A quantitative score to compare data and time series including the errors.

Ruprecht Jaenicke[1], Kexue Li[2]

[1]Institute for Physics of the Atmosphere, University, Mainz, 55099, Germany
[2]Max-Planck-Institute for Chemistry, Mainz, 55128, Germany; Presently Department of Materials, University, Oxford, OX1 3PH, UK

Correspondence to: Ruprecht Jaenicke (jaenicke@uni-mainz.de)

Abstract. In science powerful statistical means exist for analysis of time and data series of observations. However, quantifying the agreement (disagreement) between two series in terms of a number, a quantity does rarely exist to our knowledge. In addition, to our awareness, the inaccuracy (error) of the measurements, observations, and models is not included in the characterization of the agreement/disagreement. The comparison (extend of agreement/disagreement) mostly is made visually and thus remaining rather qualitative, disputable, and vague. This short notice is proposing a quantitative score (SCORE) or index of agreement for a comparison of time and data series including their errors.

1 Introduction.

Wenn ich zum Beispiel vermute "Im Kühlschrank könnte noch Bier sein" und ich gucke nach, dann betreibe ich im Prinzip schon eine Vorform von Wissenschaft. Großer Unterschied zur Theologie. Da werden Vermutungen in der Regel nicht überprüft. Wenn ich also nur behaupte "Im Kühlschrank ist Bier", bin ich Theologe. Wenn ich nachschaue, bin ich Wissenschaftler. Wenn ich nachsehe, nichts finde und trotzdem behaupte, es ist Bier drin – dann bin ich Esoteriker (Vince Ebert, 2017)[1]

Comparison (testing, verification) is a powerful, if not the most central instrument in physics, chemistry, biology, medicine, sociology, and environmental science, in daily use or once any scientific idea has been perceived. Measurements are repeated many times (samples are taken or drawn) in order testing and verifying a certain result. The observations certainly are showing errors but hopefully the data vary only by a slight amount. As an outcome a set of data is created:

- Theoretical calculations and observations in time are completed piece by piece, creating two series. This way the algorithms of models are tried to justify.
- Instruments running side by side for "calibration" (verification) are creating time series and seeding an idea, how good instruments do agree. Even if the instruments are manufactured identical, the data vary as a result of error and variation of the observed entity.
- Time series of the same observable are generated at different points in the past planning to develop predictions for the future.

[1] Vince Ebert, German Comedian, 2017 (http://www.vince-ebert.de/, 14 May 2017). Translation by him: "Scientific thinking is basically the testing of assumptions. For example, if I say: there is beer in the fridge and I go and check, I'm behaving like a scientist. If I say: there is beer in the fridge but don't check. I'm a theologian. If I do look into the fridge, find no beer and still say there is - then I'm an esoteric."

- Two different series are compared (correlated) hoping to get a first hint of a meaningful (functional) dependency.

Often series as function of time are graphically presented, showing how good they agree or disagree (Li et al., 2017)[2]. The degree of agreement is estimated visually by eye (!) only. Consequently, the agreement (disagreement) could be subject to debate, depending on personal preference and experience. Graphical presentations of time series "often are influenced" to boost certain conclusions. Thus visual comparisons are very problematic. It is not the purpose of this paper to develop sophisticated equations and discuss statistical tests. It is the purpose to help in quantifying what the eyes are seeing, taking into consideration the error of the data.

2 Basics.

Any physical (chemical) variable (experimental or modelled) like concentration, temperature, wind velocity v could be seen as the product of a numerical value and an unit (Cohen et al., 2008), as in "v = 12.5 m/s". All of those variables have uncertainties (errors). Those uncertainties come in two flavours, instrumental (experimental, modelled) uncertainties and the variability of the observed variable. It is therefore essential reporting the numerical value of the variable with a range of uncertainty. Usually the variable then is noted as "$v = 12.5 \pm 0.1$ m/s" or similar schemes meaning "face value" ± "uncertainty". Uncertainty in the model of Gaussian distribution means, that at a level of confidence of about 68%, the numerical value is in that range. Ascertaining the uncertainty, a number of independent experiments[3] have to be carried out (or estimated from the setup). Those measured uncertainties (residuals) are randomly distributed at best. Systematic errors should be treated separately. Unfortunately often environmental quantities are given without this information.

