LIFE SAFETY AND SECURITY

open access

Health Safety

Mini review Paper

IMPERFECT DATA AND CURRENT CLIMATE SCENARIOS: THE EFFECT OF SAMPLING, STANDARDIZATION AND CATALOGING ON THE RELIABILITY AND ACCURACY OF PREDICTIVE MODELS

Antonio Gallo, Gaetano La Bella

CORRESPONDANCE: Antonio Gallo
e-mail: antoniogallo@iemest.eu
Phone number: +39091333913

Received: May 16th, 2024
Revised: June 21th 2024
Accepted: June 28th, 2024

Premise

The accuracy of climate datasets is essential for understanding climate change and making future predictions, one of the most pressing and complex environmental challenges facing the scientific community in the 21st century.

While the Earth continues to naturally change its "breathing", as widely documented in paleoclimatic studies, it is imperative to have climate datasets that are not only complete in the sequence of data and in their territorial representation, but that are representative of the entire Earth's surface, are coherent and representative of reality.

Climate datasets play a crucial role not only in representing, monitoring and quantifying current and past climate changes, but also as key inputs for climate models used to forecast future trends. These forecasts underpin governments' strategic decisions in planning and implementing economic policies aimed at mitigating climate change.

Internationally renowned organizations such as NASA GISTEMP, NOAA, HadCRUT and Berkeley Earth have collected, reorganized and made available global temperature datasets, constantly updated and reconstructed over a time series that extends for about 200 years. These datasets represent a fundamental resource for climatologists, who use them to define the global temperature curve and identify the evolutionary trends of the climate at the planetary level.

Despite their relevance, the analysis of climate datasets highlights discrepancies, sometimes significant, in the results obtained. These differences mainly derive from the different mathematical methodologies adopted for the homogenization and filtering of the data, necessary to guarantee continuity and representativeness over larger areas than the original acquisition locations.

These discrepancies, present differently in each dataset, significantly affect the reconstruction of the curves that describe the trend of global warming. This impact is reflected not only on the precision of climate forecasting models, but also on mitigation and adaptation strategies, with significant repercussions on the community, unaware of the implications of such uncertainties.

This article aims to highlight how the process of homogenization of climate data, although considered essential to ensure the coherence of temperature time series, may have introduced, occasionally or accidentally, errors in the estimation of climate parameters. Such errors, in fact, have contributed to the formulation and diffusion of climate models potentially distorted with respect to reality, compromising their accuracy and reliability.

Starting with an in-depth analysis of the specific techniques used by the Institutes cited for the measurement and homogenization of climate data, this work aims to identify and describe, with accuracy, the anomalies generated and present in global temperature

datasets. The goal is to bring these aspects to the attention of the scientific community, proposing a moment of shared reflection that can define the guidelines and scientific tools applicable for a coherent and critical review of the methodologies adopted so far.
We are fully aware of the complexity that characterizes climate systems and, even more, of the crucial importance of having homogeneous, accurate and complete datasets. Only through solid and reliable datasets will it be possible to support, with coherence and concreteness, any political and scientific decisions related to the changes in progress.
Furthermore, this proposal will allow for the development of future projections based on renewed, unbiased models, capable of providing a more reliable and useful vision to effectively address the global challenges related to climate change.
The value of this work lies precisely in the invitation to the entire scientific community to join in a constructive discussion on a topic of such great global relevance. This discussion should go beyond any economic or political interest, and should place lucid, rigorous and impartial scientific reasoning at the center of the debates.
The aim is to stimulate a critical review of the currently dominant climate models, which have been defined, as described, on datasets whose quality presents margins of uncertainty. This review would represent a necessary step to restore solidity and transparency to climate research, offering more adequate tools to interpret the present and correctly look at the future of our planet.

