Das Wirrwarr von Klimamodellen

geschrieben von Chris Frey | 4. Juli 2020

Ein Gedankenexperiment

Hier ist das Gedankenexperiment: Wir wollen eine Verbindung herstellen, die irgendwie Farbe erzeugt (der Mechanismus, wie sie das macht, ist nicht wirklich relevant). Wir wollen jedoch speziell eine gut definierte Farbe, die von der Anwendung, für die sie verwendet werden soll, vorgeschrieben wird. Sagen wir einen Türkis-Ton.

Nun haben unsere Geologen und Chemiekollegen einige Mineralien und Verbindungen vorgeschlagen, die Kandidaten für unser buntes Unternehmen sein könnten. Leider gibt es keinerlei Informationen darüber, welche Farben diese Stoffe produzieren. Dieser Umstand wird noch dadurch verstärkt, dass die Minerale äußerst selten und daher extrem teuer sind, während synthetische Minerale wirklich schwierig herzustellen und daher noch teurer sind. Wie gehen wir also vor, wie finden wir die besten Verbindungen, die wir ausprobieren können? Es kommt nicht in Frage, von jeder der vielen Verbindungen ein Muster zu bekommen und jede von ihnen auf die Farbe zu testen, die sie erzeugt. Was wir also tun, ist, die Physik des Farbe erzeugenden Prozesses für jede der vorgeschlagenen Verbindungen zu modellieren, um diejenigen zu finden, die Türkis erzeugen, falls es welche gibt. Das klingt einfach genug, ist es aber nicht, weil es mehrere verschiedene Codes gibt, nämlich insgesamt 5, die vorgeben, eine solche Simulation durchzuführen, jeder mit seinen eigenen zugrunde liegenden Annahmen und Eigenheiten. Wir führen diese Codes für die vorgeschlagenen Verbindungen durch und stellen fest, dass die Farben, die sie prognostizieren, leider für einzelne Verbindungen und im Allgemeinen überall uneinheitlich sind.

Nehmen Sie zum Beispiel die Verbindung Novelium1. Die vorhergesagten Farben reichen von Gelbgrün bis zu tiefem Violett mit einigen wenigen dazwischen wie Grün, Blau oder Ultramarin, ein Frequenzbereich mit dem Faktor 1,3; ähnlich für die anderen Kandidaten. In dieser Situation ist der einzige Weg vorwärts ein Experiment. Also graben wir tief in das Budget hinein und besorgen uns eine Probe von Novelium1, um zu sehen, welche Farbe es tatsächlich produziert. Es stellt sich als orange-rot heraus, was ziemlich enttäuschend ist. Wir sind wieder da, wo wir angefangen haben. Und wegen unserer Haushaltsbeschränkungen sind wir an dem Punkt, an dem wir aufgeben müssen.

Hier wollen wir ein Mitglied unseres Teams vorstellen. Nennen wir es Mike. Mike ist ein bisschen aufdringlich, weil ihm völlig klar ist, dass wir, wenn wir unser Ziel erreichen sollten, den einen oder anderen prestigeträchtigen Preis bekommen würden, etwas, worauf er ziemlich scharf ist. Er schlägt Folgendes vor: Wir nehmen das Modell, das die Farbe vorhergesagt hat, die der tatsächlichen Farbe am nächsten kommt,

das ist das Modell, das uns Gelb-Grün gegeben hat, und passen seine Parameter so an, dass es stattdessen Orange-Rot vorhersagt. Das ist nicht allzu schwierig, und nach ein paar Tagen, in denen er auf einer Tastatur herumtüftelt, kommt er mit einem angepassten Modell, das die beobachtete Farbe erzeugt. Geschicklichkeit rundherum, bis auf ein oder zwei weitere skeptische Teammitglieder, die darauf bestehen, dass das neue Modell validiert werden muss, indem es die Farbe der Verbindung Novelium2 korrekt vorhersagen lässt. Da der Preis darauf beruht, ist dies eindeutig ein Muss, also kratzen wir den Boden des Budgets ab und wiederholen die Übung für Novelium2. Das angepasste Modell sagt Gelb voraus. Das Experiment ergibt Orange.

