HAL
archives-ouvertes.fr

Some Regression Problems in Solar-Terrestrial Sciences: Learning from Mistakes

Thierry Dudok de Wit

To cite this version:

Thierry Dudok de Wit. Some Regression Problems in Solar-Terrestrial Sciences: Learning from Mistakes. EAS Publications Series, EDP Sciences, 2014, 66, pp.77-87. 10.1051/eas/1466007 . insu-02986473

HAL Id: insu-02986473

https://hal-insu.archives-ouvertes.fr/insu-02986473

Submitted on 3 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SOME REGRESSION PROBLEMS IN SOLAR-TERRESTRIAL SCIENCES: LEARNING FROM MISTAKES

T. Dudok de Wit[1]

Abstract. We address three timely regression analysis problems in solar-terrestrial observations: the identification of trends in observations that exhibit a high level of internal variability, the choice of explanatory variables in the multilinear regression of climate data, and the identifi-cation of power laws in power spectral densities. In all three of them we focus on some common mistakes, and on how these may help facilitate critical reading of research in the field.

1 What not to do in regression analysis

In his seminal book on numerical methods, Acton, 1970 has a short interlude on what not to compute, which has inspired many scientists. As a small tribute to that unique interlude, we consider here some regression analysis problems as they are encountered in the context of solar-terrestrial physics, and highlight some common mistakes. These mistakes are of course generic, and are only a tiny subset of a large rensemble that includes issues such as the lack of awareness on the assumptions behind the regression methods, the prediction outside of a relevant range, the propagation of uncertainties of input variables to the regression model prediction, which may be even more uncertain, and neglecting the bias introduced by choosing an inadequate model (King 1986; Rong 2000; Berk 2004; Good & Hardin 2012).

Here, we concentrate on three timely issues that have been hotly debated in recent years. The first one is about identifying a drift in ionospheric observations, which is important for assessing the existence of long-term changes that may be related to global climate change. Technically, the problem is about the choice of the explanatory variables. The second issue is about the identification of a solar signature in climate records. Here the problem deals with the impact of collinearity on the regression analysis. The third and last example is about the identification of power laws in power spectral densities, with in addition the location of cutoff

[1] LPC2E, UMR 7328 CNRS-University of Orl´eans, 3A avenue de la Recherche Scientifique, 45071 Orl´eans Cedex 2, France; e-mail: ddwit@cnrs-orleans.fr

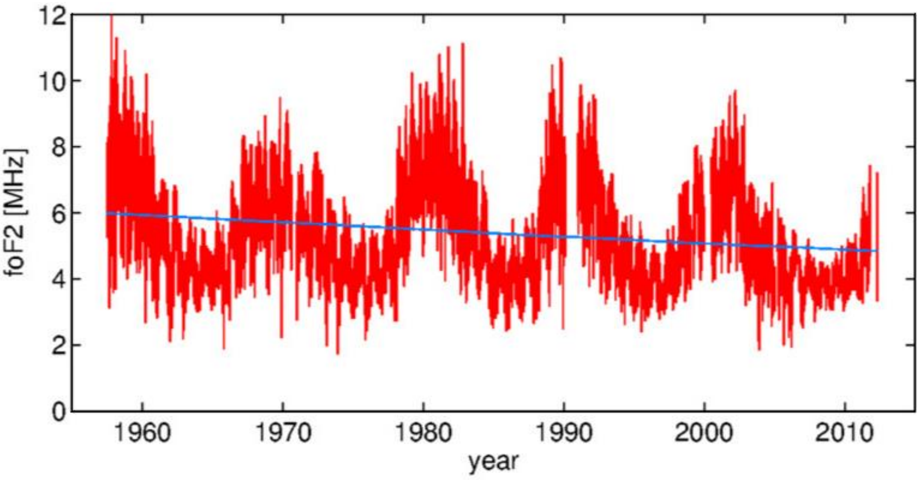

Fig. 1. Raw critical frequency foF2, measured daily at noon at the Juliusruh station (54.6N, 13.4E). The straight line has been computed without any prior reduction of data to solar or geomagnetic activity.

frequencies. This is again a classical regression problem, with too many cases of erroneous conclusions. These simple illustrations are meant to facilitate critical reading of research in the field, and to help avoid being led astray by mistakes we all make.