In addition, environmental variables (and variables in models) are mostly subject to variations in time and space. Sometimes these variations are "included" in the uncertainty of the quantity and sometimes not. These variations depend mainly on the atmospheric residence time of the variable and the distribution of the sources. Variables with a short residence time show large variabilities, while those with long residence times, like oxygen, vary only slightly (Hamrud, 1983). This variability might be veiled, especially in integrating observations, like collecting particles (up to several weeks) on filters for later chemical and/or biological analysis. If the variable is measured over extended distances (using attenuation of electromagnetic radiation for instance), mainly the variability in space is veiled. Measurements of variables over a certain time form a time series. In such a time series the data are not necessarily independent from each other. That is easy seeing, because frequently recurrent observations of temperature at any meteorological station never fill the complete possible range of temperatures at that station. Statistical evaluations fail, if data are not independent from each other. Regardless of this, "statistical evaluations" are carried out in publications and reports. Any environmental observation should carry both uncertainties, like "$physical\ variable = \left(face\ value \pm error \pm variability\right)$ unit" or "$v = \left(12.5 \pm 0.1 \pm 0.5\right)$ m/s", for example.

The standard deviation s is calculated from a number n of independent measurements and serves as an estimate of the parameter (σ) of the Gaussian distribution.

[2] This reference serves as an example only. The content of the paper has no special meaning for this paper.
[3] Experiments, not only observations

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \quad (1)$$

with x_i= measured or calculated value
n = number of the independent measurements
i = counter

s is the standard deviation of a normal distribution with the mean m. Only the positive values of this equation are used in this work. The standard deviation should be calculated from a set of independent data. For the above calculation the rule (n-1) of estimating s in the case of a small sample number n of measured values has been used. It is the general understanding, that every data x_i of later series carries that error s as well. That means that even outliers are carrying that error (they might be "cleaned in later statistical treatment").

In numerical modelling likewise errors (approximation errors) exist because numerical calculations are prone to truncation, round-off, limited calculation time, imperfect assumptions, and others. The resulting data also could be characterized by a value x_i and an error s.

In evaluating data sets, often Pearson's correlation coefficient r and a regression function are calculated. The correlation coefficient r delivers values between -1 and 1. However, the correlation coefficient r is not taking into account the error of the individual data. So data sets with different measuring uncertainties could result in identical correlation coefficients. Regression functions could be calculated from those data with uncertainty ranges of their coefficients. This gives an assessment how good the data follow the regression function. Again, the error of the measured values is not entering these calculations.

If the error of any data point is assumed as "standard deviation", each data point could be understood as part of a Gaussian distribution (with the standard deviation based on the error). Two separate data points then might intersect in their Gaussian distribution to certain extend. This intersection could be used as a quantitative measure of agreement. If the data points are extensively separated, there might be "no intersection". But, measurements with large errors (inaccurate measurements with large uncertainties) tend intersecting more frequently and more broadly than accurately measured data. Using only the intersection as a quantitative measure of agreement would foster inaccurate measurements and degrade accurate measurements.

It is the assumption in this paper, that in time series, the time is "accurate", meaning no error in time is assumed. For many cases that certainly apply, as for daily temperatures or tree rings (they are counted). It remains to be discussed, how an uncertainty in time should be handled. In data series, close data points could be independent. In time series, that mostly is not the case (as easily could be seen in temperature time series).

3 Proposal.

To improve the comparison of two data sets or time series, we propose using an empirical equation with the errors of any pair of data points of the two time (data) series defining a penalty factor PF. This way the set of data points could be characterized with a SCORE, a quantitative value, how well they compare. To allow an extensive application of the proposed equation, great care has to be taken avoiding any singularities as could happen in data points close to zero or negative. The use of the SCORE permits data (like temperature) extending into the range below zero.

$$SCORE(\text{unit}_1; \text{unit}_2) = PF \cdot \frac{1}{s_1\sqrt{2\pi}}\int e^{-\frac{(x-m_1)^2}{2s_1^2}}\,\mathrm{d}x \cap \frac{1}{s_2\sqrt{2\pi}}\int e^{-\frac{(x-m_2)^2}{2s_2^2}}\,\mathrm{d}x \quad (2)$$

with - $SCORE(\text{unit}_1; \text{unit}_2)$ = proposed quantitative measure comparing data pairs in the range 0 to 1.
- $\text{unit}_{1|2}$ = "used unit"
- $m_{1|2}$ = measured (determined) values of the data pairs. In contrast to other use, $m_{1|2}$ are measured data, not averages. $m_{1|2}$ might carry units, they also could have negative values.
- $s_{1|2}$ = errors of the data pairs. Those errors carry the same units as the measured values. As usual, the errors could have been determined separately or estimated.
- ∩ = intersection. The intersection describes which fractions of the Gaussian (error) distributions of the pairs do overlap. That fraction might get values between 0 and 1.

