Imperial College London

How well does a mechanistic model based upon biochemical principles fit a dataset of thermal responses of individual fitness in phytoplankton?

Samuel Thompson

11/03/2015

1 Introduction

The temperature dependence of metabolic rate is a key determinant of the maximum population growth rate in many systems, a feature that has long been known to biologists and ecologists. The traditional Malthusian exponential growth equation for a population N, can be written as a function of time as

$$N(t) = N(0)e^{rt} \tag{1}$$

where r changes dependent upon the environmental variables. The maximum value, r_{max} has long been linked to both the body size of the animal and the temperature of the environment. With computer simulations allowing for unprecedented increases in the ability to use mathematical models to describe these systems, various mechanistic and phenomenological models have been developed. Models have been developed linking body size, temperature and population growth (Savage et al., 2004; Chen and Liu, 2010) and even characterising

genetic divergence based on temperature (Allen et al., 2006) . The two models of interest as examples of mechanistic and phenomenological models are, respectively, the Schoolfield model (Schoolfield et al., 1981) and the Gaussian-Gompertz model (Martin and Huey, 2008).

The Schoolfield model is defined as

$$B(T) = B_0 \frac{e^{\left(\frac{-E}{k}\left(\frac{-1}{283.15T}\right)\right)}}{1 + \frac{E}{E_D - E} e^{\frac{E_D}{k}\left(\frac{1}{T_{pk}} - \frac{1}{T}\right)}} \tag{2}$$

where $k = 8.617 \times 10^{-5} eVK^{-1}$ is the Boltzmann constant, $B(t)$ is the value of the trait at temperature t, (growth for the datasets of interest), and B_0 is the trait value at 283.15K, or 10 °C. E and E_D are the activation and deactivation energies, which control the rise and fall of the curve. T_{pk} is the temperature where the trait value is maximal. The temperature value T is defined in Kelvin.

The Schoolfield model is mechanistic, with each parameter has a biological representation based on the Arrhenius euqation and Eyring's theoretical equation (Schoolfield et al., 1981). On the other hand, the Gaussian-Gompertz model is phenomenological, with the parameters B_{max} being the maximum of the trait measurements and θ simply a quantity that improves the fit. Neither parameters have any biological significance, and temperature is here measured in celsius. The model is defined as

$$B(T) = B_{max} e^{\left(-E(T - T_{pk})^2 - e^{(E_D(T - T_{pk} - \theta)}\right)} \tag{3}$$

In order to assess how well the mechanistic model can fit the dataset of thermal responses, it was compared against the Gaussian-Gompertz model and a cubic model an example of another phenomenological model. The cubic model was simply defined as the general cubic equation,

$$B(T) = a + bT + cT^2 + dT^3 \tag{4}$$

with a, b, c and d not having any biological bearing. The cubic model can therefore be used to compare to the other models, as it has a similar number of parameters.

2 Method

To investigate the ability of the mechanistic model to fit thermal responses, a dataset was used from 855 experimental studies on phytoplankton from around the world, compiled by Chen and Liu (2010) and Thomas et al. (2012). For each experiment, growth traits were recorded for a number of individuals at different temperatures, which would then be fitted using Non-linear Least Squares (NLLS) analysis. This approximates the model to a linear model and refines the parameters over successive iterations to minimise the difference between the model trait values and actual trait values. Models are then selected by the Akaike Information Criterion (Akaike, 1973), and Bayesian Information Criterion (Schwarz, 1978) to decide which produces the best fit for the dataset.

Before fitting, the data was first analysed to produce an estimate to aid model fitting, based off the maximum trait values, and using a linear model to estimate B_0, B_{max}, T_{pk}, E and E_D. The starting values for these parameters would be used to speed up model fitting and were based off the gradient, intercept and minimum value of the linear model, which disregarded points with temperature greater than the maximal trait value. The linear model produced to estimate these parameters uses a logarithmic function to linearise the points to a straight line, as shown in Figure 1.

At this point, the datasets were constrained to remove those with too few data points for meaningful model generation. A minimum of 5 unique points was required to ensure that the model had fewer parameters than the number of data points, otherwise over-fitting would occur. Recordings without trait or temperature values were also removed, as for the purpose of model fitting they would be meaningless. After filtering out these datasets, a total of 651 datasets were available to find parameter estimates for and begin fitting to the three models.

