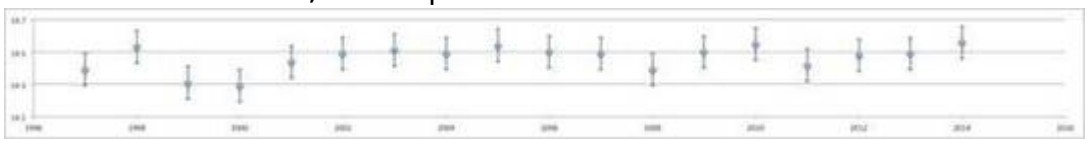
Unsicherheits-Bandbreiten, Fehlerbalken und CIs

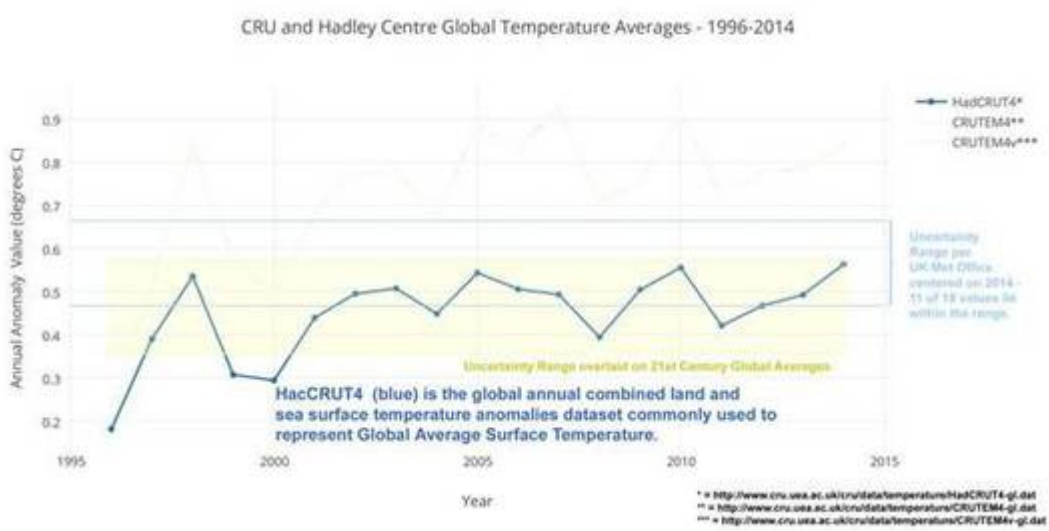
geschrieben von Kip Hansen | 16. Februar 2015

Die *NYTimes* sowie tausende anderer Nachrichtenportale haben sowohl lautstark proklamiert, dass 2014 „das wärmste Jahr jemals“ gewesen ist, aber auch die Klagen gegen diese Proklamationen. Einige, wie der Opinion Blog der NYTimes Dot Earth haben unverhohlen beides abgedeckt.
Dr. David Whitehouse hat über die GWPF in seinem Beitrag bei WUWT hier widersprochen. Sein Beitrag trägt den Titel [übersetzt] „UK Met.-Office: 2014 war NICHT das wärmste Jahr jemals wegen der ‚Unsicherheits-Bandbreiten‘ der Daten“:
„Der HadCRUT4-Datensatz (zusammengestellt vom Met.-Office und der CRU an der University of East Anglia) zeigt, dass die Temperatur des vorigen Jahres um 0,56°C (±0,1°C*) über dem langzeitlichen Mittel lag (1961 bis 1990). Nominell platziert dies das Jahr 2014 zusammen mit dem Jahr 2010 auf Rekordniveau, aber die Unsicherheits-Bandbreiten bedeuten, dass es nicht möglich ist definitiv zu sagen, welches der vielen letzten Jahre das Wärmste war“. Und unten auf der Seite: „*0,1°C ist die 95%-Unsicherheits-Bandbreite“.
Der Essay von David Whitehouse enthielt u. A. das folgende Bild: die jährlichen Mittel nach HadCRUT4 mit Balken, die die Unsicherheits-Bandbreite von ±0,1°C repräsentieren:



