Forecasting the Carbon Pulse.

John Reid*[1] and Joachim Dengler[2]

[1]33 Charlotte Street,New Norfolk,Tasmania 7140,Australia,
johnsinclairreid@gmail.com
[2]Reichensteinstr. 54, 69151 Neckargemünd, Germany,
joachim.dengler@googlemail.com

December 28, 2020

Introduction

In the early days of economic modelling it was found that complicated models performed poorly when used for out-of-sample forecasting. Regression models with few independent variables often produced better forecasts than did the big models[1][2]. Climate models are no less complicated than economic models but, until now, this aspect of them has been ignored. The fundamental question such models are designed to answer is the extent to which increases in atmospheric greenhouse gas concentrations affect the Earth's climate.

Here we investigate the autoregression of global average temperature on past temperature and its regression on anthropogenic emissions. All coefficients are estimated simultaneously to avoid omitted-variable bias[3] and the complexities of de-convoluting a moving average component are avoided by decimating the data.

We vet regression models by testing residuals for self-correlation. The successful models are then used to forecast global average temperature from Hubbert curves[4] of emission estimates. We find that forecast temperature maxima are only weakly dependent on remaining fossil fuel reserves.

Some Pitfalls of Regression

Figure 1 shows the the time series of annual global average temperature anomaly, T, and the natural logarithm of atmospheric carbon dioxide concentration, Γ, plotted against time, where

$$\Gamma = \ln([\mathrm{CO}_2]) \tag{1}$$

where $[\mathrm{CO}_2]$ is the atmospheric carbon dioxide concentration in parts per million.

Clearly there is a strong relationship between them. It is commonly assumed that this is a regression relationship of the form:

$$Z_i - \bar{z_i} = a(x_i - \bar{x_i}) + \Xi_i \tag{2}$$

where Z_i is a random variable whose realization is the temperature, T_i, $x_i = \Gamma_i$, a is the regression coefficient $\{\Xi_i\}$ is a sequence of unselfcorrelated random variables with zero mean and $\bar{z_i}$ and $\bar{x_i}$ are sample means. Ordinary least squares gives the regression coefficient estimate, $\hat{a}$, as 2.778. An estimate of climate sensitivity (the increase in T caused by a doubling of $[CO_2]$), is then $\hat{S} = \hat{a}.\ln(2) = 1.93$ in agreement with the IPCC high confidence range.

This argument is flawed. Statistical tests show that the sequence of residuals is self-correlated. There is a strong serial correlation of the residuals because the global average temperature in any given year is well correlated with the global average temperature from the preceding year even when the dependence on Γ

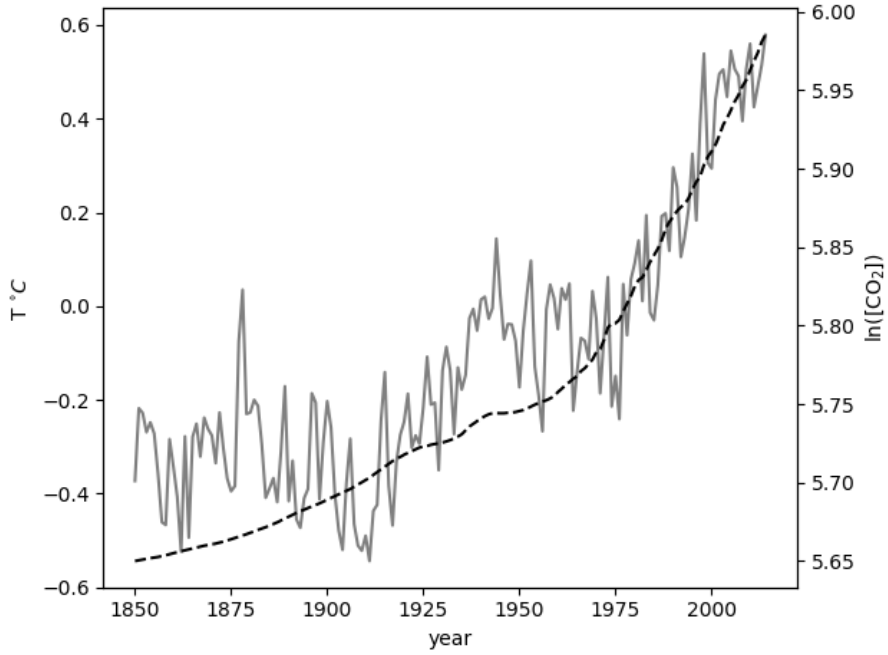

Figure 1: Annual global average temperature anomaly, T, (grey curve) and the natural logarithm of atmospheric carbon dioxide concentration, $\ln([CO_2])$, (black, dashed curve) plotted against time.

is allowed for. This is not taken into account by (2) so giving rise to omitted-variable bias[3]. Least squares regression returns biased estimates when all the exogenous variables are not included. Estimates of the regression coefficient, $\hat{a}$, and climate sensitivity, $\hat{S}$, derived from (2) are biased statistics.