2 Trend determination: Is the sky falling down?

The increasing concentration of greenhouse gases in the atmosphere causes enhanced cooling of the upper atmosphere, which should result in changes of atmospheric parameters. One of these is a trend in the ionosphere. This at-mospheric layer, however, is highly variable and driven by solar and geomagnetic activity, whose levels also changed throughout the 20th century. If a trend exists at all, then it is most likely hidden in the natural variability. Disentangling these various signatures is an interesting but challenging regression problem (Laˇstoviˇcka et al. 2011). One the most representative ionospheric parameters is the critical frequency, called foF2, of the F2 layer. This layer has the densest electron concen-tration in the ionosphere, and extends from about 200 km to 500 km.

There has been a long quest for trends in foF2 (Laˇstoviˇcka et al. 2006). Ideas have progressively evolved as the pitfalls in the analyses have came to light. Figure 1 shows one of the longest records available, which is from the Julius-ruh station in Northern Germany. The trend in foF2 is heavily dominated by an 11-year solar cycle modulation, by an annual and semi-annual modulation that is related to the inclination of the Earth, and by more impulsive events that are due to geomagnetic activity. The earliest studies started by fitting a simple line to the

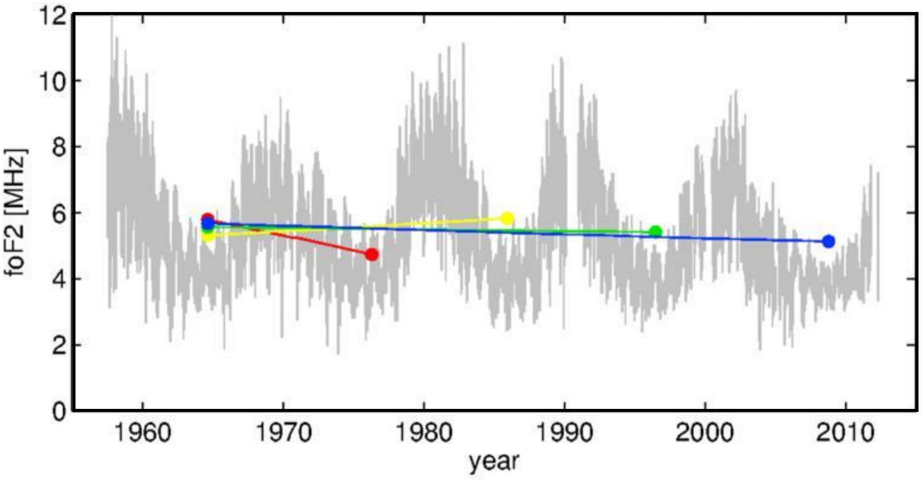

Fig. 2. Same plot as Figure 1, showing in addition four trends (dashed lines) that were estimated by considering intervals running between different solar minima.

time series, with the model

$$foF2(t) = a + bt + (t)$$

and then by interpreting the value of b; here, (t) stands for the residuals. However, it rapidly became evident that the slope b would heavily depend on the phasing with the solar cycle modulation. Both positive and negative trends were reported, depending on the endpoints of the time interval.

A first improvement consisted in applying trend analyses only to full periods of the solar cycle, as illustrated in Figure 2. However, it then became evident that the results would be biased by the non stationary properties of the solar cycle. Later models thus included the sunspot number $S(t)$ as a proxy for solar activity

$$foF2(t) = a + b\,t + c\,S(t) + (t)$$

with variants. For example, should the slope b be considered as the trend, or should we fit instead the simpler model

$$foF2(t) = a + c\,S(t) + (t)$$

and search for a trend in the residuals (t)? It took some time to realise that the last two models lead to different conclusions because missing influential variables bias the results. Successive improvements followed unabated, and several hun-dreds of models were tested, including more advanced ones, involving for example multiscale decompositions.