Das Journal *Nature* verfolgte lange Zeit eine Politik, die darauf bestand, dass Graphiken mit Fehlerbalken das beschreiben, was die Fehlerbalken repräsentieren. Ich dachte, dass dies in diesem Falle gut wäre, um genau zu sehen, was das Met.-Office mit „Unsicherheits-Bandbreite“ meint.
In den FAQ sagt das Met.-Office:
„Es ist unmöglich, die globale mittlere Temperaturanomalie mit perfekter Genauigkeit zu berechnen, weil die zugrunde liegenden Daten Messfehler enthalten und weil die Messungen nicht den gesamten Globus abdecken. Allerdings ist es möglich, die Genauigkeit zu quantifizieren, mit der wir die globale Temperatur messen können, und das ist ein wichtiger Bestandteil bei der Erstellung des HadCRUT4-Datensatzes. **Die Genauigkeit, mit der wir die globale mittlere Temperatur des Jahres 2010 messen können, liegt bei rund einem Zehntel Grad Celsius**. Die Differenz zwischen den mittleren Schätzungen für die Jahre 1998 und 2010 liegt bei etwa einem Hundertstel Grad Celsius, was weit unterhalb der Genauigkeit liegt, mit dem beide Werte jeweils berechnet werden können. Dies bedeutet, dass wir nicht sicher wissen können – jedenfalls allein aufgrund dieser Information – welches Jahr das Wärmere war. Allerdings beträgt die Differenz zwischen den Jahren 2010 und 1989 etwa 4 Zehntel

Grad Celsius, so dass wir mit viel Vertrauen sagen können, dass das Jahr 2010 wärmer ausgefallen ist als das Jahr 1989 oder tatsächlich irgendein Jahr vor 1996“. (Hervorhebung von mir).
Ich applaudiere dem Met.-Office zu seiner Offenheit und klaren Sprache in diesem einfachen Statement.
Und jetzt zu der Frage, die sich aus dieser Abbildung ergibt:



Diese Graphik wurde aus Daten erstellt, die direkt vom UK Met.-Office stammen, „unberührt von menschlichen Händen“ (keine Zahlen wurden händisch kopiert, umgeschrieben, gerundet oder anderweitig modifiziert). Ich habe die CRUTEM4-Daten vom Festland allein grau dargestellt und sie gerade noch sichtbar zum Vergleich stehen gelassen. Links zu den öffentlich zugänglichen Datensätzen werden sie in der Graphik gezeigt.
Ich habe einen Text und zwei graphische Elemente hinzugefügt:
a. In hellblau die Unsicherheitsbalken für den Wert 2014, rückwärtig ausgedehnt auf den gesamten Zeitraum,
b. Ein hellrosa Band, die Breite der Unsicherheit für diese Größe in der Graphik auf eine Weise überlagert, um deren Maximalwerte abzudecken.
Und jetzt die Frage:
Was bedeutet diese Illustration wissenschaftlich?
Genauer: Falls die Zahlen in **Ihren Fachbereich** fallen würden – Ingenieurswesen, Medizin, Geologie, Chemie, Statistik, Mathematik, Physik – und das Ergebnis längerer Messreihen wären, was würde es Ihnen bedeuten, dass:
a. 11 der 18 Mittelwerte des jüngsten Mittelwertes 2014 innerhalb der Unsicherheits-Balken liegen?
b. alle außer drei Werten (1996, 1999), 2000) überlagert werden können mit einem Band von der Breite der Unsicherheits-Bandbreite für die gemessene Größe?
Es wäre schön, wenn es Antworten gäbe aus so vielen Forschungsbereichen wie möglich.

Bemerkungen des Autors zu diesem Artikel: Ich habe keine Meinung bzgl. bestimmter Interessen hierzu – und auch keine besondere Erfahrung. (Oh, ich habe eine Meinung, aber die ist nicht gut fundiert). Mich würden Meinungen von Personen mit Forschungserfahrungen in anderen Bereichen