The proposed penalty factor is *PF*.

$$PF = e^{-\frac{s_1}{std_1}} \cdot e^{-\frac{s_2}{std_2}} \cdot e^{-\frac{(\overline{m_1}-\overline{m_2})^2}{std_1 \cdot std_2}} \cdot e^{-\frac{(m_1-m_2-(\overline{m_1}-\overline{m_2}))^2}{std_1 \cdot std_2}} \quad (3)$$

with - $\overline{m_{1|2}}$ = arithmetic means of the measuring series.

- $std_{1|2}$ = standard deviation of the measuring series.

With $(\overline{m_1} - \overline{m_2})^2$ a systematic error between the measuring series is considered.

$SCORE(\text{unit}_1; \text{unit}_2)$ is getting values of zero and greater. Zero means no agreement of the pairs. $SCORE(\text{unit}_1; \text{unit}_2) = 1$ would indicate "perfect agreement" of the pairs. One has to keep in mind that $SCORE(\text{unit}_1; \text{unit}_2)$ also might carry the influence of the variance of the measured or calculated variables. Any comparison and $SCORE(\text{unit}_1; \text{unit}_2)$ has to be seen in that light.

In averaging the *SCORE* of the individual pairs, a *SCORE* of the whole data series could be calculated.

$$\overline{SCORE} = \frac{1}{n}\sum_{i=1}^{n} SCORE_i \quad (4)$$

Of course, that averaged $\overline{SCORE}$ also has a standard deviation.

To explore the proposed *SCORE*, first a comparison of two identical time series is presented.

For this example, the instruments used and their performance are of negligible interest. In the sample, $s_{1|2}$ = 0.3°C has been assumed and $SCORE$(°C;°C) = 0.9283 is calculated. $SCORE$(°C;°C) values for different $s_{1|2}$ are shown in Table 1. As expected, $SCORE$(°C;°C) is increasing and approaching 1 as the measurements are getting more accurate and the errors $s_{1|2}$ are becoming smaller. So accurate measurements are rewarded and less accurate are degraded.

Exploring the properties of *SCORE*, this data set could be used for a kind of autocorrelation: Calculating the *SCORE* as function of time lag. The time series with an $s_{1|2}$ = 0.3°C has been selected. Time lag 0 results in values as above. Time lag 1 has a low *SCORE* with a "reasonable good" correlation r^2 = 0.8811. This could be an indication, that *SCORE* is rather sensitive (Table 2).

Expanding those simple examples, the SCORE of two (one step) successive atmospheric temperature (2 m above ground) measuring series has been calculated. That results in *SCORE* (°C;°C) = 0.1871 ± 133 %. That might serve as an indication of the temperature regime of the two years. The average temperature of one year was 15.3°C, while the other showed 14.4°C.

4 Applications and Discussions, Examples.

In acquainting with the *SCORE* and getting a sense, about the values and sensitivity, a selection of data and time series from instruments and models have been asked for, calculated and are presented.

4.1 Optical Aerosol Particle Counters and Condensation Nuclei Counters.

The size distribution (number concentration density as function of particle size of an equivalent sphere) of the atmospheric aerosol is one of its key parameters. It determines (among other parameters like index of refraction, chemistry, etc.) the colour of the sky, radiation balance, penetration of particles into the lungs, "water" precipitation efficiency of the atmosphere. Over the years a large variety of instruments has been developed and is offered by manufacturers using different properties of the aerosol to be measured (Baron and Willeke, 2001). Among them, optical particle counters are often used.

The ideal comparison would be running two identical instruments (fresh from the production line) side by side whether under laboratory or ambient conditions. Most research groups do not have access to such identical instruments. They often have only a collection of different instruments. So, it should be the exclusive privilege of any prominent producer to have such data at hand[4].

A comparison of two instruments is very crucial. It not only informs about the performance and the errors of the instruments, it also gives an impression about the variability of the aerosol. The atmospheric aerosol has a rather short residence time and consequently a rather large variability.

The Grimm 1.107 monitor is designed for airborne particulate size concentration measurements using 90° laser light scattering. Aerosol passes through a plane beam produced by a laser diode. A pulse height analyser detects the scattered optical signals and classifies them. The measuring range is given as 0.25 to 32 µm particle diameter in 31 size bins. The Grimm 1.109 monitor is very similar[5]. In principal 31 data pairs could be used, but not all of them have been reported in the example.[6] Similar me3asurements are published in (Kaaden et al., 2009).