Wir gaben auf.

Was bedeutet das nun?

Können wir irgendetwas Nützliches hieraus ableiten? Um das herauszufinden, müssen wir drei Fragen beantworten.

Erstens: Was wissen wir nach der ersten Phase des Projekts, der Modellierungsübung, bevor wir das Experiment durchführten? Bedauerlicherweise lautet die Antwort: nichts Nützliches. Bei 5 verschiedenen Ergebnissen wissen wir nur mit Sicherheit, dass *mindestens 4 der Modelle falsch sind*, aber nicht welche. Selbst wenn die von uns gewünschte Farbe (Türkis) auftaucht, wissen wir immer noch nichts. Denn wie kann man sicher sein, dass der Code, der ihn erzeugt, das „richtige Ergebnis“ ist, wenn man die Ergebnisse der *a priori* gleichwertigen anderen Modelle berücksichtigt? Das kann man nicht. Wenn ein Modell uns Türkis gab, könnte es einfach ein glücklicher Zufall sein, wenn das Modell selbst noch fehlerhaft ist. Allein die Tatsache, dass die Modelle sehr unterschiedliche Ergebnisse liefern, sagt uns daher, dass *höchstwahrscheinlich alle Modelle falsch sind*. Tatsächlich ist es sogar noch schlimmer: *Wir können nicht einmal sicher sein, dass die von Novelium1 erzeugte Echtfarbe im Bereich von Gelbgrün bis Violett liegt*, selbst wenn es ein Modell gäbe, das die von uns gewünschte Farbe erzeugt. Im Anhang gebe ich eine einfache, auf Wahrscheinlichkeit basierende Analyse zur Unterstützung dieses und der folgenden Punkte.

Zweitens, was wissen wir nach dem unerwarteten Ausgang des eigentlichen Experiments? Wir wissen nur mit Sicherheit, dass alle Modelle falsch sind (und dass es nicht die gesuchte Verbindung ist).

Drittens: Warum ist Mikes kleiner Trick so kläglich gescheitert? Was ist da passiert? Die Parametereinstellung des ursprünglichen, ungefilterten Modells verkörpert das beste Verständnis der Physik – von seinen Machern, wenn auch unvollständig, aber das ist nicht wirklich relevant –, die ihm zugrunde liegt. Durch Modifizierung dieser Parameter wird dieses Verständnis verwässert, und wenn die „Feinabstimmung“ weit genug geht, verschwindet es vollständig, so wie die Grinsekatze verschwindet, je mehr man sie ansieht. Ein solches Modell im Nachhinein so zu

optimieren, dass es an die Beobachtungen angepasst werden kann, ist daher gleichbedeutend mit dem Verzicht auf die Behauptung, dass man die dem Modell zugrundeliegende relevante Physik versteht. Jeglicher Vorwand, das Thema wirklich zu verstehen, verschwindet aus dem Fenster. Und mit ihm geht jede Vorhersagekraft, die das ursprüngliche Modell gehabt haben könnte. Ihr Modell ist gerade zu einer weiteren sehr komplexen Funktion geworden, die in einen Datensatz eingepasst wurde. Bekannt ist der Ausspruch des Mathematikers und Physikers John von Neumann über eine solche Praxis: „Mit vier Parametern kann ich einen Elefanten anpassen, und mit fünf kann ich ihn dazu bringen, mit dem Rüssel zu wackeln“. Bei dem optimierten Modell handelt es sich wahrscheinlich um ein neues, falsches Modell, das zufällig eine Übereinstimmung mit den Daten ergab.

