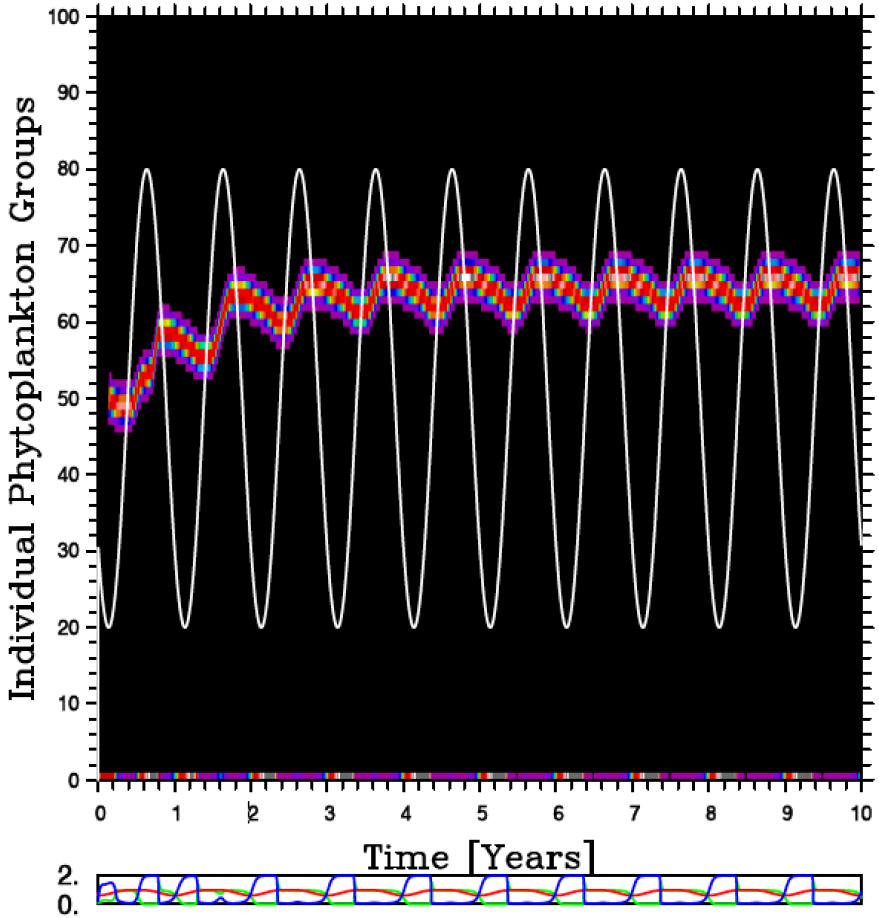
Follows et al., 2007



*Bass Becking, 1934

Linking parameters to traits for diversity and acclimation

Model with 100 phytoplankton groups



Gaussian Trait-based model (2-3 equations)