interessieren.
Dies ist *keine* Diskussion der Frage „war 2014 das wärmste Jahr?“ oder irgendetwas dergleichen. Einfache Wiederholungen zu den Glaubensartikeln von beiden Seiten der Globale-Erwärmungs-Kirche würden nichts Wesentliches zu dieser Diskussion beitragen und sollten anderswo gepostet werden.
Um mit Judith Curry zu sprechen: Dies ist ein technischer Beitrag – dazu gedacht, eine Diskussion zu erhalten über wissenschaftliche Methoden zur Erkennung dessen, was Unsicherheits-Bandbreiten, Fehlerbalken und CIs uns über die Forschungsergebnisse verraten können und verraten. Bitte die Kommentare auf diesen Bereich beschränken; vielen Dank.
Soweit dieser Artikel. Auf Bitte von Herrn Limburg werden hierzu abweichend von der üblichen Praxis auch noch ein paar Kommentare übersetzt. Fachlich berühren diese mich nur peripher, aber sie zeigen, wie man sachlich, konstruktiv und trotzdem kontrovers diskutieren kann. Die Hyperlinks hinter den Bezeichnungen der Kommentatoren wurden entfernt. – Anm. d. Übers.
Einige Kommentare zu diesem Artikel:
oz4caster
Ich glaube, dass sie zu vertrauensselig sind hinsichtlich einer Genauigkeit von 0,1°C für jüngste globale Temperaturanomalien. Mein Tipp wären mindestens 0,2°C bis 0,3°C bis zu 0,5°C während der letzten Jahre und möglicherweise bis zu 1,0°C in früheren Jahren in dem Datensatz. Ich habe den Verdacht, dass die größten Unsicherheitsquellen die sehr geringe räumliche Abdeckung, Repräsentativität der Messungen, Veränderungen der Messorte und „Homogenisierung“ sind, die die Unsicherheit eher vergrößern anstatt sie zu verringern. Der genaue Aufstellungsort ist kritisch für repräsentative Messungen, und das USCRN hilft, dieses Problem anzugehen, allerdings nur für einen sehr kleinen Teil des Globus‘.
Erwiderung hierzu:
george e. smith
Aus dem Artikel: „Der HadCRUT4-Datensatz (zusammengestellt vom Met.-Office und der CRU an der University of East Anglia) zeigt, dass die Temperatur des vorigen Jahres um 0,56°C (±0,1°C*) über dem langzeitlichen Mittel lag (1961 bis 1990)“
Nun, mit dieser Feststellung habe ich einige Probleme.
Zunächst: die Bezugsperiode 1961 bis 1990. Dies schließt bequemerweise jene Periode in den siebziger Jahren ein, als man die Klimakrise in Gestalt einer bevorstehenden Eiszeit kolportiert hat, zusammen mit abenteuerlichen Vorschlägen, Ruß auf dem arktischen Eis zu verstreuen, um die vermeintliche Eiszeit abzuwehren. Globale Hungersnöte wurden von genau der gleichen Bande von Kontrollfreaks vorhergesagt, die jetzt versucht, das globale Überkochen zu verhindern.
Außerdem liegt die Hälfte dieser Bezugsperiode vor der Satellitenära, die, glaube ich, etwa um das Jahr 1979 begann, was fast genau zusammenfällt mit der Ausbringung der ersten ozeanischen Messbojen, die in der Lage waren, simultan Messungen der Wassertemperatur an der Oberfläche (bis zu 1 m Tiefe) sowie Lufttemperaturen über der Oberfläche