Eine Anwendung auf Klimamodelle

Bewaffnet mit den Erkenntnissen aus der vorstehenden Geschichte sind wir nun in der Lage, einige grundsätzliche Aussagen zu den IPCC-Klimamodellen zu machen, zum Beispiel zu der Gruppe der 31 Modelle, die das CIMP6-Ensemble bilden (Eyring et al., 2019; Zelinka et al., 2020). Die interessierende Größe ist der Wert der Gleichgewichts-Klimasensitivität [Equilibrium Climate Sensitivity (ECS)], die erwartete langfristige Erwärmung nach einer Verdoppelung der atmosphärischen CO_2-Konzentrationen. Die vorhergesagten ECS-Werte im Ensemble umfassen einen Bereich von 1,8°C am unteren Ende bis 5,6°C am oberen Ende, ein kolossaler Faktor 3 im Bereich, der von den 31 Modellen mehr oder weniger gleichmäßig belegt wird. Die Natur mag listig, ja sogar hinterhältig sein, aber sie ist nicht bösartig. Es gibt nur einen „wahren“ ECS-Wert, der der Verdoppelung der CO_2-Konzentration in der realen Welt entspricht.

Können wir eine Erklärung zu diesem Ensemble abgeben? Nur diese beiden Beobachtungen:
Erstens, höchstwahrscheinlich sind all diese Modelle falsch. Diese Schlussfolgerung ergibt sich logisch aus der Tatsache, dass es viele a priori gleich gültige Modelle gibt, die nicht gleichzeitig richtig sein können. Höchstens eines dieser Modelle kann richtig sein, aber angesichts der übrigen 30 falschen Modelle stehen die Chancen schlecht, dass überhaupt ein Modell richtig ist. Tatsächlich lässt sich zeigen, dass die Wahrscheinlichkeit, dass keines der Modelle korrekt ist, bis zu 0,6 betragen kann.

Zweitens können wir nicht einmal sicher sein, dass die wahre ECS im Bereich der von den Modellen erfassten ECS-Werte liegt. Die Wahrscheinlichkeit, dass dies der Fall ist, liegt bei 1,0-0,6=0,4, was bedeutet, dass die Wahrscheinlichkeit, dass die wahre ECS in dem von den Modellen abgedeckten Bereich liegt, etwa 2 bis 3 beträgt (und damit die Wahrscheinlichkeit, dass die wahre ECS außerhalb des Bereichs liegt). Die oft gemachte Annahme, dass der „wahre“ ECS-Wert irgendwo im Bereich der Ergebnisse der Modelle im Ensemble liegen muss, basiert auf einem

logischen Trugschluss. Wir haben absolut keine Ahnung, wo das „wahre“ Modell – Nummer 32, das „Experiment“ – innerhalb oder außerhalb der Spanne landen würde.

Es sind einige Einschränkungen zu machen. Was bedeutet es zum Beispiel: Das Modell ist „falsch“? Es bedeutet, dass es unvollständig sein könnte – es fehlen Konzepte oder Prinzipien, die in ihm vorhanden sein sollten – oder umgekehrt überkomplett – mit Dingen, die vorhanden sind, aber nicht vorhanden sein sollten – oder dass es Aspekte enthält, die einfach falsch oder falsch kodiert sind, oder alles davon. Da viele Ensemble-Modelle ähnliche oder sogar identische Elemente haben, könnte man ferner argumentieren, dass die Ergebnisse der Ensemble-Modelle nicht unabhängig sind, dass sie korreliert sind. Das bedeutet, dass man die „effektive Anzahl“ N der unabhängigen Modelle berücksichtigen sollte. Wenn N = 1 wäre, würde dies bedeuten, dass alle Modelle im Wesentlichen identisch sind, wobei der Bereich von 1,8°C bis 5,6°C ein Hinweis auf den intrinsischen Fehler wäre (was eine ziemlich schlechte Darstellung wäre). Wahrscheinlicher ist, dass N irgendwo im Bereich von 3 bis 7 liegt – mit einer intrinsischen Streuung von z.B. 0,5°C für ein einzelnes Modell – und wir sind wieder bei dem hypothetischen Beispiel oben.

Die Wahrscheinlichkeit von etwa 3 zu 2, dass keines der Modelle richtig ist, dürfte politisch interessant sein. Würden Sie mit diesen Quoten viel von Ihrem hart verdienten Geld auf ein Pferd setzen? Ist es klug, die Energieversorgung und damit die gesamte Wirtschaft Ihres Landes auf solche Quoten zu setzen?