The ARX Regression Model

In what follows, with no loss of generality, all variables will be assumed to have zero mean.

Environmental temperatures are laregly dependent on the transport of heat and other forms of energy by turbulent processes[5]. The relationship between the heat flux to and from a reservoir and its temperature is given by Fourier's heat equation in the form:

$$\rho_0 c_p \frac{\partial T}{\partial t} = -K_0 \frac{\partial T}{\partial x} + Q(x,t) \tag{3}$$

where ρ_0 and c_p are the density and specific heat of the reservoir, T is its temperature, t is the time, x is a spatial variable with the units of length, K_0

is the conductivity and Q is the flux of heat or other form of energy.

In the present case the reservoir is the mixed layer of the ocean. Because the specific heat of water is much larger than that of air, the top 2.5 m of the mixed layer holds as much heat as the entire atmosphere above it. The globally averaged mixed layer depth[6] is about 100m, so that the heat capacity of the mixed layer is 40 times that of the atmosphere. Seventy percent of the Earth's surface is ocean. Effectively, mixed layer temperatures comprise the bulk of globally averaged surface air temperature measurements and the two are highly correlated.

Integrating (3) over the entire mixed layer and including a radiative forcing term, $G(t)$, for the heat trapping effect of greenhouse gases gives Newton's Law of Cooling for the mixed layer:

$$\rho c_p \frac{\partial T}{\partial t} = -KT + G(t) + q(t) \tag{4}$$

where ρ, c_p are the density and specific heat of sea water and K is a global average of all forms of cooling which are approximately proportional to temperature differences. In finite difference form (4) becomes

$$T_i = a.T_{i-1} + b.\Gamma_i + \xi_i \ \ , \ \ i = 1, ..., N \tag{5}$$

where $t = i\Delta t$, Γ_i is given by (1)[7], , $\xi_i = \Delta t q_i/(\rho c_p + K\Delta t)$ and a and b are constants to be estimated from the N data values. This can be further generalized to

$$Y_i = \alpha_0 x_i + \sum_{j=1}^{p} \alpha_j.y_{i-j} + \sum_{k=1}^{q} \beta_j \Xi_{i-k} \ \ , \ \ i = 1, ..., N \tag{6}$$

where T is represented by both Y and y, $x_i = \Gamma_i$ is the exogenous variable and the Ξ_i are unselfcorrelated random variables with zero mean. The regression coefficients α_0, α_j and β_j are to be estimated from the data and p and q are small positive integers. Here upper case symbols represent random variables and lower case symbols are constants.

This is the ARMAX(p,q) model (for 'autoregressive moving average with exogenous variable'). There are software packages for parameter estimation available under the aegis of the major programming languages. Unfortunately some of these, such as the Python *Statsmodels* package, are flawed because they estimate the exogenous parameter, α_0, prior to estimating the other parameters, leading to omitted-variable bias mentioned above.

Note the distinction between the random variable Y_i and the sample values y_{i-j} which are constants. Equation (6) is a state space representation[8] describing states of the system at a succession of discrete instants; the random variable, Y_i, at one instant becomes the constant, y_i, in the following instant. The direction of time is important in regression, which, unlike correlation, allows causality to be inferred.

Estimation of the MA coefficients, $\{\beta_i\}$, requires an iterative Kalman filter method which does not always converge. The second, moving average summation in (6) is a blurring function, which is non-zero when the sampling interval, Δt, is too small. The given time series can be downsampled or 'decimated' by q to give a new time series with sampling interval $q\Delta t$ with little loss of information and (6) becomes

$$Y_m = \alpha'_0 x_m + \sum_{n=1}^{p} \alpha'_n . y_{m-n} + \Xi'_m \quad , \quad m = 1, ..., M \tag{7}$$

where $m = qi$, $qM \leq N$ and

$$\Xi'_m = \sum_{k=1}^{q} \beta_j \Xi_{m-k} \tag{8}$$

so that $\mathrm{E}(\Xi'_m \Xi'_n) = 0$ for $m \neq n$ because the summations do not overlap. The $\{\Xi'_m\}$ in (7) are therefore unselfcorrelated with zero mean. This is a simple case of a Renormalization Group Transformation[9].