For the Schoolfield model, the parameters B_0, E, E_D and T_{pk} were all allowed to vary.

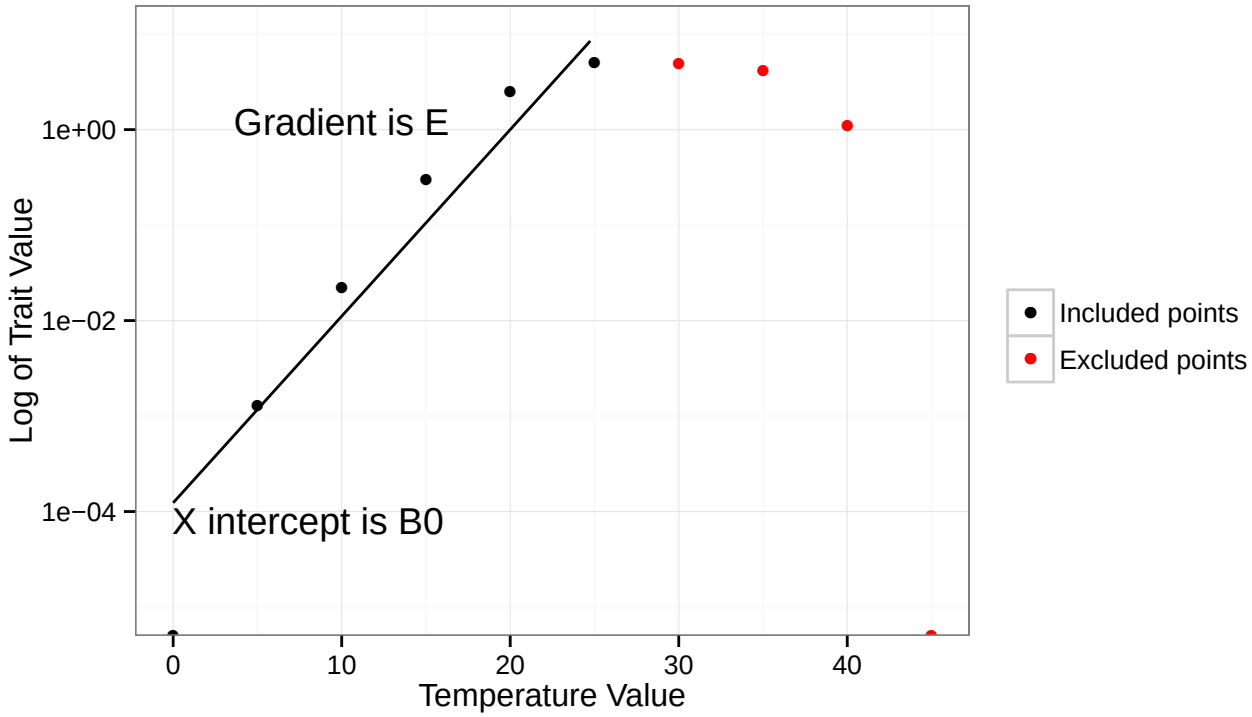

Figure 1: Example plot of initial linear model to produce parameter values. E_D is then estimated as $\frac{1}{2}E$ and T_{pk} is estimated as the maximal trait value observed.

The E and E_D values constrained to positive values as they represent biological activation energies, which are assumed to be positive, and have $E > E_D$.

The Gaussian-Gompertz model has parameters B_{max}, E, E_D and T_{pk}, and all were allowed to vary over any number except for B_{max}, which was constrained to the experimental maximum trait value as it has a phenomenological interpretation. For the cubic model all parameters, a,b,c and d were allowed to vary, as none have any biological interpretation.

During the fitting process, `Python` (Rossum and Drake, 2011) was used with the `lmfit` function. For the cubic and Schoolfield models, non-linear least squares fitting was performed using the Levenberg-Marquardt algorithm (Levenberg, 1944). However, this didn't produce fantastic fits for the Gaussian-Gompertz models, so the Nelder-Mead algorithm (Nelder and Mead, 1965) was instead utilised here.