The Critical Role of Scientific Data Accuracy

The accuracy of climate datasets, according to scientific recommendations on the criteria for defining a scientific dataset[1], assumes a first phase process of scrupulous data collection, through a vast network of terrestrial, oceanic and satellite meteorological stations equipped with appropriate, valid and maintained sensors, distributed over a vast area and a second phase dedicated to their management and cataloguing.
However, in order for the sampled data to be considered reliable over time, it is necessary that each detection sensor is subject to periodic calibration and maintenance, in order to be able to acquire the data sought with an objective and rigorous precision, whose congruence and coherence can be demonstrated over time.
Sampling data from uncalibrated and unverified stations and with non-standardized procedures would produce uncertain and inhomogeneous datasets, inadequate for their scientific use and, even more so, for the definition of climate forecasting models.
It is therefore essential that the data are sampled using reliable instrumentation that provides appropriate measurement standards and that their mathematical treatment, for the sole purpose of cataloguing, is performed using non-invasive homogenisation methodologies, adequately documented and described in detail in a specific note accompanying the datasets.
Furthermore, it is considered convenient to distribute the sampled and untreated data together with the organized dataset, in such a way as to allow a personal evaluation, an appropriate choice of the logical treatment and, finally, a specific use for the definition of the forecasting model.
Subsequently, anyone, starting from a common dataset known to the scientific community in the mathematical criterion that gave rise to it, will be able to choose to apply scientific methodologies, even robust ones, for the identification, understanding and forecasting of long-term climate patterns.
The main international institutions that today collect, analyze and distribute global climate data are NASA GISTEMP, NOAA, HadCRUT and Berkeley Earth.
These institutions, from what is assumed from the consultation of the sequences, employ unequal methodologies both for the sampling of the data, both for their homogenization, and for their cataloguing. Specifically:

- **NASAGISTEMP**(Goddard Institute for Space Studies), is part of the Goddard Institute for SpaceStudies (GISS), a NASA laboratory based in New York.

The dataset consists of surface temperature data sampled from a network of 6,300 measuring stations distributed across the planet.

[1]The recommended criteria for defining a scientific data set according to international guidelines and practices are designed to ensure the quality, integrity, reproducibility and usefulness of the data. These criteria ensure that scientific data can be used for rigorous analysis and allow reliable conclusions to be drawn, increasing the credibility and impact of research. The main criteria concern:

1. Accuracy and Validity. Data must be used using scientifically valid and verifiable methods. Measurement instruments and collection protocols must be calibrated and certified. An analysis of the error or uncertainty associated with each data point must be included.

2. Completeness. Datasets must include all data necessary to represent the phenomenon studied without omissions that could introduce bias. The time and geographical series must be extended as much as possible to ensure the representativeness of the phenomenon.

3. Consistency. Data must be homogeneous and coherent over time and space, with attention to changes in methodology or instrumentation. Homogenization processes must be documented to avoid discontinuities or errors in interpretation.

4. Transparency and Traceability. Each step of the collection, homogenization and analysis process must be clearly documented. The provenance of the data (origin, sources, measurement tools) must be traceable and verifiable.

5. Accessibility and Openness. Data must be made available to the scientific community in standard and open formats, compatible with the main analysis platforms. The data license must allow responsible use and reuse for research purposes.

6. Reproducibility Other researchers must be able to reproduce the results using the same dataset and methodologies. Detailed metadata describing the collection methods, analysis procedures, and algorithms used must be provided.

7. Metadata Quality Each dataset must be accompanied by detailed metadata, which includes information on: collection and analysis protocols, origin and characteristics of the tools, any modifications or interventions on the data.

8. Periodic Update. Datasets should be updated regularly to incorporate new data or to improve existing data. Updated versions should be accompanied by documentation that clarifies the changes made.

9. Representativeness. The data must adequately represent the phenomenon of interest and must be generalizable, avoiding distortions due to non-representative samples.

10. Uncertainty Management. Uncertainties in data must be quantified, made explicit and, when possible, reduced by rigorous statistical methods. Correction and homogenization algorithms must be shared and validated.

11. Standardization. The data must be organized according to accepted international standards, such as those defined by bodies such as the World Meteorological Organization (WMO), ISO, or other industry authorities.

However, to ensure global data coverage by reducing the large gaps present in relation to the rather limited number of monitoring stations, the Institute has integrated the data collected by their stations with different types of other measurement sources, such as terrestrial meteorological stations, ocean buoys, data collected by ships, research stations located in Antarctica and, for several decades, satellite data.
Therefore, the NASA GISTEMP datasets are treated through analysis processes that allow to define complete datasets defined on advanced statistical methods based on corrections for bis and inconsistencies. These processes, in fact, correct and complete the information of the 6300 measurement stations, with mathematical artifices also of a certain importance, also defined to remove any distortions deriving from changes in the location of the meteorological stations, improvements in measurement technology, and other environmental ones such as the urban heat island effect, spatial interpolation, as well as to extend the presence of data in those sectors where there are no real data. The treatment of areal extension of the datasets is structured on different interpolation techniques that allow to extend the existing information to a homogeneous global grid, through a temporal normalization process.
Finally, mathematical adjustments are made to ensure that the data are comparable over time and space, harmonizing the data series collected from different sources and correcting for non-climatic discontinuities due to sources that present local singularities.