The comparison (Figure 2) shows very typical aerosol size distributions with 22 data pairs. The regression coefficient $r^2 = 0.9980$ points toward an excellent agreement. Most experimenters and manufacturers would praise such a comparison. And indeed, especially in the view of examples following later, the *SCORE* (cm^{-3}; cm^{-3}) = 0.5419 does reflect a good agreement. The *SCORE*s standard deviation in this example is 0.1405 (26 %). The measuring error is based on counting statistics. The counting error could even be lowered by increasing the measuring time (the example has measured in a 6 s time slot). That would result in a trade in of counting statistics and variability of the aerosol. Any extended measuring time includes an averaging.

Using a research flight with the Russian aircraft Geophysica, two Condensation Nuclei Counters (COPAS) have been used in EUPLEX (European Polar Stratospheric Cloud and Leewave Experiment), Kiruna 2003[7]. Details are discussed elsewhere (Weigel et al., 2009). Two COPAS instruments (COPAS1, COPAS2) have been placed, with identical inlet systems, in one aircraft container. The instrumental temperatures have been adjusted, so particles greater than 10 nm are activated.

The time series of February 2, 2003 (10 – 30° E; 68 – 76° N) consist of 286 data pairs. The flight elevation during measurements ranged from 11500 – 12000 m. The outside temperature fluctuated around 190 K. The correlation coefficient results in $r^2 = 0.9400$, a very good value for observations on aircrafts. For COPAS1 and COPAS2 a *SCORE*(cm^{-3};cm^{-3}) = 0.316979 is an excellent value calculated so far. Those two instruments compared very well.

---

[4] Early in 2017, the major sellers of optical particle counters in Europe were unable providing such data.
[5] http://www.grimm-aerosol.com/company/grimm-aerosol/index.php, 9 Nov 2016
[6] Data provided by Prof. Dr. Konrad Kandler,
http://www.geo.tu-darmstadt.de/iag/personen/mitarbeiter_details_geo_4444.de.jsp
[7] Data provided by Dr. Ralf Weigel, https://www.blogs.uni-mainz.de/fb08-ipa/aerosol-und-wolkenphysik/

4.2 "Global" Temperatures and Tree Rings.

There are numerous efforts to retrieve past global atmospheric temperatures from proxy-data. One (out of many) method is using tree rings (Wilson et al., 2016) resulting in an atmospheric temperature time series. It is tempting to compare such time series with "observed" instrumental temperatures. Like observed instrumental temperature data, tree rings provide a very precise annual time dating. Tree ring observations could extend over a very long time scale and thus could look far into the past.

"N-TREND" is an initiative by dendroclimatologists to improve large-scale reconstructions of temperatures[8]. To single out the temperature dependence of tree ring growth and lower the influence of available water on it, only series north of 40°N (and high altitude) have been selected. For the same reason, summer data only have been used (MJJA – May, June, July, and August). Trees grow on land only, what certainly limits a global approach. The existing time series extends from year 750 to present. It is certainly a keen step concluding from data at high altitude (where the selected trees are growing) to surface temperatures (sea level, where the comparison is made).

As "observed" temperatures the "Met Office Hadley Centre" observations datasets HadCRUT4 time series has been selected. It has been screened for Northern hemisphere, land only, and JJA (June, July, and August), as this is the only selection available to us[9]. Errors are calculated from the upper and lower bounds of the 95% confidence intervals from the combined effects of all the uncertainty sources (i.e. the range within which the anomalies are very likely to occur according to the corresponding uncertainty model). These uncertainty sources are: station and grid-box sampling, bias uncertainties and coverage uncertainties.

Both time series are deviations of temperature relative to 1961-1990. Both time series present the data with 95% confidence intervals. For both time series, the time range 1880 to 2010 has been used. That comparison is somehow limited. Figure 4 is showing the results.

In looking at the time series, they seem to agree very nicely. The up and downs are following each other and both series show higher temperatures toward the 2020th. The correlation coefficient looks rather promising, if the difficulties in getting those time series are taken into account. However, in "non-visual" (SCORE), quantitative terms, the agreement is not very good. One of the reasons is certainly the "large" uncertainty ("error") of the tree ring proxies.