The model summarized by (7) can be termed an ARX(p) model for 'autoregressive with exogenous variable'. Its parameters and their confidence limits can be estimated using ordinary least squares. The sequence of residuals, $\{\xi'_m\}$, is given by

$$\xi'_m = y_m - \left(\hat{\alpha}'_0 x_m + \sum_{n=1}^{p} \hat{\alpha}'_n . y_{m-n} \right) \quad , \quad m = 1, ..., M \tag{9}$$

where y_m is the sample value or 'realization' of Y_m, $\hat{\alpha}'_0$ and $\hat{\alpha}'_n$ are the coefficient estimates and the $\{\xi'_m\}$ are to be tested for self-correlation.

The diffusion of carbon dioxide from the atmosphere into the deep ocean reservoir is governed by a similar equation to (4), viz.:

$$\frac{\partial C}{\partial t} = -DC + E(t) + r(t) \tag{10}$$

where C is the atmospheric concentration of CO_2, D is a diffusion coefficient, $E(t)$ is the production rate and $r(t)$ is a random component analogous to $q(t)$ in (4). This too can be generalized into an ARX model of the form (7) with $Y = C$ and $x = E$.

Data Sources

Time series in the form of annual averages of the relevant variables, E, C and T were downloaded from the Web in November 2020.

The global average temperature anomaly data, T. were taken from the `HadCRUT.4.5.0.0` data set[10].

Model	Exogenous Variables	Q	P
ARX(0)	E_i only	1424.6	0.0000
ARX(1)	E_i,C_{i-1}	95.4	0.0000
ARX(2)	E_i,C_{i-1},C_{i-2}	120.0	0.0000
ARX(3)	E_i,C_{i-1}, ... ,C_{i-3}	91.3	0.0000
ARX(4)	E_i,C_{i-1}, ... ,C_{i-4}	71.0	0.0000
ARX(5)	E_i,C_{i-1}, ... ,C_{i-5}	71.4	0.0000

Table 1: Ljung-Box parameter, Q, and its probability, P, for ARX models of undecimated CO_2 concentration, C_i, and CO_2 emissions data, E_i.

Carbon dioxide concentrations, C, were taken from the University of Melbourne Greenhouse Gas Factsheet[11] supplemented with recent values from Mauna Loa.

Global fossil fuel emissions, E, were downloaded from CDIAC, the Carbon Dioxide Information Analysis Center[12].

Finding Models

The *sine qua non* of all regression models is that the innovation, Ξ, be unselfcorrelated. It is an assumption which can be validated by testing whether the sample residuals, $\{\xi'_m\}$, given by (9), are self-correlated. If they are, then the random variables, $\{\Xi'_m\}$, in (7) cannot be assumed to be unselfcorrelated and (7) is not a valid regression model.

Under the Ljung-Box test[13], a test parameters, Q, is computed from the sample autocorrelation of $\{\xi'_m\}$ at lag k out to some maximum lag, k_{max}. Under the null hypothesis that the residuals are unselfcorrelated, Q has a χ^2 distribution with $k-n$ degrees of freedom where n is the number of regression coefficients fitted. From this a probability, P, can be found and suitable values of q, the decimation factor, and p, the number of autoregressive coefficients, decided by trial and error as the smallest values which satisfy this test.

We wish to use an ARX model (7) to find firstly, (i) the regression of atmospheric carbon concentration, C, on anthropogenic carbon emissions, E, and then, (ii) the regression of global average temperature anomaly, T, on the logarithm of atmospheric carbon concentration, Γ. In step (i), E is the exogenous variable. In step (ii) Γ is the exogenous variable.

Step (i): applying Ljung-Box to the residuals given by (9) for ARX(p), p= 0, ... ,5) gives the results shown in Table 1. Probabilities are zero for all values of p indicating that the null hypothesis that the residuals are unselfcorrelated can be rejected. Both time series were then decimated by 2. The results are shown in Table 2. The probability, P, for the ARX(1) model has a value of 0.4259 indicating that the null hypothesis that the residuals are unselfcorrelated cannot be rejected. The ARX(1) case, is a good fit to the data and constitutes a valid regression model for C on E.

