Comparison of Performance of Models Fitting Organisms in Different Habitats

Jiqiu Wu

j.wu18@imperial.ac.uk

Department of Life Sciences

Contents

Abstract

Accurate prediction of the impact of climate change on metabolic rate has overwhelmingly practical siginificance. This study shows that the phenomenal model has stronger predicting ability. And much more attention and patience are supposed to pay to mechanistic models. In order to get the best prediction, scholars should choose models and model selection criteria with caution.

1 Introduction

Metabolism theory provides a foundation for using first principles of physics, chemistry, and biology to link the performance from the molecular level(Allen et al. 2006) the ecology level of population(Savage et al. 2004), community and ecosystem(Brown et al. 2004). The theoretical basis is the core of metabolism consists of a small number of reactions that are ultimately dependent on temperature(Gillooly et al. 2001).

Climate change is projected to have significant effects on public health(Woodward et al. 2014), agriculture(Howden et al. 2007), and fishing(Brander 2007). Prediction of the impacts of climate change on freshwater, marine, and terrestrial environments can assist the management of agriculture, fishery, and wildlife conservation(Prowse et al. 2009).

Here, a large dataset called BioTraits(Della et al. 2013), one phenomenological models, Cubic model, one mechanistic models with three forms, Schoolfield model, Schoolfield model for low temperature, and Schoolfield model for high temperature model, three selection criteria, AIC, BIC, and AICc were used to study two main questions: (1) Is there any difference in the performance of organisms in different habitats fitted by different models? (2) If the answer of Question (1) is YES, so what is the best model in different habitats?

In this study, the results show: (1) The differences in performances exist; (2) In most cases, phenomenological model fits much better than mechanistic models. By different model selection, the best mechanistic model varies under different environments.

Consequently, researchers should choose the most appropriate model selection criteria according to the habitat where the research organism live in, the size of sample, and the number of parameters, and choose the best model cautiously.

2 Methods

2.1 Data

BioTraits(Della et al. 2013) that consists of 25826 rows of data from 2165 researches from small scale, like photosynthesis rate and respiration rate to large scale, like population growth rate. The habitats included marine, terrestrial, and freshwater environments. At kingdom level, the organisms were composed of archaea, bacteria, chromista, metazoa, fungi, plantae, and protista.

But a large number of low quality data exist, NA, zero, and negetive value. If the body temperature of consumer were missing, and ambient temperature replaced. Besides, the researches with less than 8 data points were filtered out.

2.2 Start values calculating for Schoolfield model

Before fitting School model to the dataset, the starting value of the parameters were calculated. First, the optimal temperature at which the trait value is highest. Before and after this optimal temperature, linear regression analysis were conducted twice respectively.

E is the slope for the part after the optimal temperature, and its biological interpretation is the activation energy. E_h is the slope before the optical temperature, and the biological interpretation is the de-activation energy of enzyme at high temperature. T_h is the at which the enzyme is 50% high-temperature deactivated, and was determined as the mean value of the maximal and optical temperature. E_l is the slope for the part before the optimal temperature, and the biological interpretation is the de-activation energy of enzyme at low temperature. T_l is the at which the enzyme is 50% low-temperature deactivated, and was determined as the mean value of the minimal and optical temperature.

Schoofield model would be discussed further in the following models section.

2.3 Models

2.3.1 Phenomenological model

Generally, the thermal performance curves have unimodal shape. Cubic model(Callaway 1959) are often used to study. This is Cubic model:

$$B = B_O + B_1 * T + B_2 * T^2 + B_3 * T^3 \tag{1}$$

The parameters: B_O, B_1, B_2, and B_3 don't have biological interpretations. This equation gives the trait value at the given T temperature.