Today, there is a global consensus for the trend in foF2 at Juliusruh to be relatively weak and negative, of the order of −0.02 to −0.015 MHz/year. However, these results are not conclusive yet, neither from a statistical, nor from a physical point of view. Some of the statistically important questions that remain to be addressed in order to move from a mere description of observations to actual statistical inference (with hypothesis tests), are:

- Measurement errors, and their propagation to the model parameters, have been largely ignored so far. The errors (and the natural variability) in foF2 are heteroscedastic, reaching a maximum at the time when solar activity also peaks.

- There is no single good proxy for solar activity, but rather an ensemble of them. The most frequently used proxies are the sunspot number and the solar radio flux at 10.7 cm, which have quite different statistical properties. One emanates from a counting process, with Poisson-like noise, whereas the other has normal noise. Not surprisingly, the two lead to different results.

- The most critical issue probably is the proper choice of the influential vari-ables. Missing influential variables are known to inflate type II errors, which are the failure to reject a false null hypothesis.

The first two questions suggest that Simpson's paradox (Pearl 2009) may also play a role here. In this paradox, a trend may change quite substantially depending on how the data are aggregated. Since both the statistical and the physical properties of the observations change with their phasing with respect to the solar cycle, the search for a better model, including a rethinking of the way the data are sampled, is definitely needed.

3 Multilinear regression: Desperately searching for solar signatures

One of the key issues in climate research is to scientifically ascertain the mecha-nisms that are responsible for recent climate change. The Sun is in the spotlight because the level of its contribution to global warming has far reaching political implications. Although there is a clear consensus today on the prevalent effect of anthropogenic greenhouse gases in global warming, the quantification of the solar contribution remains a difficult and ongoing challenge (Stocker & Qin 2014). The reason for this is the extreme complexity of the dynamic response of the at-mosphere and the couplings of the mechanisms involved, which preclude simple sensitivity analyses.

In this context, several simplified approaches have been developed for testing the sensitivity of the climate system to the solar forcing. A common one is a multilinear regression analysis wherein a climate signal, typically the global surface temperature anomaly, is modelled as a linear superposition of various contributions (Lean & Rind 2008; Gray et al. 2010). Foremost among these is the concentration of greenhouse gases (GHG), volcanic cooling expressed in terms of tropospheric

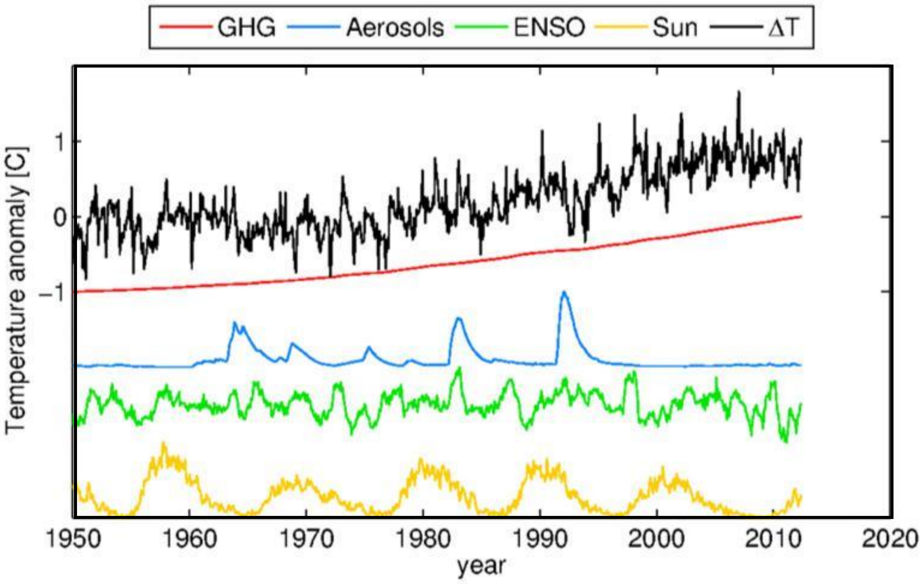

Fig. 3. The global surface temperature anomaly (the dependent variable) and its four re-gressors, from top to bottom: the concentration of GHG, the concentration of tropospheric aerosols, the strength of ENSO, and the level of solar activity. Monthly averages are used. All quantities are in arbitrary units, except for the temperature anomaly.

aerosols, and the internal variability of the climate system, whose dominant mode is the El Ni˜no Southern Oscillation (ENSO). On top of these comes the solar forcing, which is represented by the sunspot number.