Nachhersage

Ein anonymer Rezensent einer meiner früheren Schriften gab diesen offenen Kommentar ab, und ich zitiere:

Die Erfolgsbilanz der GCM’s ist insofern enttäuschend, als sie nicht in der Lage waren, den beobachteten Temperatur-Stillstand nach dem Jahr 2000 vorherzusagen, und auch nicht vorhersagen konnten, dass die Tropopausentemperaturen in den letzten 30 Jahren nicht gestiegen sind. Das Versagen der GCMs ist nicht auf Fehlverhalten zurückzuführen, aber die Modellierung des Erdklimas ist eine große Herausforderung.

Der wahre Wissenschaftler weiß, dass Klimamodelle noch in Arbeit sind. Der Pseudowissenschaftler, der unter dem Druck steht, die „Vorhersagen“ zu halten, muss einen Weg finden, um die Modelle und die Temperaturdaten der realen Welt „in Einklang zu bringen“.

Eine Möglichkeit dazu besteht darin, die Temperaturdaten in einem Prozess zu massieren, der „Homogenisierung“ genannt wird (z.B. Karl et al., 2015). Auf wundersame Weise verschwindet der „Stillstand“. Ein merkwürdiger Aspekt einer solchen „Homogenisierung“ ist, dass bei jeder Anwendung die „angepassten“ vergangenen Temperaturen immer niedriger sind, wodurch die angeblich „künstliche Erwärmung“ größer wird. Niemals

andersherum. Offensichtlich kann man diesen kleinen Handgriff nur einmal ausführen, vielleicht zweimal, wenn niemand zusieht. Aber danach wird selbst der Dorftrottel verstehen, dass er reingelegt wurde, und die „Homogenisierung“ in den gleichen Mülleimer der Geschichte werfen wie Lysenkos „Vernalisierung“.

Der andere Weg ist die Anpassung der Modellparameter an die Beobachtungen (z.B. Hausfather et al., 2019). Angesichts der vielen anpassbaren Parameter und unter Berücksichtigung von von Neumans Witz ist es nicht überraschend, dass die Modelle durch eine solche „Nachhersage“ recht gut an die Daten angepasst werden können. In der kriecherischen Mainstream-Presse herrscht überall Eifer, mit manchmal urkomischen Ergebnissen. So verkündete beispielsweise ein Korrespondent einer überregionalen niederländischen Zeitung enthusiastisch, dass die Modelle die Temperaturen der letzten 50 Jahre korrekt vorhergesagt hätten. Dies wäre wirklich eine bemerkenswerte Leistung, denn die früheste Software, die als „Klimamodell“ betrachtet werden kann, stammt aus den frühen 1980er Jahren. Eine interessantere Frage ist jedoch: Können wir erwarten, dass ein derart angepasstes Modell eine Vorhersagekraft hat, insbesondere im Hinblick auf die Zukunft? Die Antwort ist ein schallendes „Nein“.

Sind Klimamodelle nutzlos?

Natürlich nicht. Sie können sehr nützlich sein als Werkzeuge zur Erforschung jener Aspekte der Physik der Atmosphäre und des Klimasystems, die noch nicht verstanden werden oder deren Existenz noch nicht bekannt ist. **Wofür man sie nicht verwenden kann, sind Vorhersagen**.

[Hervorhebung vom Übersetzer]

References:
Eyring V. *et al.* Nature Climate Change, **9**, 727 (2019)
Zelinka M. *et al. Geophysical Research Letters,* **47** (2020)
Karl T.R., Arguez A. *et al.* Science **348,** 1469 (2015)
Hausfather Z., Drake H.F. *et al.* Geophysical Letters, **46** (2019)
[Es folgt eine interessante Beschreibung der Bestimmung der Wahrscheinlichkeit, mit der ein bestimmtes Modell korrekt ist. Das ist das tägliche Brot eines Wetter-Prognostikers. Ein Faktor wird hierbei aber nicht erwähnt: Die Erfahrung eines seit Jahrzehnten praktizierenden Wetter-Prognostikers ist ein sehr gewichtiger statistischer Faktor, ist doch Erfahrung nichts anderes als Statistik. Dennoch, kein seriöser Prognostiker würde eine Prophezeiung, die über höchstens eine Woche hinausgeht, als Prognose bezeichnen – und sie schon gar nicht zu einer Grundlage schwer wiegender Entscheidungen machen.
Diese allgemein theoretische Betrachtung wird hier nicht mit übersetzt. Anm. d. Übers.]