2.3.2 Mechanistic models

A biological understanding of the response of metabolic rate to temperature is critical. Consequently, Schoolfiel model is used. This is full Schoolfiel model(Schoolfield et al. 1981):

$$B = \frac{B_0 e^{\frac{-E}{k}(\frac{1}{T} - \frac{1}{283.15})}}{1 + e^{\frac{E_l}{k}(\frac{1}{T_l} - \frac{1}{T})} + e^{\frac{E_h}{k}(\frac{1}{T_h} - \frac{1}{T})}} \tag{2}$$

Where k is the Boltzmann constant $(8.617 * 10^{-5} eVK^{-1})$, B the value of the trait at a given temperature T (K) (K = ^{o}C + 273.15), while B_0 is the trait value at 283.15 K (10^{o}C) which stands for the value of the growth rate at low temperature. E_l and E_h are the enzyme's low-temperature de-activation energy (eV) and high-temperature de-activation energy (eV). T_l and T_h are the at which the enzyme is 50% low-temperature deactivated and 50% high-temperature deactivated. E is the activation energy (eV).

In some cases, low temperature inactivation and high temperature are not available, the simplified models for low temperature and high temperature are more appropriate. This is the Schoolfield model for low temperature:

$$B = \frac{B_0 e^{\frac{-E}{k}(\frac{1}{T} - \frac{1}{283.15})}}{1 + e^{\frac{E_l}{k}(\frac{1}{T_l} - \frac{1}{T})}} \tag{3}$$

This is the Schoolfield model for high temperature:

$$B = \frac{B_0 e^{\frac{-E}{k}(\frac{1}{T} - \frac{1}{283.15})}}{1 + e^{\frac{E_h}{k}(\frac{1}{T_h} - \frac{1}{T})}} \tag{4}$$

2.4 Model selection criteria

Candidate models are compared to one another by evaluating the relative support in the observed data for every model(Johnson & Omland 2004). Akaike information criteria (AIC)(Burnham & Anderson 2004), Bayesian information criterion (BIC)(Schwarz et al. 1978), and Small sample unbiased AIC(AICC)(Burnham & Anderson 2004).

AS for AIC criteria, after models were fitted within each research, the minimal AIC was determined, and the delta AIC was calculated. Finally, the relative weight of every model was calculated. The best model is the model with most support(Burnham & Anderson 2004).

2.5 Statistics analysis

P values were calculated using Kruskal-Wallis test(Vargha & Delaney 1998) to compare the significant difference in multiple groups. In two groups, p values were calculated using Mann-Whitney test(Ruxton 2006). The reason is that the normal distribution and consistent variance can't be guaranteed.

2.6 Visualizaton

The histogram suing point geom with the size of samples within one research, the boxplot with weight supported by AIC, BIC, and AICC, and model fitting figure with one specific research were plotted using ggplot2 (Wickham 2016) in R.

2.7 Computing languages

Bash 4.3.48: It was used to combine the pipeline together and compile the report into pdf format.

Python 3.5.2: It was used to model fitting. Some helpful packages were also used, $lmfit$(Newville et al. 2016), $pandas$ (McKinney 2011), $numpy$(Van Der Walt et al. 2011), and $scipy$ (Millman & Aivazis 2011).

R 3.2.3: It was used to pre-process data, calculate the starting value, plot the results and calculate the p values. Some helpful packages were used, $dplyr$ (Wickham et al. 2015), $plyr$(Wickham 2009), $ggplot2$ (Wickham 2016), and $reshape2$(Wickham 2012).

3 Results

3.1 Data

After removing the low quality data, the modified Biotraits contains 17525 rows of data from 971 researches. Here are the summaries of the modified Biotraits. In this dataset, Terrestrial and freshwater environments dominate over marine environments.

Table 1: Summary of researches included in different habitats. Terrestrial and freshwater environments dominate.

Habitats	Number of researches
Terrestrial	443
Marine	51
Freshwater	274

In this dataset, Bateria, fungi, metazoa, and plantae dominate over archaea, chromista, and protista.

Table 2: Summary of researches on organisms at Kindom level. Bateria, fungi, metazoa, and plantae dominate.