3 Results

As a measure of the decency of fits for each of the datasets, individual R-squared (R^2) values were calculated, and graphs plotted overlaying each of the models for every dataset. It appeared that for data with simple sigmoidal increases in trait value with temperature, likely as no experimental values existed for temperatures greater than the T_{pk}, all three models fitted with low errors and a high level of fit. An example is shown in Figure 2A.

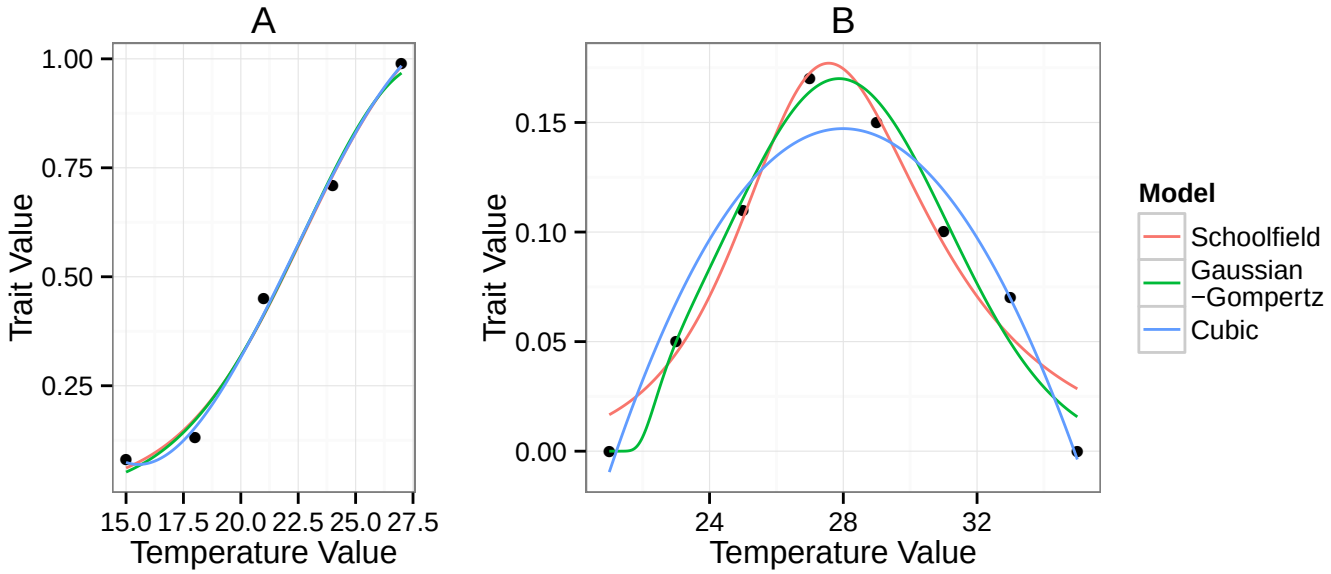

Figure 2: **A)** Example of the sigmoidal shape of some of the datasets which easily produced fits. Here the data was of *Synechococcus* species (Mackey et al., 2013). **B)** Example of all models fitting well, with a defined peak as seen in some datasets. Here the data was of *Prorocentrum mexicanum* species (Morton et al., 1992)

Other models with well defined peaks also produced good fits for all three models, as shown in Figure 2B. In these cases, most often the Schoolfield model produced the best fit as measured by the AIC and BIC, likely due to the same number of variables that exist in the model, but increased fit. In these cases, it appears that a mechanistic model fits the dataset better than the two phenomenological models. Moreover, there were many cases where the cubic and Schoolfield model produced reasonable fits for the data, whilst the Gaussian-

Gompertz model struggled with low numbers of data points and very low trait values.

The jumps seen in the model produced in Figure 3 would likely not be representative of real-world systems. The smoother curve produced by the Schoolfield model is more likely to be biologically accurate, as well as allowing for much more accurate extrapolation outside of the experimental data range.