This simple approach has received considerable attention in the literature. Many cases of misuse have been reported too (Benestad & Schmidt 2009). A typical case is illustrated in Figure 3, in which the monthly-averaged global sur-face temperature anomaly ΔT is expressed as a combination of four contributions

$$\Delta T(t) = b_0 + b_1 GHG(t) + b_2 AEROSOL(t) + b_3 ENSO(t) + b_4 S(t) + (t)$$

which are GHG, aerosols, ENSO and the Sun; stands for the residual. For a com-plete description of the various observables, see for example (Lean & Rind 2008). Fortunately, all regressors are almost totally uncorrelated here, so that collineary is not an issue. However, examples abound wherein the number of regressors, or collinearity, become serious issues (Crooks & Gray 2005). Well-established ap-proaches, such as partial least squares then apply (Montgomery et al. 2012), but are often ignored in practice.

There are three frequent problems which we wish to concentrate on. The first one deals with the identification of the solar signal through bandpass filtering. On the time scales of interest (months to decades), the solar contribution mostly consists of a monochromatic signal, in contrast to all other explanatory variables.

The solar forcing indeed has a period of about 11 years, and a slowly varying amplitude. A simple strategy for emphasising its signature in the highly dynamic climate signal then consists in applying a 11-year bandpass filter, see for example (Gleisner & Thejll 2003).

Let $\Delta T^*(t)$ then denote the bandpass filtered version of the dependent variable (in our case the global surface temperature anomaly) and $S^*(t) \approx S(t)$ that of the sunspot number. In that case, we should have approximately

$$\Delta T^*(t) = b_4 S^*(t) + {}^*(t).$$

The major advantage of this filtering is the elimination of all other regressors, assuming that they have little spectral power content around 11 years. Unfortu-nately, such a filtering artificially enhances the correlation between $\Delta T^*(t)$ and the solar signal, thus giving the false impression that there truly is a solar signature hidden in the climate noise. Examples abound, wherein false correlations were reported because of such filtering (Coughlin & Tung 2006). One solution would be to consider the regression coefficient b_4, which is less impacted.

A second common problem arises from adding to the regressors all known contributions to the temperature anomaly, based on the widespread belief that this should eventually reduce the residual error, and consequently improve the description of the solar contribution. This is yet another example wherein model bias is reduced at the expense of a larger variance. Typical regressors, in addition to the four ones we just presented, are various internal modes of the climate system, such as the North Atlantic Oscillation (NAO). Unfortunately, these modes are not fully independent, and actually occasionally synchronise. Therefore, the model rapidly tends to become ill-conditioned. Orthogonalisation may cure here the numerical problem, but will not solve the physical one. What we need is a robust strategy for identifying the most significant regressors. This is once again a common problem in linear regression, but certainly not a trivial one.

Various interesting solutions have been developed for that purpose. One of them is based on the error reduction ratio (ERR), which is a criterion for pruning the set of regressors (Korenberg et al. 1988). Using matrix notation, let y be the vector containing the dependent variable, X the matrix of regressors, b the model coefficients, and e the residuals:

$$y = Xb + e.$$

The matrix $X = WA$ can be decomposed into the product of an upper triangular matrix A and a matrix with orthogonal columns W. Then

$$\begin{aligned} y &= Xb + e \\ &= (XA^{-1})(Ab) + e \\ &= Wg + e \end{aligned}$$

which leads to

$$\frac{1}{N} y^T y = \frac{1}{N} \sum_i g_i^2 w_i^T w_i + \frac{1}{N} e^T e.$$

The three terms respectively describe the variance of the dependent variable, of the explanatory variables, and the unexplained variance. From this, we define the ERR as

$$ERR_i = \frac{g_i^2 w_i^T w_i}{y^T y}$$

with

$$\sum_i ERR_i + \frac{e^T e}{y^T y} = 1.$$

The strategy then consists in determining the ERR for each regressor, and select-ing those that offer the highest ERRs until their sum reaches a predefined level. This criterion is widely used for trimming linear and nonlinear parametric models (Billings 2013), but dissemination outside of that field has been slow.