To a certain extent, the tree ring temperature data depend on observed instrumental data (on the HadCRUT4 set?) set as well, as those data are used developing the temperature time series from tree rings. That means, the data sets are not independent from each other.

5 Conclusion.

An empirical equation has been proposed to calculate a SCORE, comparing two data or time series of observed or modelled variables including the errors. The proposed SCORE is a sensitive parameter. It permits a quantitative ranking of any comparison of two series as compared to visual (qualitative) comparison. The selected examples are from different fields. It remains for further applications to explore the full realm of

[8] https://ntrenddendro.wordpress.com/, 20 March 2017. The data, including the errors, have been provided by Prof. Dr. Jan Esper, Institute of Geography, University of Mainz, https://www.blogs.uni-mainz.de/fb09climatology/staff-and-students/jan-esper/

[9] http://www.metoffice.gov.uk/hadobs/crutem4/data/diagnostics/hemispheric/northern/CRUTEM.4.5.0.0.nh_JJA, 1 May 2017

SCORE. One could easily expand the SCORE to two dimensional (maps) or even three dimensional comparisons. Two dimensional comparisons are presently often colour coded and the visual comparison thus depends on the colour sensitivity of the eye of the reader (what, if the reader carries colour blindness). A quantitative ranking would foster a better comparison.

6 References

Baron, P. A., and Willeke, K.: Aerosol measurement, principles, techniques and applications, John Wiley & Sons, 2001.

Cohen, E. R., Cvitas, T., Frey, J. G., Holmström, B., Kuchitsu, K., Marquardt, R., Mills, I., Paves, F., Quack, M., Stohner, J., Strauss, H. L., Takami, M., and Thor, A. J.: Quantities, units and symbols in physical chemistry, in, 3rd ed., IUUPAC & RSC Publishing, Cambridge, 2008.

Hamrud, M.: Residence time and spatial variability for gases in the atmosphere, Tellus B, 35B, 295-303, 10.1111/j.1600-0889.1983.tb00034.x, 1983.

Kaaden, N., Massling, A., Schladnitz, A., Müller, T., Kandler, K., Schütz, L., Weinzierl, B., Petzold, A., Tesche, M., Leinert, S., Deutscher, C., Ebert, M., Weinbruch, S., and Wiedensohler, A.: State of mixing, shape factor, number size distribution, and hygroscopic growth of the saharan anthropogenic and mineral dust aerosol at tinfou, morocco, Tellus, 61B, 51-63, 2009.

Li, Q., Zhang, L., Xu, W., Zhou, T., Wang, J., Zhai, P., and Jones, P.: Comparisons of time series of annual mean surface air temperature for china since the 1900s, Bulletin of the American Meteorological Society, 98, 699-728, 2017.

Weigel, R., Hermann, M., Curtius, J., Voigt, C., Walter, S., Böttger, T., Lepukhov, B., Belyaev, G., and Borrmann, S.: Experimental characterization of the condensation particle counting system for high altitude aircraft-borne application. , Atmospheric Measuring Techniques, 2, 243-258 10.5194/amt-2-243-2009 2009.

Wilson, R., Anchukaitis, K., Briffa, K., Büntgen, U., Cook, E., D'Arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B., Hegerl, G., Helema, S., Klesse, S., Krusic, P., Linderholm, H. W., Myglan, V., Osborn, T., Rydval, M., Schneider, L., Schurer, A., Wiles, G., Zhang, P., and Zorita, E.: Last millennium northern hemisphere summer temperatures from tree rings: Part i: The long term context. , Quaternary Science Reviews, 134, 1-18, 10.1016/j.quascirev.2015.12.005, 2016.