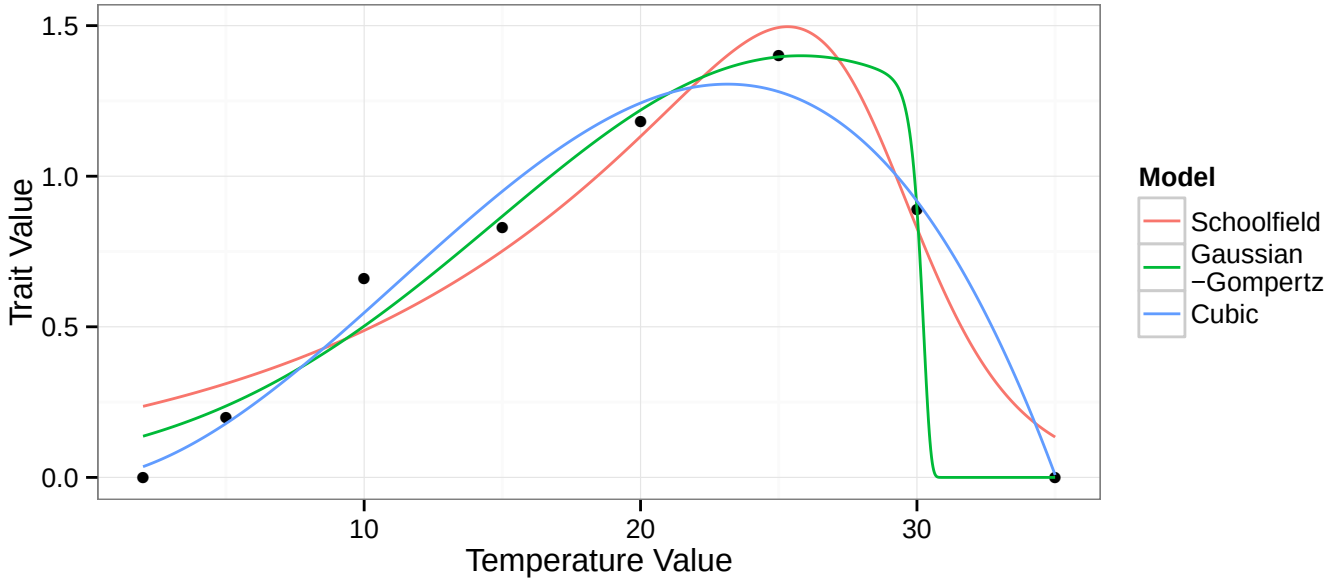

Figure 3: Gaussian-Gompertz model producing poorer fits for trait values equal to 0. The data is of *Pandorina morum* from Moss (1973).

Looking at all the datasets, by analysing the spread of the R^2 values, it gives an indication of how many of the models actually produced meaningful fits. Plotting the density of the R^2 values for every model (Figure 4) is interesting due to the high accuracy of the cubic fit. This is likely because it has the temperature as a parameter raised to multiple powers and therefore has more freedom to accurately define the path of the curve.

By defining a "good" fit to be one with greater than 0.75 R^2 value, we can then determine the total number of models in every dataset that produce good fits (Figure 6). Unsuprisingly given the R^2 distribution, the number of poor fits in the Gaussian-Gompertz model was far higher than for the other two, with the cubic model appearing slightly more successful than