A third interesting problem is the frequent association between regression models and measures of correlation. On many occasions, when the influential variables are not properly known, or too numerous, then the correlation between one of them and the dependent variable is used as an alternative for their link, see for example (Sfˆıcˇa & Voiculescu 2014). The usual gauge of correlation is Pearson’s parametric correlation coefficient. Spearman’s non parametric rank correlation coefficient has received surprisingly little attention, although it is more appropriate for handling variables that have a monotonic but nonlinear relationship (Feigelson & Babu 2012).

Unfortunately, correlation coefficients are frequently interpreted as a measure of the influence of the explanatory variables on the dependent variable, which is incorrect (King 1986). Indeed, nothing attributes causal or sensitivity assumptions to the correlation coefficient. (Cook & Weisberg 2009) express this idea in a more elaborate way when stating that regression analysis aims at understanding “as as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors”.

There are few cases, however, in which the correlation coefficient does bring some value to other statistical tools. One of them is the comparison of two equa-tions with the same dependent variable but not the same set of explanatory vari-ables. Fortunately, this is precisely what many of us are after when seeking to identify the most influential variables.

4 Power laws: Are there lines everywhere?

Power laws are ubiquitous in nature, and are often considered as a signature of self-similarity, possibly caused by critical processes (Sornette 2004; Aschwanden 2013). A quantity x follows a power law if it is drawn from a probability distribution

$$p(x) \propto x^{-\alpha}.$$

In astrophysics, power laws often show up in distributions, for which the Poisson statistics apply (Maschberger & Kroupa 2009; Andreon & Hurn 2013). Here, we consider instead empirical data for which $p(x)$ is not necessarily a distribution. The noise properties differ, and so do the solutions for estimating the slope, or scaling parameter α. Typical examples $p(x)$ are power spectral densities, for which $p(\omega) \propto \omega^{-\beta}$.

In most power law regression problems there are two objectives: one is to estimate the slope α, such as the $-5/3$ spectral index in Kolmogorov's model of fully developed turbulence. A second, and often even more important objective is about locating the transition from this power law to another functional expression, which then tells us about the characteristic scale of the system. There should be a third objective, which is to determine whether the power law model is actually the best one, but this aspect is often left out.

The quest for power laws is particularly frequent in solar wind turbulence. Here, the power spectral densities of the electric or magnetic fields are the key to the understanding of the underlying microphysical processes. Of particular inter-est are their extension down to small (so-called kinetic) scales, where dissipation processes set in (Sahraoui et al. 2009; Alexandrova et al. 2013).

Figure 4 illustrates a typical power spectral density from the solar wind. Fol-lowing common practice, we identify power laws by the approximately straight line fit in a log-log representation. The human eye is indeed remarkably good in detecting such straight lines. It can also be easily led astray by other types of distributions, such as log-normal or exponential. Examples abound, in which so-called evidence for power laws could be equally well, or even better explained by other distributions. As a rule of thumb, straight line fits should not be applied to ranges that cover less than a decade. Even a decade, however, may be too optimistic, not to mention the excessive number of digits used to express the value of the slopes.

Let $\{x_i\}_{i=1,\dots,n}$ be the observed values of a random variable, which is known to follow a power law in the interval $x_{max} \geq x_i \geq x_{min}$. Under the condition where x_{max} x_{min}, the maximum likelihood estimator of the slope is

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}$$

Unfortunately, this estimator is very sensitive to the initial guess of x_{min}, whose underestimation can severely impact the value of $\hat{\alpha}$. Clauset et al. (2009) provide useful guidance for properly estimating both the interval and the slope by max-imum likelihood. But this is only part of the challenge. The second and more difficult challenge is their validation, which still remains a largely unsolved prob-lem when considering the bounds of the interval. Bootstrapping is one of the ap-proaches that may help us estimate confidence intervals (Feigelson & Babu 2012).