Step (ii): an ARX model of the regression of global average temperature, T,

Model	Exogenous Variables	Q	P
ARX(0)	E_i only	513.5	0.0000
ARX(1)	E_i,C_{i-1}	28.5	0.4359
ARX(2)	E_i,C_{i-1},C_{i-2}	28.6	0.3830
ARX(3)	E_i,C_{i-1}, ... ,C_{i-3}	24.5	0.5483
ARX(4)	E_i,C_{i-1}, ... ,C_{i-4}	24.3	0.5049
ARX(5)	E_i,C_{i-1}, ... ,C_{i-5}	22.0	0.5796

Table 2: Ljung-Box parameter, Q, and its probability, P, for ARX models of CO_2 concentration, C_i and CO_2 emmissions, E_i, when both time series have been decimated by 2

Model	Exogenous Variables	Q	P
ARX(0)	T_i vs ln(C_i) only	158.4	0.000
ARX(1)	T_i vs ln(C_i),T_{i-1}	60.9	0.0003
ARX(2)	T_i vs ln(C_i),T_{i-1},T_{i-2}	32.9	0.1995
ARX(3)	T_i vs ln(C_i),T_{i-1}, ... ,T_{i-3}	31.2	0.2214
ARX(4)	T_i)vs ln(C_i),T_{i-1}, ... ,T_{i-4}	32.5	0.1436
ARX(5)	T_i vs ln(C_i),T_{i-1}, ... ,T_{i-5}	34.2	0.0807

Table 3: Ljung-Box parameter, Q, and its probability, P, for ARX models of global average temperature and CO_2 concentration data decimated by 2

on the logarithm of atmospheric CO_2 concentration is also required. These time series were also decimated by 2 to keep the same sampling interval throughout. The results of the Ljung-Box test for these ARX models are shown in Table 3. The ARX(2) model was chosen as the simplest model with unselfcorrelated residuals. Estimates of the regression coefficients of (7), $\hat{\alpha}'_0$, $\hat{\alpha}'_1$ and $\hat{\alpha}'_1$ for the two selected models are shown in Table 4

Forecasting Global Average Temperature

Under the assumption of covariance stationarity, regression coefficients estimated from the data, such as those of Table 4, can be substituted back into (7) to yield expectation values of Y outside the domain of the original data over the domain in which the exogenous variable, x, is known or can be estimated.

Model	$\hat{\alpha}'_0$	$\hat{\alpha}'_1$	$\hat{\alpha}'_2$
C_i vs E_i, C_{i-1}	0.210	0.969	
T_i vs ln(C_i),T_{i-1},T_{i-2}	1.549	0.213	0.299