(bis zu 3 m) durchzuführen. 2001 zeigten diese Bojendaten (über etwa 20 Jahre), dass Wasser- und Lufttemperaturen nicht gleich und auch nicht korreliert waren. Warum sollte irgendjemand sich auch nur vorstellen, dass beides doch der Fall war?
Fazit: Ich glaube keinen „globalen“ Klimatemperaturen vor dem Jahr 1980.
Und warum hört man mit dem Vergleich 25 Jahre vor heute auf?
Warum verwendet man nicht das Mittel ALLER glaubwürdigen Daten, die man hat? Anderenfalls wären die Zahlen der Bezugsperiode reine Rosinenpickerei.
Man verleihe also den HadCRUT-Daten vor 1980 keinerlei Glaubwürdigkeit oder irgendwelchen anderen Angaben, die man später aus diesen nutzlosen frühen Daten ableiten könnte.
Und schließlich glaube ich nicht, dass irgendeine Strategie bzgl. Stichproben legitim sind, steht dies doch im Gegensatz zur Theorie von Stichproben (man wäre nicht in der Lage, dies zu lesen, falls das nicht der Fall wäre).
Epiphron Elpis
Erwiderung hierauf:
Die Schlussfolgerungen sind unabhängig von der Bezugsperiode. Und irgendeine andere gewählte Bezugsperiode würde die gleichen Trends und die gleichen Temperaturdifferenzen zeigen.
Kip Hansen
Antwort an oz4caster: Ich denke, dass die Unsicherheits-Bandbreite von 0,1°C zu klein ist selbst für Messungen aus jüngster Zeit. Aber für diese Diskussion lasse ich das mal so durchgehen – es ist fast ein Wunder, dass sie eine solche Unsicherheits-Bandbreite überhaupt zugeben. Ich arbeite langfristig an einem Beitrag, der aktuelle Original-Messfehler erkundet bei der Temperatur der Welt mit der Zeit, und was dies für die globalen Mittelwerte bedeutet.
Zum Beispiel weiß ich, dass die ‚krigging‘ [wie übersetzt man dieses Adjektiv?] Ergebnisse von BEST nur maximal zu 0,49°C akkurat sind.
davidmhoffer
Das *Erste*, was ich bei einer solchen Graphik fragen würde ist, ob das berechnete Mittel überhaupt relevant ist. Da die Temperatur nicht linear mit Strahlung (W/m²) variiert, ist es möglich, unterschiedliche räumliche Temperaturverteilungen zu erhalten, die zwar identische Mittelwerte aufweisen, aber sehr unterschiedliche Energiebilanzen.
Beispiel: Zwei Punkte mit einer Temperatur von 280 K bzw. 320 K würden einen Mittelwert von 300 K zeigen und ein Strahlungsgleichgewicht von 471,5 W/m². Aber zwei Punkte mit einer Temperatur von 300 K an beiden Stellen würden ebenfalls einen Mittelwert von 300 K, jedoch ein Strahlungsgleichgewicht von 459,3 W/m² aufweisen.
Damit im Hinterkopf lassen die Fehlerbalken nicht nur keine Schlussfolgerung hinsichtlich des Temperaturtrends zu, d. h. ob er positiv oder nicht bedeutungslos ist, sondern die Fehlerbreite der Gleichgewichts-Energiebilanz ist wegen der nichtlinearen Beziehung zwischen beiden viel größer. Da AGW auf der Prämisse basiert, dass zunehmendes CO2 die Energiebilanz der Erde ändert, ist der Versuch, die Art und Weise zu quantifizieren, mit der es das tut, die Mittelung eines

Parameters, der keine direkte Beziehung zur Energiebilanz hat. Damit bleibt die Graphik zurück als selbst bedeutungslos hinsichtlich statistischer Genauigkeit und auch der Physik.
Antwort darauf:
Steven Mosher
Zitat: „Das *Erste*, was ich bei einer solchen Graphik fragen würde ist, ob das berechnete Mittel überhaupt relevant ist“.
Tatsächlich ist sie überhaupt kein Mittelwert von Temperaturen.
Obwohl die meisten Menschen, die der Klimadiskussion folgen, dies nicht verstehen (tatsächlich verstehen die Meisten, die solche Mittelwerte erstellen, das selbst nicht).
Was ist das globale Temperaturmittel wert, wenn es nicht wirklich ein Mittel ist.
Mathematisch handelt es sich um eine Vorhersage. Es ist eine Vorhersage dessen, was man an instrumentenfreien Standorten messen würde.
„Stationsbeobachtungen werden allgemein verwendet, um Klimavariablen auf Gitternetzen vorherzusagen (predict), wobei der statistische Terminus „Vorhersage“ (prediction) hier benutzt wird als „räumliche Interpolation“ oder „räumlich-zeitliche Interpolation“. Er sollte nicht mit „forecasting“ verwechselt werden*. Gründliche Begutachtungen von in Meteorologie und Klimatologie verwendeten Interpolationsverfahren wurden jüngst von Price et al. (2000), Jarvis und Stuart (2001), Tveito et al. (2006) und Stahl et al. (2006) vorgelegt. Die Literatur zeigt, dass die am meisten verwendeten Interpolationsverfahren in Meteorologie und Klimatologie folgende sind: Verfahren größter Nachbarschaft, Splines [?], Regression und Kriging [?]; aber auch neurale Netzwerke und *Machine learning techniques* [?].
*[*Ich glaube zwar, mich in der englischen Sprache ganz gut auszukennen, aber der hier angesprochene Bedeutungsunterschied zwischen den Wörtern „prediction“ und „forecast“ ist mir unbekannt. Vielleicht weiß jemand Näheres? Auch die am Ende dieses Absatzes stehenden und mit einem Fragezeichen versehenen Fachbegriffe sind mir unbekannt. Anm. d. Übers.]*
Aus: Räumlich-zeitliche Interpolation täglicher Temperaturwerte für globale Landgebiete mit einer Auflösung von 1 km. – Autoren: Milan Kilibarda1,*, Tomislav Hengl2, Gerard B. M. Heuvelink3, Benedikt Gräler4, Edzer Pebesma4, Melita Perčec Tadić5 and Branislav Bajat1.
Wenn man also liest, dass das globale Mittel für Dezember 2014 15,34°C beträgt, bedeutet das Folgendes:
Wenn man mit einem perfekten Thermometer an zufälligen Stellen des Globus‘ misst, wird eine Schätzung von 15,34°C den Fehler minimieren.
Wählt man 1000 zufällige Stellen ohne Thermometer aus, minimiert eine Vorhersage (prediction) den Fehler.
Antwort hierzu:
Danny Thomas
Steven,
danke für jene Beschreibung. Um mit BEST so fair wie möglich zu sein – es scheint, als bieten sie mit einiger Sicherheit eine vernünftige Evaluierung ihrer Analyse für 2014 an als eines der fünf wärmsten Jahre.
Kip Hansen sagte, dass wir auf der Diskussion der „Unsicherheits-