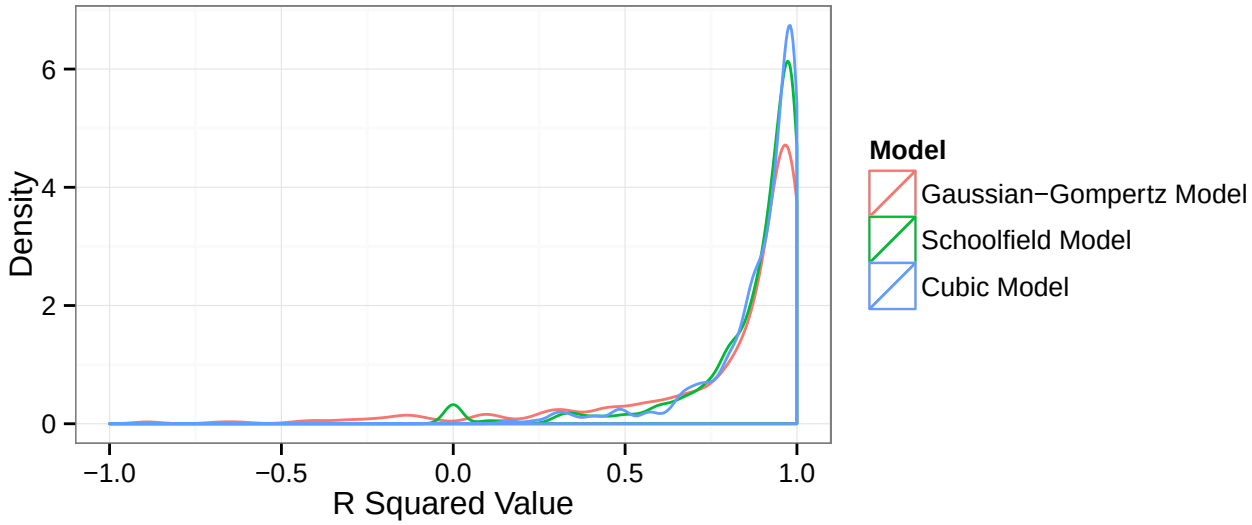

Figure 4: Spread of the R^2 values for the different models.

the Schoolfield model for producing good quality fits. Again, this can be attributed to the higher-power terms in the cubic function. Using the AIC and BIC as measures of the quality of a the model allows us to select the best model for each dataset (Figure 5). 32% of datasets had the best fit with the Schoolfield model, 22% with the Gausian-Gompertz model and the remaining 46% saw the best fit with the cubic model.

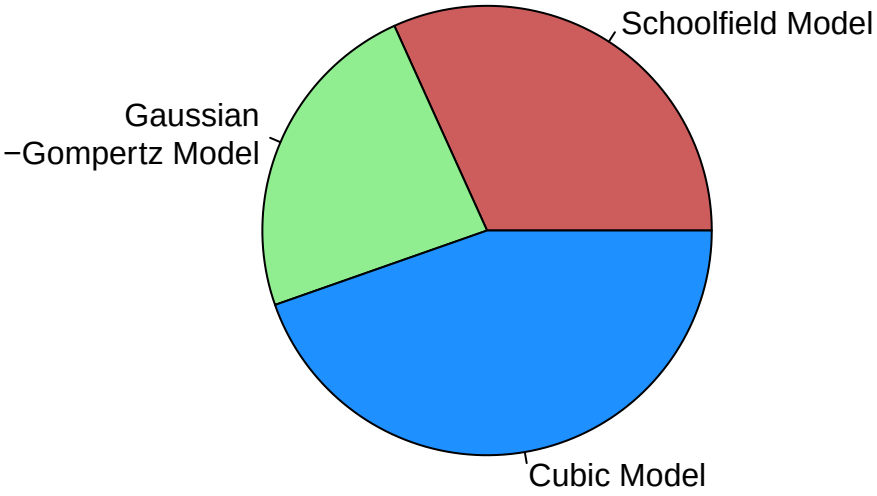

Figure 5: Best model selection for the datasets.