Kingdoms	Number of researches
Archaea	24
Bacteria	205
Chromista	16
Fungi	122
Metazoa	236
Plantae	366
Protista	1

Also, the size of samples within one research was summarised. The minimal is 8, because researches with less than 8 data point were filtered out. And the maximal is 637. Most researches consists of less than 50 samples.

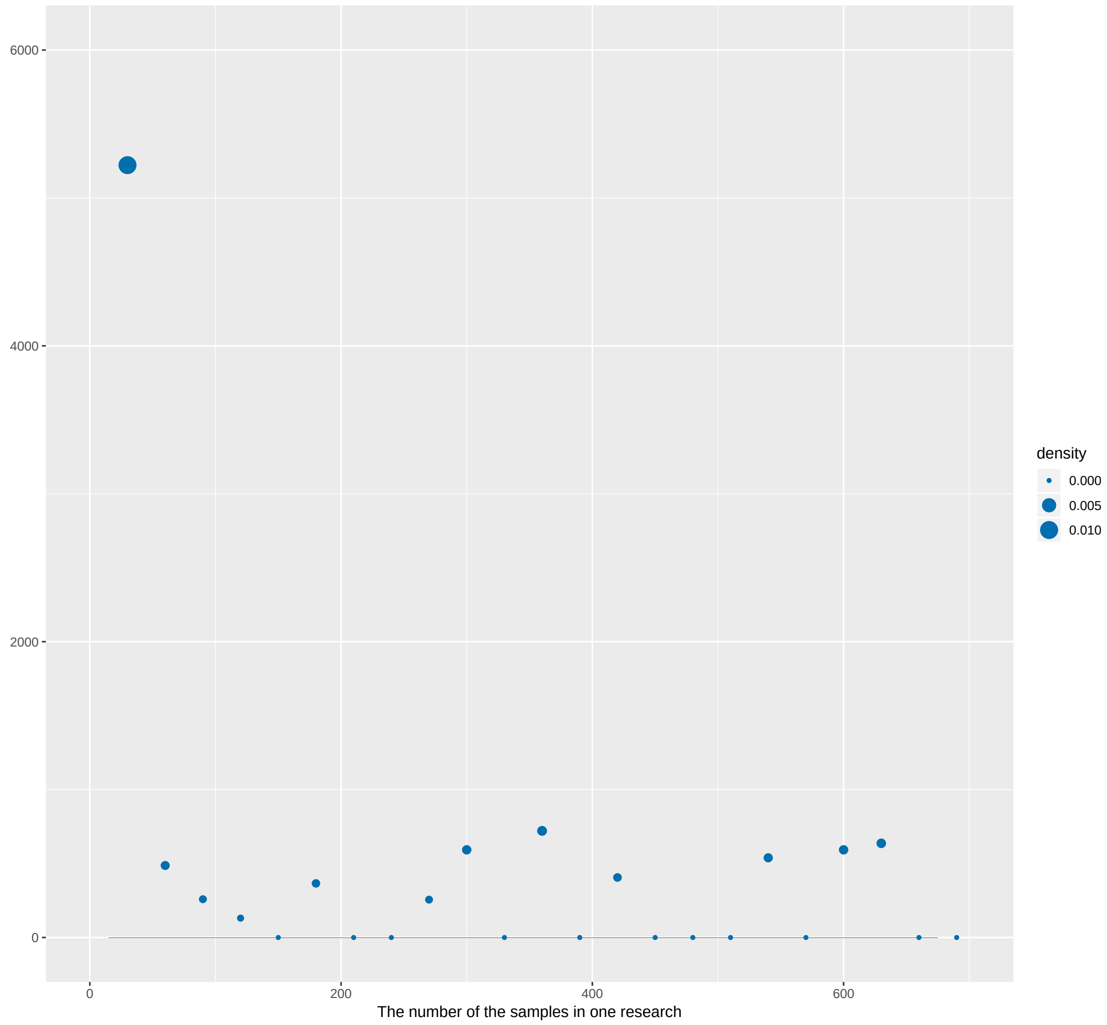

Figure 1: The size of samples within one research, most researches contains less than 50 samples.