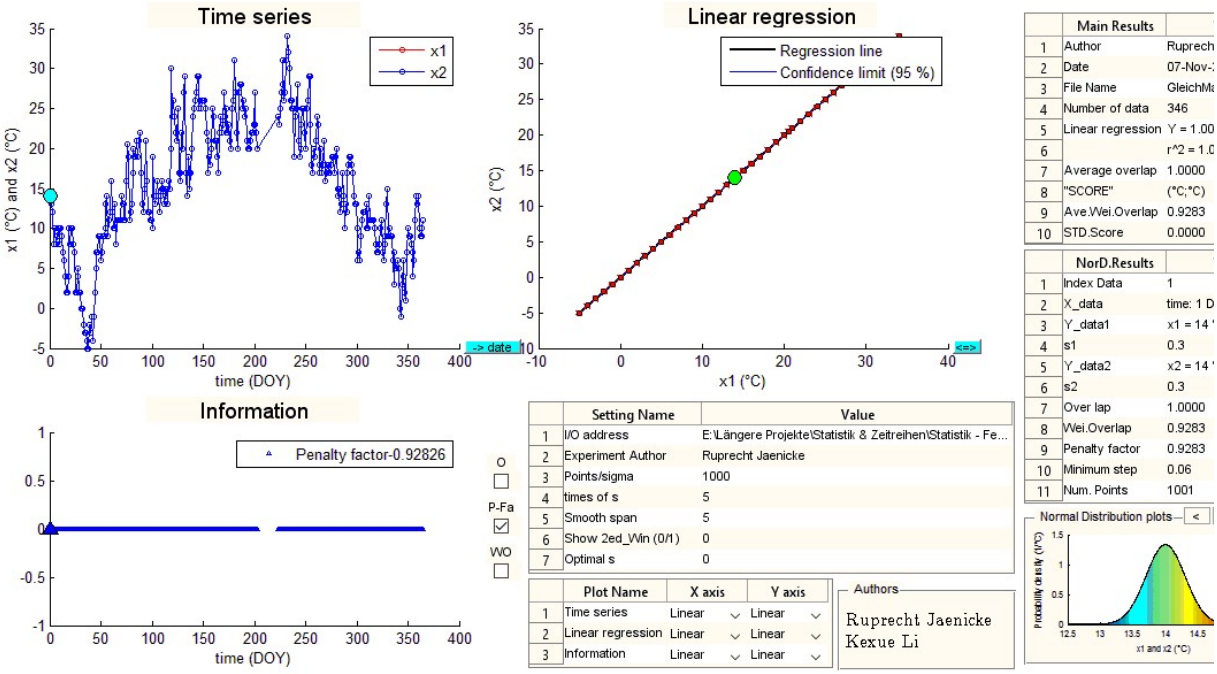

Figure 1: One (almost) complete year of daily outside air temperature measurements (2 m above ground). For testing, both time series ($x_{1|2}$) contain an identical data set (panel "Time series") as function of DOY (only one colour is seen, because the series are identical. The error of reading temperatures has been assumed to be $s_{1|2}$ = 0.3°C. The regression function (panel "Linear regression") is shown. The squared correlation coefficient r^2 = 1 (panel "Main Results"), as expected. In the lower right hand corner the Gaussian density (at any particular DOY of the time series) is showing the perfect overlap. The calculation of intersection of the Gaussian densities has been carried out in the range $\pm 5 \cdot s_{1|2}$ ("times of s" in "Setting Name"). The penalty factor is one (panel "Information" indicates 'P-Fa' = individual penalty factor – 0.92826. It has been scaled for better reading of the graph). This perfect agreement of both time series results in *SCORE*(°C;°C) = 0.9283.[10] *SCORE*(°C;°C) = 1 is missed, because of the error of the temperature reading assumed.

$s_{1\|2,}$ °C	*SCORE*(°C;°C)
0.01	0.9975
0.03	0.9926
0.10	0.9755
0.30	0.9283
1.00	0.7802

Table 1: *SCORE*(°C;°C) for different s1|2.

time lag, days	*SCORE*(°C;°C)	Standard deviation, °C	Correlation coefficient r^2
0	0.9283	0.0000	1.0000
1	0.1604	0.3159	0.8811
2	0.1175	0.2849	0.7815
3	0.0945	0.2571	0.7242
4	0.0972	0.2618	0.6893
5	0.1020	0.2703	0.6645
6	0.0869	0.2428	0.6688
7	0.0715	0.2250	0.6413
8	0.0908	0.2571	0.6083
9	0.0993	0.2583	0.5874

Table 2: Development of SCORE, its standard deviation, and correlation as function of time lag.

[10] The algorithm has been written in MATLAB R2014a. The figures are printouts and might contain more information than needed in the actual case.