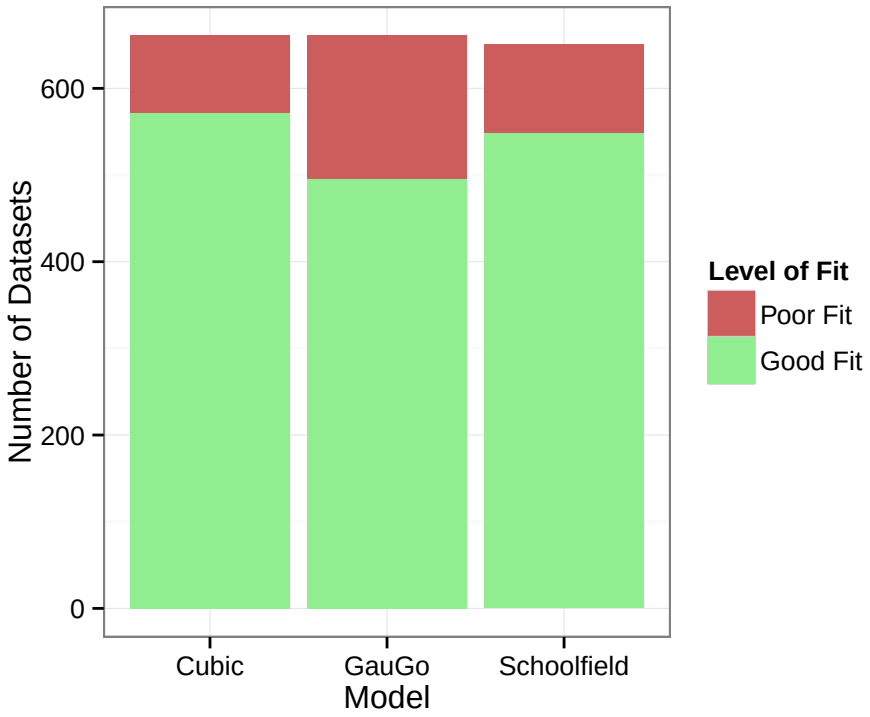

Figure 6: Number of datasets that produce good and poor fits for each model category.

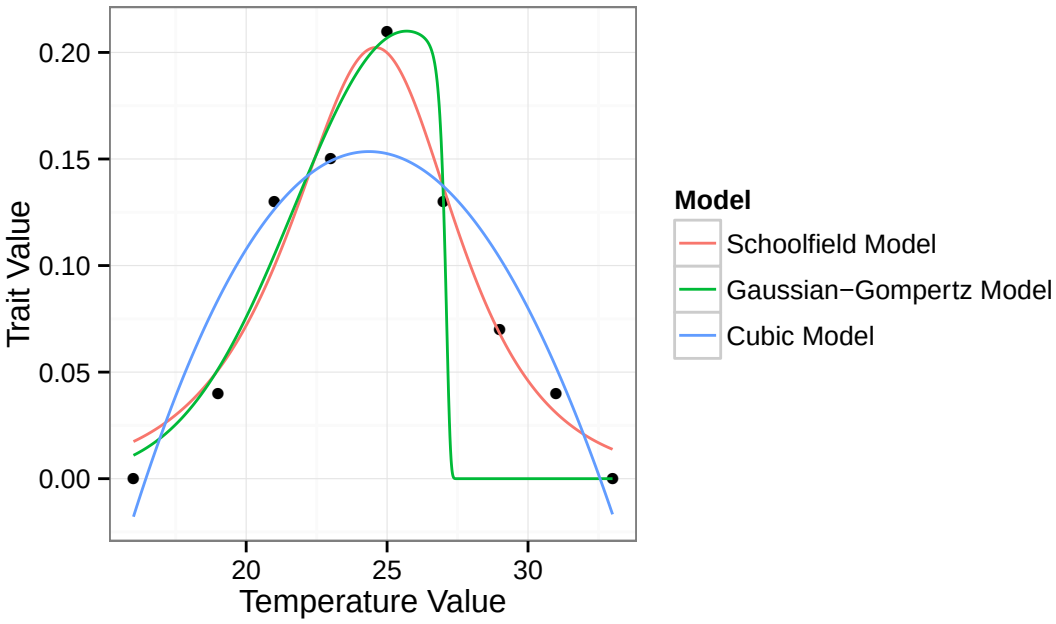

Figure 7: Examples of varying strategies of fits for different models.

4 Discussion

From much of the data produced, it appears that with a better distribution of R^2 values, the cubic model produces the best fit for the most datasets. However, as the cubic model is entirely mathematical, with none of the terms having a basis in biological attributes, it makes producing a model for a different organism without full data very difficult.Contrastingly, given measurements of E, E_D, B_0 and T_pk, a Schoolfield model could be generated for any organism, outlining the advantage of a mechanistic model.

Furthermore, many of the curves for cubic models, whilst producing a higher R^2 value than the other models, had peaks far lower than the recorded maximal trait value and produced a curve which likely wouldn't work if there were more data points recorded. It could be argued that for many datasets, including those for *Pandorina morum* (Figure 3) and for *Ostreopsis siamensis* (Figure 7), the cubic fit does not produce biologically accurate curves. The smooth

curve does not peak near the value for the recorded maximum trait, and the curve falls off extremely quickly at the minimal trait values. For temperatures outside the range of those recorded, this would mean high negative trait values, which would be very unlikely biologically. The tapering curves displayed at extreme temperature values by both the Gaussian-Gompertz and Schoolfield models would be far more likely to accurately fit experimental data.