3.2 Comparisons in performances of models fitting in habitats

Overall, the performances of all the models fitting in different habitats have significant difference, the p value < 1.534e-07. Cubic model has the best fitting performance over other three Schoolfield models. Within the three forms of Schoolfield models, the Schoolfield model for high temperature has the best performance.

In freshwater environments, under AIC, BIC, and AICC, the four models have different performances. All the pairwise p value < 0.01, only one p value between Schoolfield model and Schoolfield model under AICC is 0.01296. And Cubic model fits best, especially under AICC.

In marine environments, only under AICC, the Cubid model has significant difference with other three Schoolfield models. Under AIC and BIC, there isn't significant difference with the four models. The reason could be penalty of the number of parameters from AICC(Johnson & Omland 2004).

In terrestrial environments, under AICC, the Cubid model has significant difference with other three Schoolfield models. Under AIC, the Cubic model has significant difference with the Schoolfield model for high temperature and he Schoolfield model for low temperature, the p values are < 0.009082, and < 6.75e-08, respectively. And the Schoolfield model has significant difference with the Schoolfield model for high temperature and he Schoolfield model for low temperature, as well. The p values are < 0.03602, and < 1.13e-04, respectively. Under BIC, except p value between Cubic model and he Schoolfield model is > 0.05, other p value are signigicant(p values are not shown).

In a word, phenomenological model has better performance than mechanistic models. Under AICC, the weight value of three forms of the Schoolfield models are exceedingly low. It is likely that there are too many parameters(Johnson & Omland 2004). With the three forms, the Schoolfield model for high temperature and he Schoolfield model for low temperature have relatively good performance in marine environments.