This problem of estimating the power law is equally challenging when $p(x)$ is not a distribution. However, a new and even more challenging problem arises: the error-in-variable now cannot be neglected anymore. That is, the error in the

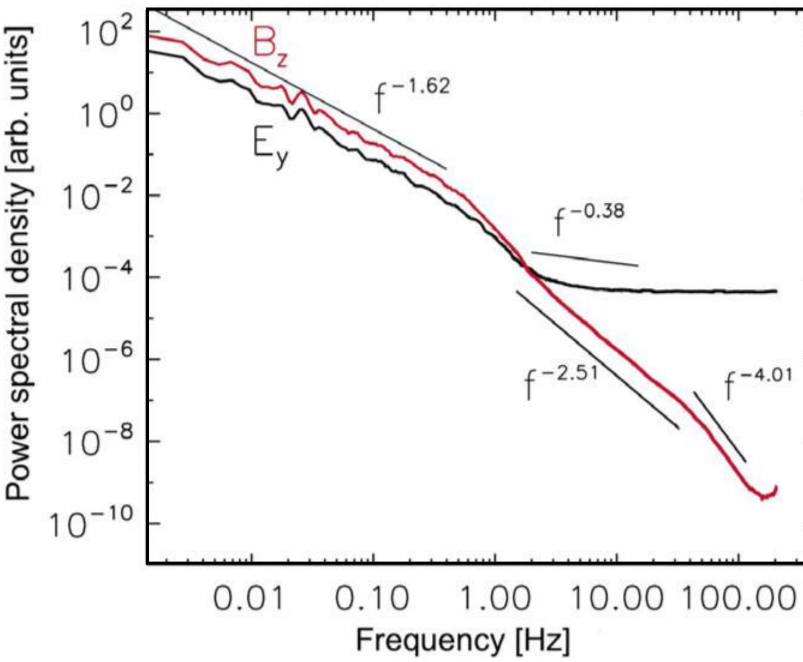

Fig. 4. Power spectral density of a component of the electric field (E_y) and a component of the magnetic field (B_z) as measured in solar wind turbulence, respectively by the EFW and FGM/STAFF-SC instruments onboard the Cluster 4 spacecraft. The straight lines are power law fits of the spectra. Figure adapted from Sahraoui et al. (2009).

independent variable x can substantially affect the conclusions, in addition to the error in the dependent variable $p(x)$.

Various solutions have been proposed for dealing with this error-in-variable case. For least squares regression, there is the natural generalisation to total least squares. Let us, however, take one step back. With standard maximum likelihood approaches, we are answering the question: "How do the data I observed match my model?". Actually, we should rather ask "What do I know about the model parameters even before collecting the data?". The difference is not just philosophical, because the two questions lead to substantially different approaches, a frequentist one in the first case, and a Bayesian one in the second case (Gelman et al. 2013). A discussion on their differences is definitely beyond the scope of this text. The main asset of the Bayesian approach lies in its ability to explicitly add prior information that may help further constrain the power law model. In cases where the statistics on the number of events is poor, such as with low photon counts, the Bayesian approach excels in extracting information from noisy data (van Dyk et al. 2001). In our case, the Bayesian setting enables the estimation of the best range $[x_{min}, x_{max}]$ and slope α simultaneously, rather than sequentially.

However, it would be fair to conclude that the numerical implementation of the solution can be challenging too.

5 Conclusion

Not surprisingly, regression analysis is key concept solar-terrestrial sciences, with implications that may sometimes have far-reaching societal consequences. In many applications, there is a widespread belief that issues such as collinearity, and lack of knowledge of what the influential variables should be, are some state of nature against which nothing can be done. The examples we have shown, however, suggest that there is much room left for improvement – provided we can learn from these mistakes and not repeat them. Therefore, in some sense, the mistakes that were risen in this short article, should be seen as a strong incentive for going back to basics and determining whether our model matches our true assumptions. There is no doubt here that the Bayesian framework will strongly help in a near future, as it has already in astrophysics.