Nevertheless, in terms of raw fitting ability over the datasets provided, the R^2 distribution and AIC values of the models leads to our conclusion being that the non-mechanistic cubic model best fits the data. Interestingly, when the best model is calculated just for models with more than 15 datasets, the proportion of datasets best fitting each model does not change much (Figure 8). This is potentially due to clustering of the data points in the optimum range, meaning that extreme values of temperatures aren't as important when calculating the best fit, removing the advantage that might have been predicted for the Gaussian-Gompertz and Schoolfield models.

Even so,the advantage of having a mechanistic model can be hotly debated (O'Connor et al., 2007), especially as the data provided here has the Schoolfield model generally producing a better fit than the counterpart phenomenological model, the Gaussian-Gompertz model. If we remove the cubic model as a comparative, a larger proportion fit the Schoolfield model (Figure 9). The spread of fits is also better, with Figure 4 showing a greater peak for high R^2 values for the Schoolfield model than the Gaussian-Gompertz model.

Overall the results suggest that mechanistic models have great potential to accurately fit thermal responses in phytoplankton as well as providing a biologically meaningful basis for the model. Similar model fitting for experiments with a greater number of data points over a larger range of temperature values would provide insights into the predictive level of the models produced already. These models could help indicate whether the cubic model is actually as powerful a tool as highlighted here, and possibly give further reasoning behind choosing mechanistic models for thermal responses.

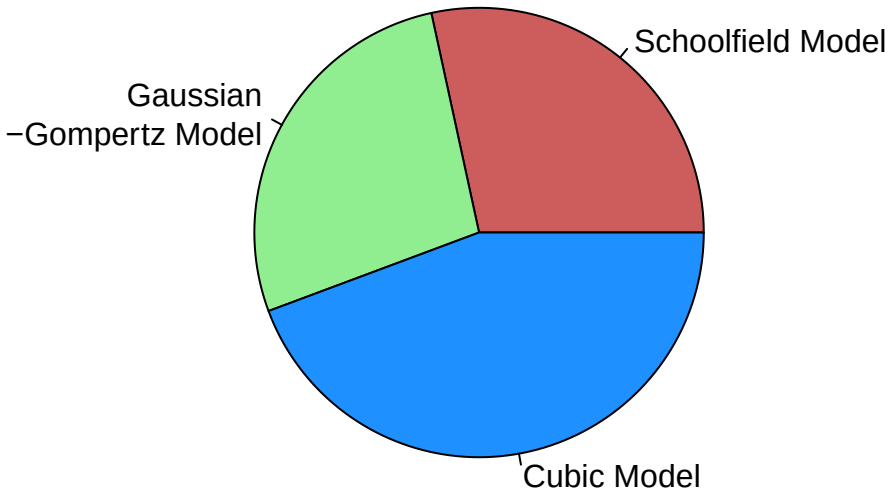

Figure 8: Best model selection for the datasets with more than 15 data points.

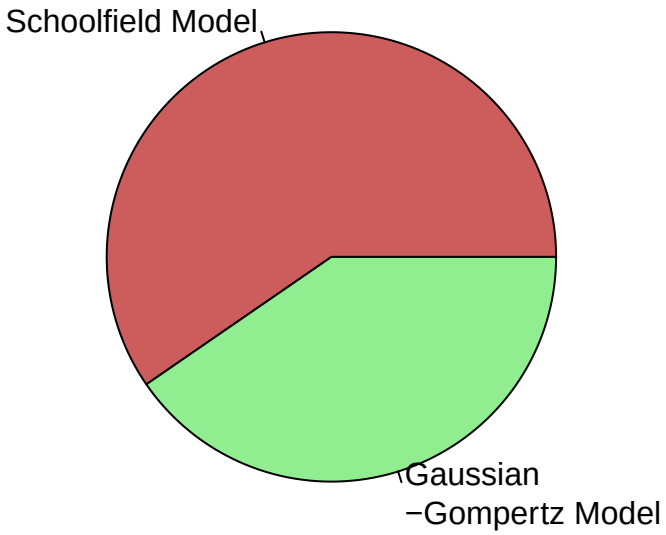

Figure 9: Best model selection between Gaussian-Gompertz and Schoolfield models.