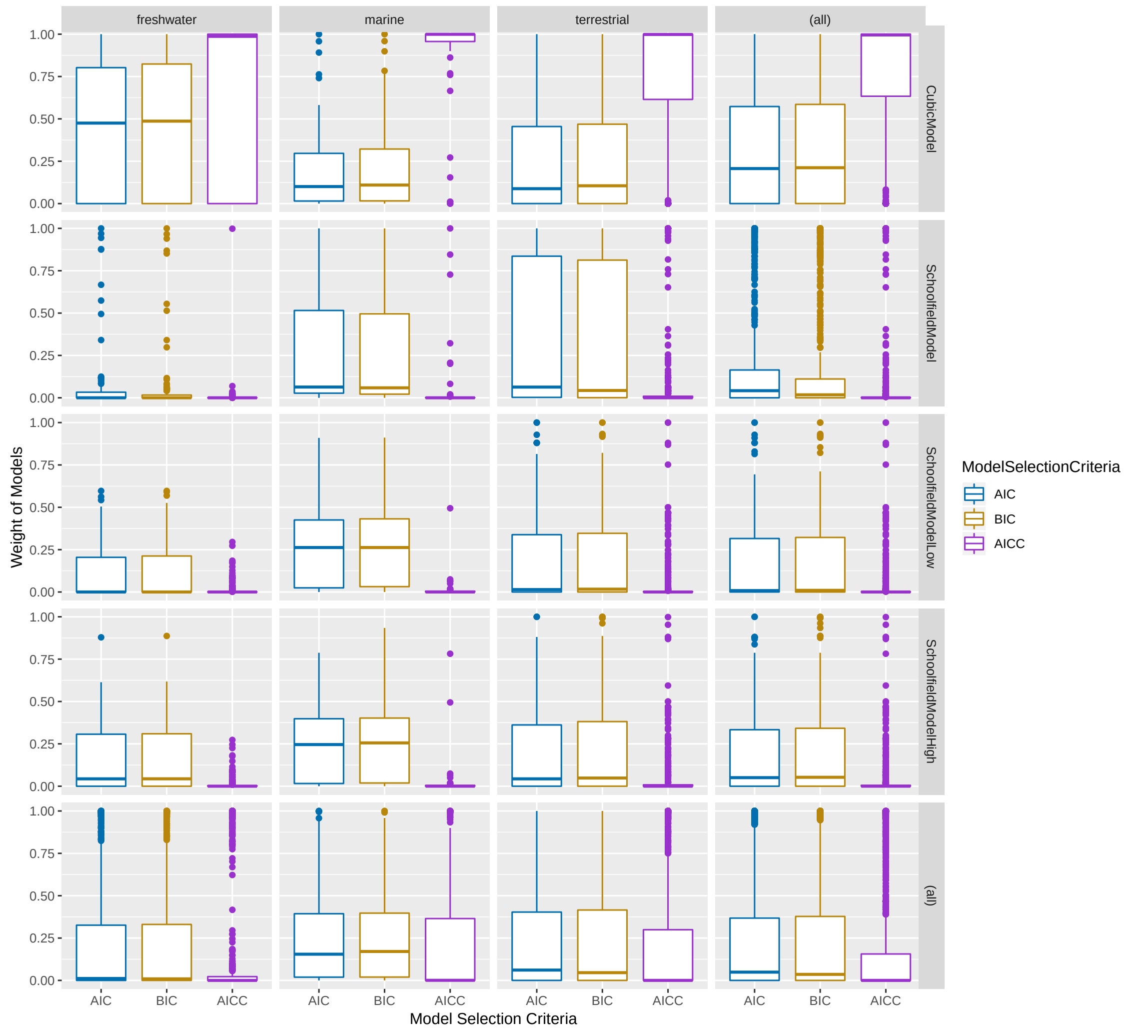

Figure 2: Comparisons in performances of models fitting in habitats. Cubic has the best performance. Under AICC, the weight value of three forms of Schoolfield model is extremely low.

3.3 Four models fitting one research

The FinalID of this research is MTD5347(Ratkowsky et al. 2005), the species is bacterium, *Escherichia.coli*, the original trait name is Square Root (Maximum Specific Growth Rate). The reason why this plot can only show three models is that the Schoolfield model for high temperature and the Schoolfield model for low temperature are quite similar and are overlapped.