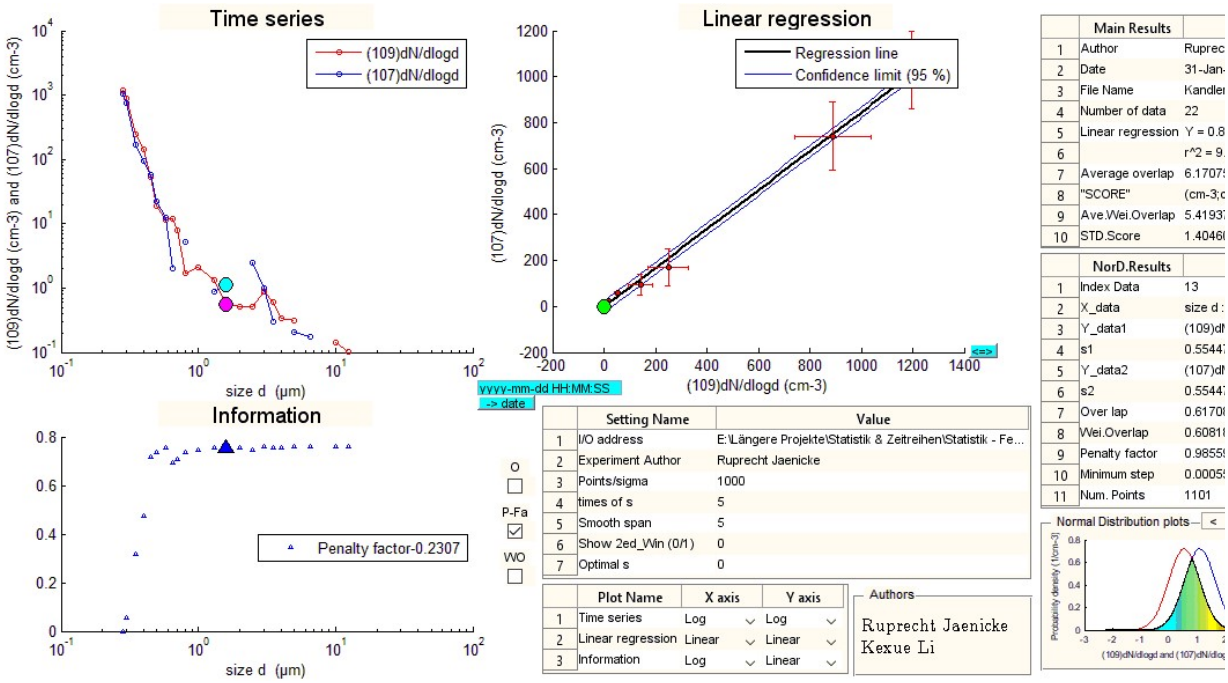

Figure 2: Comparison of Grimm 1.107 and Grimm 1.109 monitors aerosol size distribution measurements. The errors have been calculated according to the number of counted particles. That number depends on aerosol flow and counting time. The counting time was set to 6 s. The resulting *SCORE* (cm^{-3}; cm^{-3}) = 0.5419 has a standard deviation of 0.1405. The selected Normal Distribution Plot (lower right corner) shows the partial overlap of one selected pair of data.