References

Htrotugu Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, August 1973. ISSN 0006-3444. doi: 10.1093/biomet/60.2.255. URL http://biomet.oxfordjournals.org/content/60/2/255.short.

Andrew P Allen, James F Gillooly, Van M Savage, and James H Brown. Kinetic effects of temperature on rates of genetic divergence and speciation. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (24):9130–5, June 2006. ISSN 0027-8424. doi: 10.1073/pnas.0603587103. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1474011.

Bingzhang Chen and Hongbin Liu. Relationships between phytoplankton growth and cell size in surface oceans: Interactive effects of temperature, nutrients, and grazing. *Limnology and Oceanography*, 55(3):965–972, 2010. ISSN 00243590. doi: 10.4319/lo.2010.55.3.0965. URL http://www.aslo.org/lo/toc/vol_55/issue_3/0965.html.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics*, II(2):164 – 168, 1944.

Katherine R M Mackey, Adina Paytan, Ken Caldeira, Arthur R Grossman, Dawn Moran, Matthew McIlvin, and Mak A Saito. Effect of temperature on photosynthesis and growth in marine Synechococcus spp. *Plant physiology*, 163(2): 815–29, October 2013. ISSN 1532-2548. doi: 10.1104/pp.113.221937. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3793060.

Tara Laine Martin and Raymond B Huey. Why "suboptimal" is optimal: Jensen's inequality and ectotherm thermal preferences. *The American naturalist*, 171 (3):E102–18, March 2008. ISSN 1537-5323. doi: 10.1086/527502. URL http://www.ncbi.nlm.nih.gov/pubmed/18271721.

Steve L. Morton, Dean R. Norris, and Jeffrey W. Bomber. Effect of temperature, salinity and light intensity on the growth and seasonality of toxic dinoflagellates associated with ciguatera. *Journal of Experimental Marine Biology and Ecology*, 157 (1):79–90, May 1992. ISSN 00220981. doi: 10.1016/0022-0981(92)90076-M. URL http://www.sciencedirect.com/science/article/pii/002209819290076M.

B Moss. The influence of environmental factors on the distribution of freshwater algae: an experimental study: II. The role of pH and the carbon dioxide-bicarbonate. *The Journal of Ecology*, 61(1):157–177, 1973. URL http://www.jstor.org/stable/2258925.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL http://comjnl.oxfordjournals.org/content/7/4/308.short?rss=1&ssource=mfc.

Michael P. O'Connor, Stanley J. Kemp, Salvatore J. Agosta, Frank Hansen, Annette E. Sieg, Bryan P. Wallace, James N. McNair, and Arthur E. Dunham. Reconsidering the mechanistic basis of the metabolic theory of ecology. *Oikos*, 116(6):1058–1072, June 2007. ISSN 00301299. doi: 10.1111/j.0030-1299.2007.15534.x. URL http://doi.wiley.com/10.1111/j.0030-1299.2007.15534.x.

Guido Van Rossum and Fred L Jr Drake. *The Python Language Reference Manual*. Network Theory Ltd., March 2011. ISBN 9781906966140.

Van M Savage, James F Gilloly, James H Brown, and Eric L Charnov. Effects of body size and temperature on population growth. *The American naturalist*, 163(3):429–41, March 2004. ISSN 1537-5323. doi: 10.1086/381872. URL http://www.ncbi.nlm.nih.gov/pubmed/15026978.

RM Schoolfield, PJH Sharpe, and CE Magnuson. Non-linear regression of biological temperature-dependent rate models based on absolute reaction-

rate theory. *Journal of Theoretical Biology*, 0:719–731, 1981. URL http://www.sciencedirect.com/science/article/pii/0022519381902460.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 2168-8966. URL http://projecteuclid.org/euclid.aos/1176344136.

Mridul K Thomas, Colin T Kremer, Christopher a Klausmeier, and Elena Litchman. A global pattern of thermal adaptation in marine phytoplankton. *Science (New York, N.Y.)*, 338(6110):1085–8, November 2012. ISSN 1095-9203. doi: 10.1126/science.1224836. URL http://www.ncbi.nlm.nih.gov/pubmed/23112294.