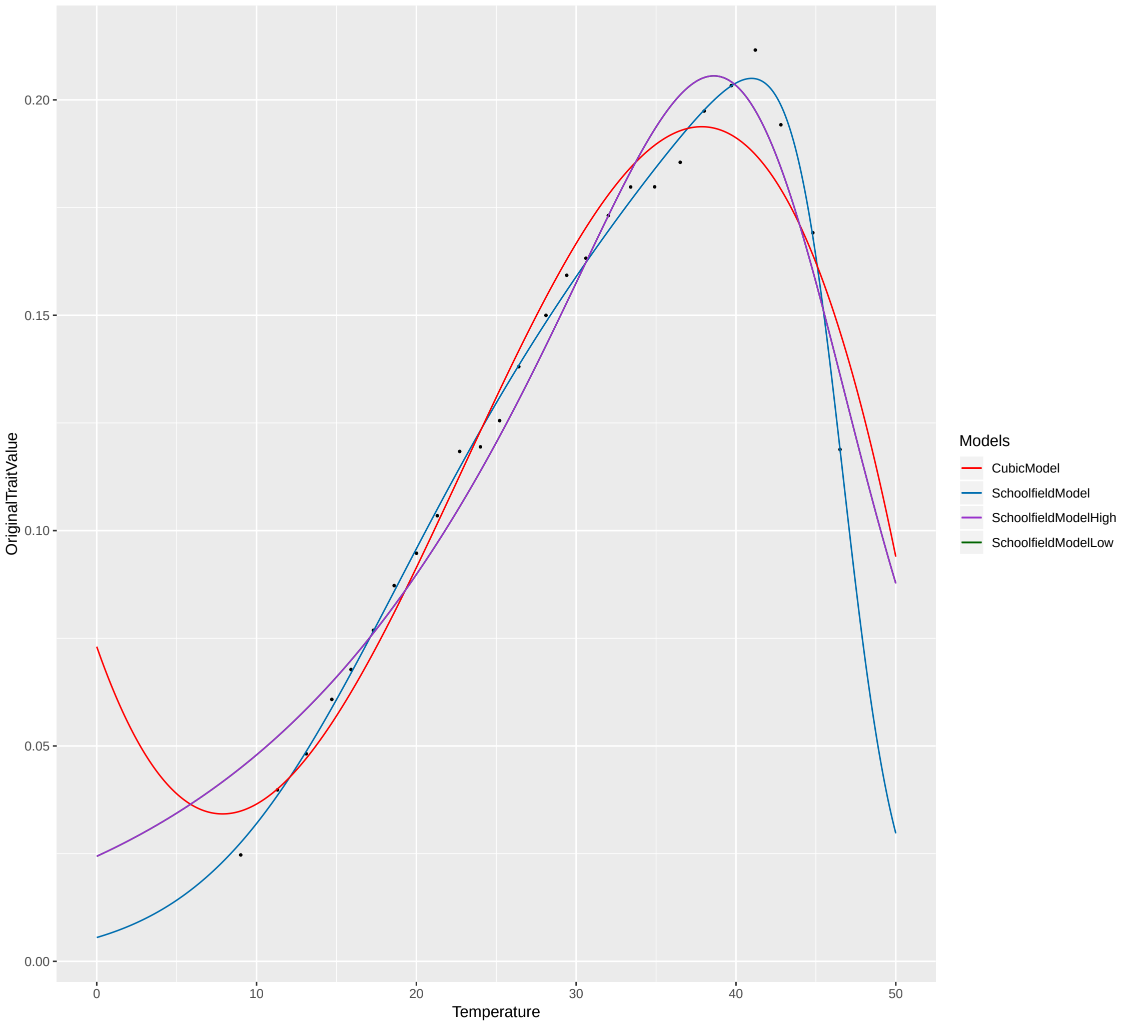

Figure 3: The four models fitted on one research. This curve shows the growth rate of *Escherichia.coli*.

4 Discussion

4.1 The reason behind the different performance

To star with, the distince temperature ranges in different habitats is considered. For instance, under water, most organisms do not experience temperatures above 25°C âĂŞ 30 °C, but in terrestrial environments, high tempratures of 36°C - 40 °C are common in summer.

However, later analysis shows the assumption is not supported. If the assumption is true, the Schoolfield model for low temperature should have much better performance than others. As shown above, that results are not drawn. Accordingly, researches on the reason are needed.

4.2 Limitations

4.2.1 Without further mining

The large dataset contains many interesting contents. But this report is without further mining. For instance, many Kingdom are included. Overall, the four models have significant difference at Kingdom level, the p value is 1.55e-10. But the pairwise significant difference were not calculated. The reason has not been considered, either.

Figure 4: Comparisons in performances of models fitting in different organisms at Kingdom level.

4.2.2 Without further understanding of the model selection criteria

The extremely low weight value of Schoolfield models is due to the penalty of the numbers of parameters from AICC. But the difference between AIC and BIC has not been discussed.

4.2.3 Visual flaw

Some of the figures are too complicated to add p values, and all the p values can't be shown at the figures.

5 Conclusion

In conclusion, phenomenological model has better performance than mechanistic models in different habitats. As for mechanistic models, in different habitats and under different model selection criteria, there is better model to choose. Researchers are supposed to choose the better models and model selection criteria carefully in terms of the specific background, habitat and so forth.

We need more researches to study to mechanisms behind the metabolic reactions and propose mechanistic models with precise prediction, which is especially essential at the time when extreme climate frequent and climate change has a critical influence on agriculture, fisheries, and health.

References

Allen, A. P., Gillooly, J. F., Savage, V. M. & Brown, J. H. (2006), 'Kinetic effects of temperature on rates of genetic divergence and speciation', *Proceedings of the National Academy of Sciences* **103**(24), 9130–9135.