Bandbreiten“ bestehen. Um dies so weit wie möglich zu achten, wären Sie willens zu diskutieren, wie die Entscheidung gefallen ist, das Arbeitsergebnis zu titulieren mit „Die Mittlere Temperatur 2014 von Berkeley Earth“, wenn in Wirklichkeit diese Arbeit prädiktiver Natur ist und nicht das enthält, was der Titel nahelegt? Mit anderen Worten, warum lautet der Titel nicht „das prädiktive Mittel…“ (hier).
MET scheint eine vernünftigere Beschreibung ihrer Arbeit und des Vertrauensniveaus abzugeben. NOAA und NASA aber nicht so sehr, wenn selbst die oberflächlichste Sicht auf ihr Vertrauensniveau entweder zu einem klaren Plan der Irreführung führt im Vergleich zu ihren Schlagzeilen oder zur Befürwortung von AGW-Propaganda trotz fehlender guter wissenschaftlicher Kommentare.
bones
Vor der neuen Mathematik war es gewöhnlich der Fall, dass das arithmetische Mittel auch die Kleinste-Quadrate-Best-Estimate war, welche zufällig verteilte Fehler minimieren würde. Man kann glauben, dass Verfahren unter Verwendung von Kriging, neuralen Netzwerken und *machine learning techniques* ein Ergebnis zeitigen, das Fehler minimiert, aber man erwarte nicht, dass ich diesen Mist glaube.
davidmhoffer
Zitat: „Tatsächlich ist sie überhaupt kein Mittelwert von Temperaturen. Obwohl die meisten Menschen, die der Klimadiskussion folgen, dies nicht verstehen (tatsächlich verstehen die Meisten, die solche Mittelwerte erstellen, das selbst nicht).“
Ich verstehe Ihren Punkt und stimme dem auch bis zu einem gewissen Grad zu, obwohl Ihre Vermutung amüsant finde, dass selbst die die Berechnungen durchführenden Menschen nicht verstehen, was das ist. Damit ändert Ihr Standpunkt aber nicht meinen. Mein Punkt ist nicht, wie man einen Mittelwert für einen bestimmten Punkt mit der Zeit berechnet, sondern was die *Änderung* in jenem Mittelwert impliziert. Man nenne es ein Mittel, man nenne es eine Vorhersage (prediction) einer zufälligen Messung – da jene Messung sich mit der Zeit ändert infolge der nichtlinearen Beziehung zwischen Temperatur und Strahlung, ist die berechnete Änderung sogar noch bedeutungsloser als die Fehlerbalken suggerieren. Die Änderung des Wertes kann nicht die Änderung der Energiebilanz repräsentieren, weil einfache Physik vorschreibt, dass kalte Temperatur-Regimes (Nacht, Winter, hohe Breiten, große Höhe) überrepräsentiert und warme Temperatur-Regimes (Tag, Sommer, niedrige Breiten, geringe Höhe) unterrepräsentiert sind.
Der Rohwert der von Ihnen illustrierten Vorhersage (prediction) ist das Eine; die *Änderung* jenes Wertes aber etwas Anderes. Jene Änderung hat keine direkte Beziehung zu der interessierenden Metrik (Änderung des Energie-Gleichgewichtes), egal wie Sie es definieren.
Walt D.
Nehmen wir einmal die Antarktis als ein Gebiet mit sehr wenigen Daten. Sie behaupten, dass eine mittlere Temperatur von 15,34°C den Schätzfehler minimiert. Höchst unwahrscheinlich besonders im antarktischen Winter. Es ist absurd zu zeigen, dass die Temperaturen am Südpol und im Death Valley als zufällige Werte angesehen werden können,