Table 4: Regression coefficients estimated from the selected models

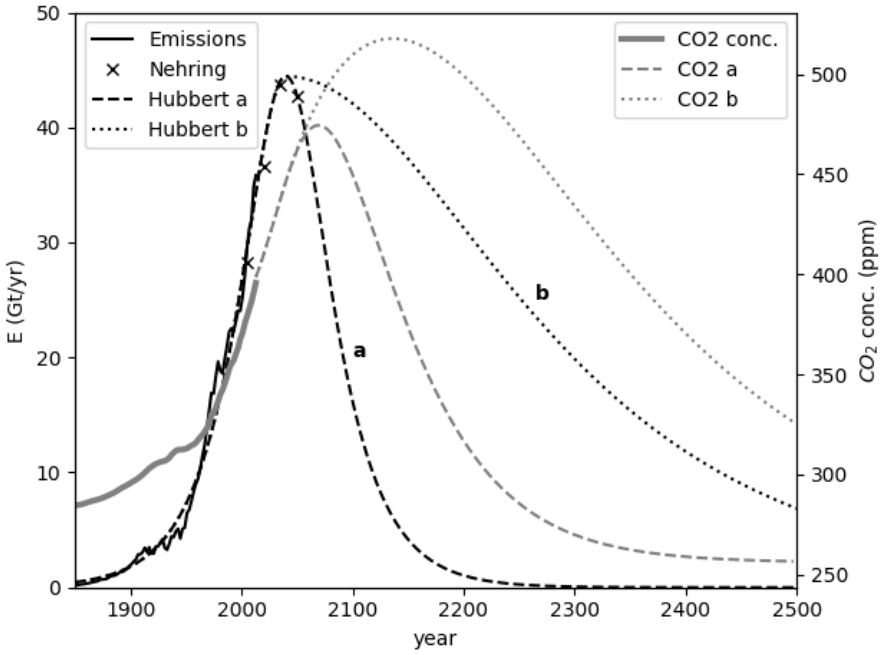

Figure 2: The solid black curve shows the recorded total carbon dioxide emission rate (Gt/yr). The crosses show recorded and predicted emission rates quoted by Nehring[14]. Curve **a** is the symmetrical Hubbert curve which best fits Nehring's data. Curve **b** is an asymmetrical Hubbert curve with decay time constant set to five times the onset constant. The grey dashed and dotted curves show the predicted CO_2 concentrations forecast from the two Hubbert emission curves.

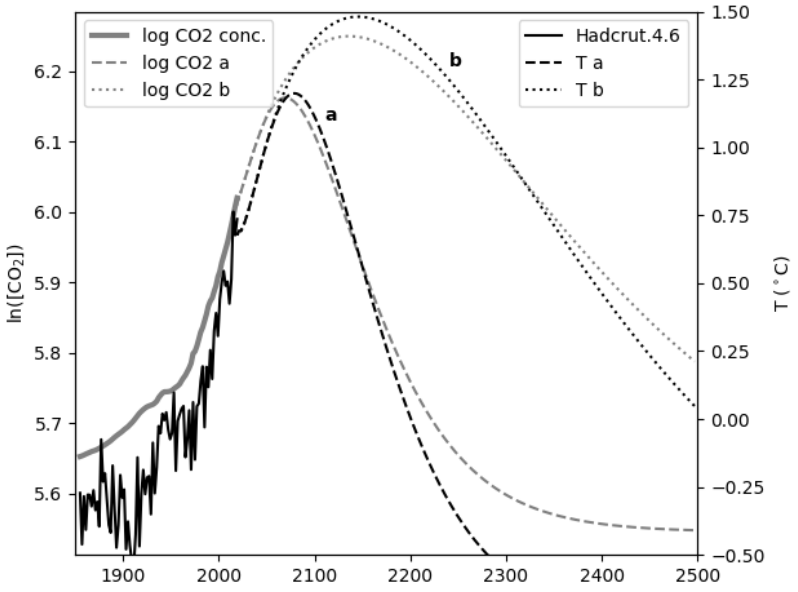

Figure 3: The grey curves show the logarithm of carbon concentrations resulting from the emission curves shown in Figure 2. The solid black curve is the observed (HadCRUT.4.6.0.0) global average temperature anomaly. The dashed black curves show the predicted global average temperature anomalies corresponding to the Hubbert curves **a** and **b** in Figure 2.

Curve	Year	T	Total
Hubbert (a)	2079	1.20°C	1824 Gt
Hubbert (b)	2145	1.48°C	6223 Gt
Pre-industrial	1850	-0.37°C	
Present Day	2019	0.74°C	

Table 5: Year and value, T, of peak temperature anomaly generated by the (a) symmetrical and (b) asymmetrical Hubbert curves shown in Figure 2. The the pre-industrial and present day values of T are shown for reference. Total emissions for the period 1850 to 2500 for the two Hubbert curves are also shown.

This can be done by iteration. Thus, from (7)

$$\mathrm{E}(Y_{M+1}) = \alpha'_0 x_M + \sum_{n=1}^{p} \alpha'_n . y_{M+1-n} \tag{11}$$

since $\mathrm{E}(\Xi'_m) = 0$ by definition. If y_{M+1} is now defined as $\mathrm{E}(Y_{M+1})$ for substitution into the right hand side of (11), the process can be repeated and the future behaviour of the endogenous variable, y, forecast over the domain for which the exogenous variable, x, is known.

Hubbert Curves

In order to forecast global average temperature we need to know the endogenous variable, Γ, a function of atmospheric carbon dioxide concentration, C, given by (1). In order to forecast C we need to know the carbon dioxide emission rate, E.

Although controversial, the Hubbert curve[4] provides a canonical starting point. The Hubbert curve for projected carbon emissions, fitted to known emissions and to the data of Nehring[14] is shown as the dashed black curve, **a**, of Figure 2. The atmospheric CO_2 concentration forecast using the iterative method is shown as the grey dashed line.