Brander, K. M. (2007), 'Global fish production and climate change', *Proceedings of the National Academy of Sciences* **104**(50), 19709–19714.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. & West, G. B. (2004), 'Toward a metabolic theory of ecology', *Ecology* **85**(7), 1771–1789.

Burnham, K. P. & Anderson, D. R. (2004), 'Multimodel inference: understanding aic and bic in model selection', *Sociological methods & research* **33**(2), 261–304.

Callaway, J. (1959), 'Model for lattice thermal conductivity at low temperatures', *Physical Review* **113**(4), 1046.

Della, A. I., Pawara, S. & Savagea, V. M. (2013), 'The thermal dependence of biological traits', *landscape* **36**, 37.

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M. & Charnov, E. L. (2001), 'Effects of size and temperature on metabolic rate', *science* **293**(5538), 2248–2251.

Howden, S. M., Soussana, J.-F., Tubiello, F. N., Chhetri, N., Dunlop, M. & Meinke, H. (2007), 'Adapting agriculture to climate change', *Proceedings of the national academy of sciences* **104**(50), 19691–19696.

Johnson, J. B. & Omland, K. S. (2004), 'Model selection in ecology and evolution', *Trends in ecology & evolution* **19**(2), 101–108.

McKinney, W. (2011), 'pandas: a foundational python library for data analysis and statistics', *Python for High Performance and Scientific Computing* **14**.

Millman, K. J. & Aivazis, M. (2011), 'Python for scientists and engineers', *Computing in Science & Engineering* **13**(2), 9–12.

Newville, M., Stensitzki, T., Allen, D. B., Rawlik, M., Ingargiola, A. & Nelson, A. (2016), 'Lmfit: non-linear least-square minimization and curve-fitting for python', *Astrophysics Source Code Library* .

Prowse, T. D., Furgal, C., Wrona, F. J. & Reist, J. D. (2009), 'Implications of climate change for northern canada: freshwater, marine, and terrestrial ecosystems', *AMBIO: A Journal of the Human Environment* **38**(5), 282–290.

Ratkowsky, D. A., Olley, J. & Ross, T. (2005), 'Unifying temperature effects on the growth rate of bacteria and the stability of globular proteins', *Journal of theoretical biology* **233**(3), 351–362.

Ruxton, G. D. (2006), 'The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test', *Behavioral Ecology* **17**(4), 688–690.

Savage, V. M., Gillooly, J. F., Brown, J. H., West, G. B. & Charnov, E. L. (2004), 'Effects of body size and temperature on population growth', *The American Naturalist* **163**(3), 429–441.

Schoolfield, R., Sharpe, P. & Magnuson, C. (1981), 'Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory', *Journal of theoretical biology* **88**(4), 719–731.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Van Der Walt, S., Colbert, S. C. & Varoquaux, G. (2011), 'The numpy array: a structure for efficient numerical computation', *Computing in Science & Engineering* **13**(2), 22.

Vargha, A. & Delaney, H. D. (1998), 'The kruskal-wallis test and stochastic homogeneity', *Journal of Educational and Behavioral Statistics* **23**(2), 170–192.

Wickham, H. (2009), 'plyr: Tools for splitting, applying and combining data', *R package version 0.1* **9**, 651.

Wickham, H. (2012), 'reshape2: Flexibly reshape data: a reboot of the reshape package', *R package version* **1**(2).

Wickham, H. (2016), *ggplot2: elegant graphics for data analysis*, Springer.

Wickham, H., Francois, R., Henry, L., Müller, K. et al. (2015), ‘dplyr: A grammar of data manipulation’, *R package version 0.4* **3**.

Woodward, A., Smith, K. R., Campbell-Lendrum, D., Chadee, D. D., Honda, Y., Liu, Q., Olwoch, J., Revich, B., Sauerborn, R., Chafe, Z. et al. (2014), ‘Climate change and health: on the latest ipcc report’, *The Lancet* **383**(9924), 1185–1189.