gewonnen aus der gleichen Verteilung mit dem gleichen Mittel und den gleichen Varianzen. Darum arbeiten die Menschen mit Änderungen der Temperatur und nicht mit den tatsächlichen Temperaturen selbst.
Jeff F
@bones: Tatsache ist, dass die Verwendung von Daten, die mit den zuvor erwähnten Verfahren eingehen, nicht wirklich neue Mathematik ist. Die Verfahren wurden schon geraume Zeit angewendet. Allerdings ist das Problem hier, dass viele Leute diese anwenden, ohne deren Implikationen zu verstehen. Mit vielen dieser Verfahren KANN NICHT gezeigt werden, dass sie Fehler minimieren – außer unter ganz besonderen Umständen. Zum Beispiel ist Kriging nur eine unverzerrte Schätzfunktion, falls der Prozess stationär ist. Schlimmer noch ist, dass es wirklich keine Rechtfertigung dafür gibt, Klimadaten überhaupt als einen stochastischen Prozess zu behandeln.
In anderen Disziplinen, wo wir diese Verfahren für die Datenanalyse anwenden, zeigen wir Beispiele, wo sie funktioniert haben, und überlassen es dem User zu entscheiden, ob die Verfahren für sein Problem geeignet sind. Aber wir stellen KEINE Behauptungen auf über die Optimalität des Verfahrens. Weil wir wissen, dass wir das nicht können. Allerdings habe ich VIEL zu viele klimawissenschaftliche Studien gelesen, die weiter entwickelte Verfahren allein zu dem Zweck angewendet haben, um eine Behauptung über den zugrunde liegenden PROZESS zu rechtfertigen, der die Daten erzeugt hat! Falls wenige Behauptungen über die statistischen Charakteristika der DATEN selbst erhoben werden können, wenn man diese Verfahren anwendet, kann man fast nichts über den PROZESS aus diesen Daten aussagen.
IvanV
@ Steven Mosher & davidmhoffer
Ich versuche gerade, Ihre Punkte nachzuvollziehen und die Unterschiede zwischen ihnen.
Steven Mosher, was Sie sagen, ähnelt einem Fall, wo man zwei Stationen in den Tropen hat mit einem Mittel von 30°C und eine nahe den Polen mit einem Mittel von 0°C. Dann würde die Best Estimate der globalen mittleren Temperatur 15°C betragen, da dies den Fehler zwischen den Messungen minimieren würde?
Davidmhoffer, aus Sicht des Energiegleichgewichtes wäre die Best Estimate der globalen Mitteltemperatur die Temperatur einer Sphäre mit gleichmäßiger Temperatur, die das gleiche Energie-Gesamtgleichgewicht aufweist?
ferdberple
Mathematisch ist es eine Vorhersage (prediction). Es ist eine Vorhersage dessen, was man an nicht mit Instrumenten ausgestatteten Stellen messen würde.
Auf dieser Grundlage wird die Stichproben-Theorie einen realistischeren Wert ergeben als die gegenwärtige Praxis, Stationen zu adjustieren, damit sie statisch aussehen.
Nehmen wir einfach mal an, dass jede Stationsablesung einmalig ist. Dass die Station selbst verlagert werden kann oder anderweitig von einer Messung zur nächsten verändert wird, und dass jeder Versuch von

Adjustierungen, um eine kontinuierliche Stationsaufzeichnung zu erhalten einfach noch mehr unbekannte Fehler einführen wird.
Es gibt keine Notwendigkeit. Da man den Wert an unbekannten Punkten vorhersagt (predict), wird eine Stichprobe an bekannten Punkten ausreichen, während die Möglichkeit eingeführter Fehler eliminiert wird. Alles, was erforderlich ist, ist ein Stichproben-Algorithmus, der zu der räumlichen und zeitlichen Verteilung der Erdoberfläche passt.
Und so weiter. Es gibt über 300 weitere Kommentare zu dem Artikel. Als Grundlage für das hier besprochene Problem soll die Übersetzung bis hier aber erst einmal reichen.
Link:
http://wattsupwiththat.com/2015/02/01/uncertainty-ranges-error-bars-and-cis/
Übersetzt von Chris Frey EIKE