References

Acton, F., 1970, Numerical Methods that work, Harper and Row (New York)

Alexandrova, O., Chen, C.H.K., Sorriso-Valvo, L., Horbury, T.S., & Bale, S.D., 2013, Space Sci. Rev., 178, 101

Andreon, S., & Hurn, M.A., 2013, Rev. Stat. Anal. Data Mining, 6, 15

Aschwanden, M.J. (ed.), 2013, Self-Organized Criticality Systems (Open Academic Press)

Benestad, R.E., & Schmidt, G.A., 2009, J. Geophys. Res. Atmospheres, 114, 14101

Berk, R.A., 2004, Regression analysis: A constructive critique, Vol. 11 (SAGE Publica-tions)

Billings, S.A., 2013, Nonlinear System Identification: NARMAX Methods in the Time, Frequency and Spatio-Temporal Domains (John Wiley & Sons, Ltd. New York)

Clauset, A., Rohilla Shalizi, C., & Newman, M.E.J., 2009, SIAM Rev., 51, 661

Cook, R.D., & Weisberg, S., 2009, Applied Regression Including Computing and Graph-ics, Vol. 488 (John Wiley & Sons, New York)

Coughlin, K.T., & Tung, K.K., 2006, J. Geophys. Res. Atmospheres, 111, 24102 Crooks,
S.A., & Gray, L.J., 2005, J. Climate, 18, 996

Feigelson, E.D., & Babu, G.J., 2012, Mod. Stat. Meth. Astron., with R Applications (Cambridge University Press, Cambridge, UK)

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., & Dunson, D.B., 2013, Bayesian Data Analysis, CRC Texts in Statistical Science (Chapman & Hall, London)

Gleisner, H., & Thejll, P., 2003, Geophys. Res. Lett., 30, 1711

Good, P.I., & Hardin, J.W., 2012, Common Errors in Statistics, 4th edition (and How to Avoid Them) (John Wiley & Sons)

Gray, L.J., Beer, J., Geller, M., et al., 2010, Rev. Geophys., 48, 1

King, G., 1986, Am. J. Polit. Sci., 30, 666

Korenberg, M., Billings, S., Liu, Y., & McIlroy, P., 1988, Int. J. Control, 48, 193

Laˇstoviˇcka, J., Mikhailov, A.V., Ulich, T., et al., 2006, JASTP, 68, 1854

Laˇstoviˇcka, J., Solomon, S.C., & Qian, L., 2011, Space Sci. Rev., 168, 113

Lean, J.L., & Rind, D.H., 2008, Geoph. Res. Lett., 35, 18701 Maschberger, T., &

Kroupa, P., 2009, MNRAS, 395, 931

Montgomery, D.C., Peck, E.A., & Vining, G.G., 2012, Introduction to Linear Regression Analysis, Vol. 821, Wiley Series in Probability and Statistics (John Wiley & Sons)

Pearl, J., 2009, Causality: Models, Reasoning, and Inference, 2nd edition (Cambridge University Press, Cambridge, UK)

Rong, Y., 2000, Environ. Forensics, 1, 213

Sahraoui, F., Goldstein, M.L., Robert, P., & Khotyaintsev, Y.V., 2009, Phys. Rev. Lett., 102, 231102

Sfˆıcˇa, L., & Voiculescu, M., 2014, JASTP, 109, 7

Sornette, D., 2004, Critical Phenomena in Natural Sciences: Chaos, Fractals Self-Organization and Disorder: Concepts and Tools, Springer Series in Synergetics, 2nd edition (Springer Verlag, Heidelberg)

Stocker, T. & Qin, D., 2014, Climate Change 2013 – The Physical Science Basis. Work-ing Group I Contribution to the Fifth Assessment Report of the IPCC (Cambridge University Press, Cambridge)

van Dyk, D.A., Connors, A., Kashyap, V.L., & Siemiginowska, A., 2001, ApJ, 548, 243