A major criticism of the Hubbert Curve is its symmetry. In reality, the decline in resources is usually significantly slower than the onset of the curve. An asymmetrical Hubbert curve was generated with the same onset and maximum as curve **a** but with a decay time five times slower. This is shown as the dotted black curve, **b**, in Figure 2. The corresponding, atmospheric CO_2 forecast is shown as the grey dotted line.

The emissions-generated, grey CO_2 concentration curves in Figure 2 were used via (1) to generate the (grey) log concentration curves **a** and **b** shown in Figure 3. These are the exogenous functions used to generate the corresponding temperature forecasts, the black dashed and dotted curves, in that figure. The coordinates of the peaks of the two curves are shown Table 5 along with pre-industrial and present day values. Also shown are the total emissions for the two

Hubbert curves for the 650 years displayed in the figures. The total emissions for the Hubbert (b) curve are more than three times those for the Hubbert (a) curve but the resulting temperature maxima differ by only 0.28°C.

Evidently peak global average temperature is only a weakly dependent on total carbon emissions. This is so because of the non-linearity of (1) and the negative sign of the first term on the right hand side of both (3) and (10). In (3) this term describes the rate at which heat is lost to space from the earth's surface and in (10) it describes the rate at which CO_2 diffuses from the atmosphere into the deep ocean.

A common definition of climate sensitivity is the temperature increase caused by a permanent doubling of CO_2 concentration. In reality, because heat loss and diffusion terms are non-zero, this would require a large emission rate to be maintained indefinitely, an unrealistic scenario.

Conclusions

Viewed on a time scale of centuries, human exploitation of fossil fuels in the industrial era is generating a pulse in atmospheric carbon concentration which, in turn, generates a pulse in global average temperature. So far we have observed only the onset of this pulse. Reasonable estimates of recoverable fossil fuel reserves suggest that the carbon pulse will reach a maximum within the next century or so. Global average temperature will follow suit with a peak value of about 2°C above pre-industrial values.

References

[1] Charles R Nelson. "The Prediction Performance of the F.R.B.-M.I.T.-PENN Model of the U.S. Economy". In: *American Economic Review* 62 (1972), pp. 902–917.

[2] R. Ashley. "On the Relative Worth of Recent Macroeconomic Forecasts". In: *International Journal of Forecasting* 4 (1988), pp. 363–376.

[3] W. H. Greene. *Econometric Analysis (5th ed.)* New Jersey: Prentice Hall, 2003, pp. 245–246. ISBN: 0-13-066189-9.

[4] M. K. Hubbert. *Energy Resources*. National Academy of Sciences, Publication 1000-D, 1962, pp. 60–83.

[5] K. Hasselmann. "Stochastic climate models, Part I, Theory". In: *Tellus* XXVIII.6 (1976), pp. 473–485.

[6] C. de Boyer Montégut et al. "Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology". In: *J. Geophys. Res.* 109.C12003 (July 2004). DOI: 10.1029/2004JC002378.

[7] Y. Huang and M. Bani Shahabadi. "Why logarithmic? A note on thedependence of radiative forcing on gas concentration". In: *J. Geophys. Res. Atmos.* 119 (2014), pp. 13, 683–13, 689. DOI: 10.1002/2014JD022466.

[8] E. J. Hamilton. *Time Series Analysis*. Princeton, New Jersey: Princeton University Press, 1994. ISBN: 978-0-691-04289-3.

[9] Q-G. Wang et al. "Improved System Identification with Renormalization Group". In: *International Conference on Control and Automation*. Hangzhou, China: ICCA, 2013, pp. 878–883.

[10] C P Morice et al. "Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset". In: *Journal of Geophysical Research* 117 (2012), p. D08101. DOI: 10.1029/2011JD017187. URL: http://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts+dSST.txt.

[11] M. Mainshausen et al. "Historical greenhouse gas concentrations for climate modelling (CMIP6)". In: *Geosci. Model Dev.* 10 (2017), pp. 2057–2116.

[12] T. A. Boden et al. "Global, Regional, and National Fossil-Fuel CO2 Emissions." In: *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A* (2017). DOI: 10.3334/CDIAC/00001_V201.

[13] G M Ljung and G E P Box. "On a Measure of a Lack of Fit in Time Series Models". In: *Biometrika* 65 (1978), pp. 297–303. DOI: 10.1093/biomet/65.2.297.

[14] Richard Nehring. "Traversing the mountaintop: world fossil fuel production to 2050". In: *Philos Trans R Soc Lond B Biol Sci* 364 (2009), pp. 3067–3079.

Author Information

Authors contributed equally.