Addendum: an analysis of probabilities
First the case of 5 models of which at most 1 can possibly be right.

What is the probability that none of the models are correct? All models are *a priori* equally valid. We know that 4 of the models are not correct, so we know at once that the probability of any model being incorrect is at least 0.8. The remaining model may or may not be correct and in the absence of any further information both possibilities could be equally likely. Thus, the expectation is that, as a matter of speaking, half a model (of 5) is correct, which means the *a priori* probability of any model being incorrect is 0.9. For N models it is 1.0-0.5/N. The probability that all models fail then becomes: F=(1-0.5/N)^N which is about 0.6 (for N > 3). This gives us odds of 3 to 2 that none of the models are correct and it is more likely that none of the models are correct than that one of them is. (If we had taken F=(1-1/N)^N the numbers are about 0.34 with odds of 1 to 2)

Now an altogether different question. Suppose one of the models does give us the correct experimental result, what is the *a posteriori* probability that this model is indeed correct, given the results of the other models? Or, alternatively, that the model is incorrect even when it gives the 'right' result (by coincidence)? This posterior probability can be calculated using Bayes' theorem,

$$P(X|Y) = P(Y|X)*P(X)/P(Y),$$

where P(X|Y) stands for the probability of X given Y and P(X) and P(Y) are prior probabilities for X and Y. In this case, X stands for 'the model is incorrect' and Y for 'the result is correct', in abbreviated form M=false, R=true. So the theorem tells us:

$$P(M=false|R=true) = P(R=true|M=false) * P(M=false) / P(R=true)$$

On the right-hand side the first term denotes the false-positive rate of the models, the second term is the probability that the model is incorrect and the third is the average probability that the result predicted is accurate. Of these we already know P(M=false)=0.9 (for 5 models). In order to get a handle on the other two, the 'priors', consider this results table:

Model	Outcome		Rate-1	Rate-2	Rate-3	Rate-4	Rate-5	Rate-6	
True	True	TT	1.0	1.0	1.0	1.0	1.0	1.0	
True	False	TF	0	0	0	0	0	0	
False	True	FT	0.8	0.5	0.2	0.1	0.05	0.01	
False	False	FF	0.2	0.5	0.8	0.9	0.95	0.99	
			0.82	0.55	0.28	0.19	0.145	0.109	P(R=true)
			0.87	0.82	0.64	0.47	0.31	0.08	P(M=f|R=t)

The 'rate' columns represent a number of possible ensembles of models differing in the badness of the incorrect models. The first lot still give relatively accurate results (incorrect models that often return the about correct result, but not always; pretty unrealistic). The last with seriously poor models which on occasion give correct results (by happy coincidence) and a number of cases in between. Obviously, if a model is correct there is no false-negative (TF) rate. The false-positive rate is given by P(R=true|M=false) = FT. The average true result expected is given by 0.1*TT + 0.9*FT = 0.82 for the first group, 0.55 for the second

and so on.
With these priors Bayes’ Theorem gives these posterior probabilities that the model is incorrect even if the result is right: 0.87, 0.82 etc. Even for seriously poor models with only a 5% false positive rate (the 5th set) the odds that a correct result was made by an incorrect model are still 1 to 2. Only if the false positive rate (of the incorrect models) drops dramatically (last column) can we conclude that a model that produces the experimental result is likely to be correct. This circumstance is purely due to the presence of the incorrect models in the ensemble. Such examples shows that in an ensemble with many invalid models the posterior likelihood of the correctness of a possibly correct model can be substantially diluted
Link: https://wattsupwiththat.com/2020/07/01/the-climate-model-muddle/
Bearbeitet von Chris Frey EIKE