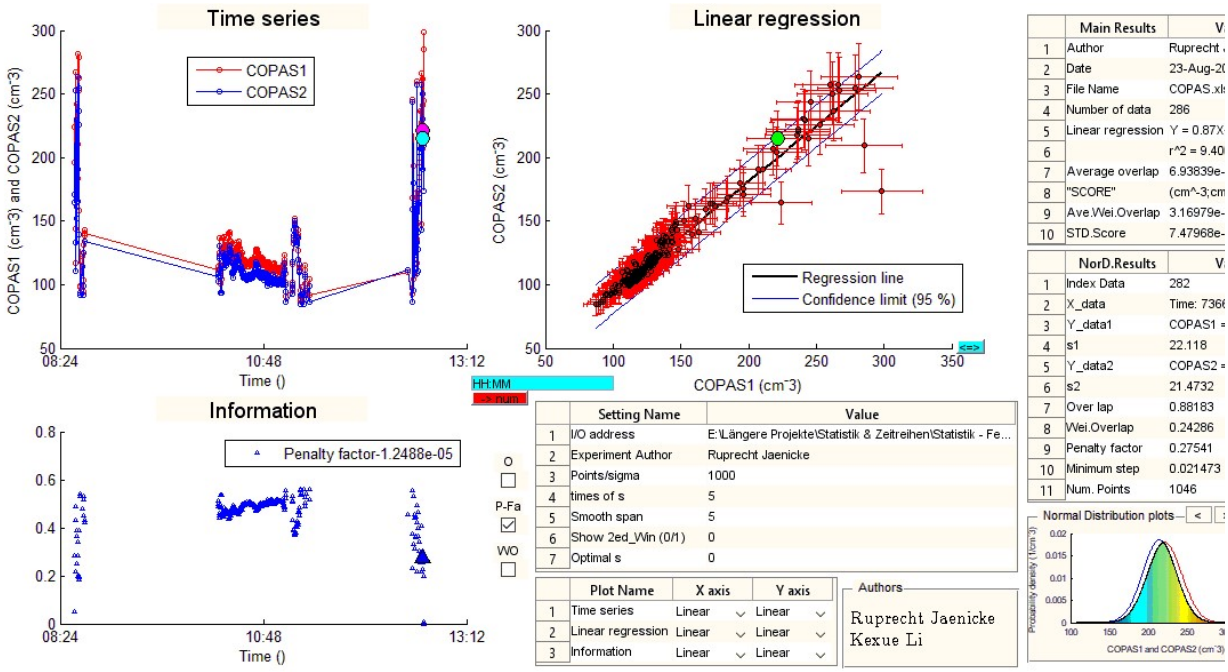

Figure 3: Condensation Nuclei Counters on the Russian research aircraft Geophysica at altitudes 11500 – 12000 m in the polar region between 68° and 76° N. The Condensation Nuclei Counters temperature was set to activate particles greater 10 nm. Only a few data pairs deviate from the main confidence limit of the regression line.

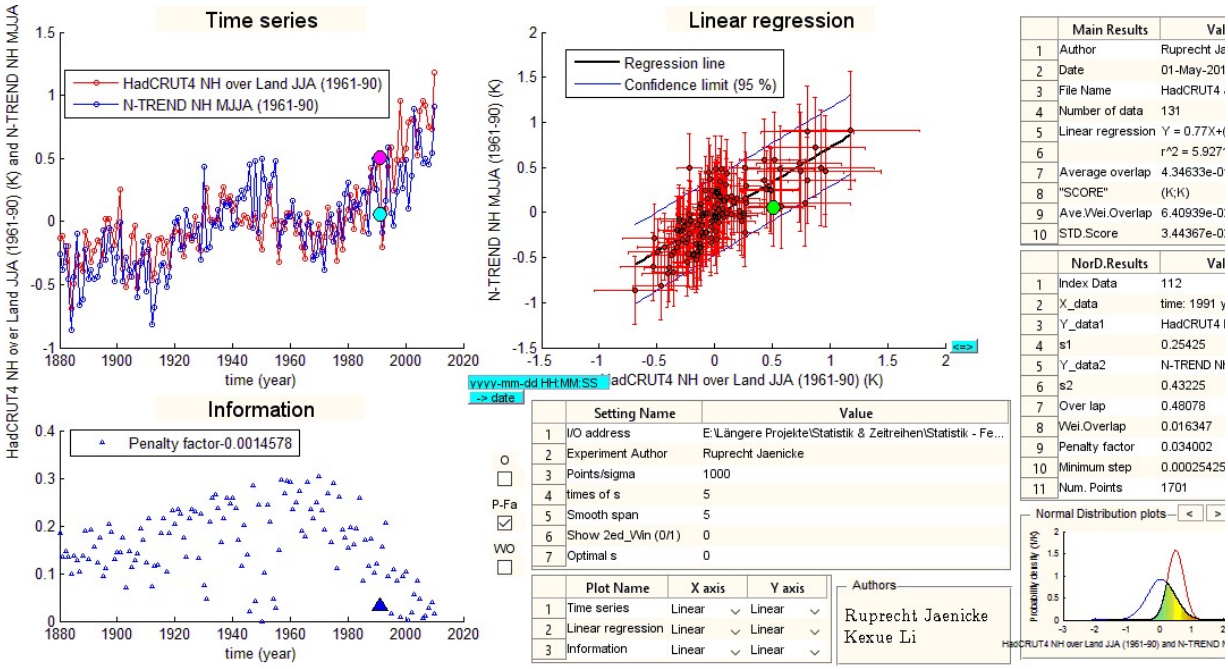

Figure 4: Time series of Northern hemisphere surface temperatures, HadCRUT4, for land only and JJA, N-TREND with limitations (northern hemisphere, extra tropical region, MJJA). Both series show the temperature deviations in K from 1961-90 and cover the time range 1880 to 2010. One year has been selected for comparison (Normal Distribution plots, lower right corner). It shows that the HadCRUT4 data have been published with rather small "errors", while the N-TREND data exhibit broader uncertainties. The correlation $r^2 = 0.59271$ looks not bad. The *SCORE* (K;K) = 0.0640939 (with a standard deviation of 3.4%) is rather small. The penalty factor (lower left corner) is getting smaller, as the years are progressing.