- **NOAA**(National Oceanic and Atmospheric Administration)is an American federal agency operating under the United States Department of Commerce that uses a large network of data collection instruments to monitor global climate conditions, which provide data on air temperature at ground level, as well as ocean buoys located in different parts of the oceans, which allow the sampling of temperature data at sea, and also satellite data that provide real-time global coverage, especially of less accessible areas.

In order to ensure the reliability of the data, the Agency has endeavoured to use rigorous methods to try to ensure the reliability of its climate data, such as data quality corrections, carried out to correct measurement errors and biases due to factors such as urbanisation, advanced interpolation techniques to ensure a homogeneous representation of global temperature in those areas without measuring stations and normalisation techniques to "adjust" data from different sources and periods, in such a way as to have a consistency of the time series.

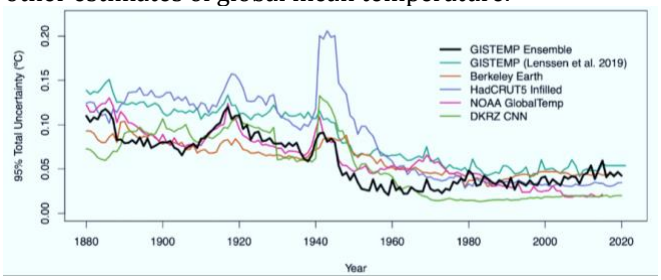
- **HadCRUT**(Hadley Centre/Temperature Climate Research Unit), a collaboration between the UK Met Office's Hadley Centre and the Climatic Research Unit (CRU) at the University of East Anglia, both based in the UK.

The data collection system is structured in land-based weather stations and sea surface temperature data, which are also limited in number.
The data analysis is conducted through data interpolations, which unlike other datasets, are not conducted on large areas without data, defining datasets that are incomplete for many areas but more truthful, in addition to the adjustments of inconsistencies applied to standardize the data over time.

- **Berkeley Earth**, an independent non-profit organization based in Californiawhich integrates data from multiple meteorological stations, through mathematical operations to correct errors and inconsistencies present in historical datasets, in order to homogenize the datasets provided, providing for the integration of the data necessary for the completion of the series from surveys taken from other sources, in order to extend the information of the temperature data over a vast terrestrial surface.

This Organization also proceeds to the analysis of datasets through advanced corrections based on sophisticated statistical analysis techniques necessary to correct biases in the data, including those due to changes in measurement sites and urban influence, in addition to extensive interpolation techniques, to extend the information in areas not covered by surveys.

The result of the above is highlighted below, in the comparison made by Lenssen et al. (2024) between the 95% confidence intervals for the global annual temperature anomaly calculated by GISTEMP and various other estimates of global mean temperature.

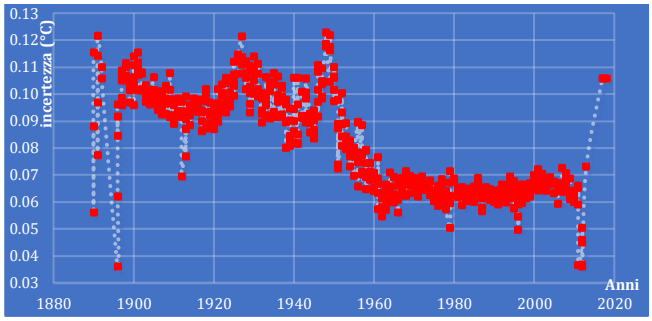


Comparison of 95% confidence intervals for the global annual temperature anomaly calculated from GISTEMP and various other estimates of global mean temperature (Lenssen et al., 2024)

Analysis of climate curves

Global climate datasets released by NASA GISTEMP, NOAA, HadCRUT and Berkeley Earth play a fundamental role in the field of climate forecasting, as they constitute the input for the definition of climate models that form the basis of current international political decisions.
As seen, these datasets, having to compensate for the multiple inequalities due both to the sampling systems and to their distribution and precision, are subject to composite mathematical treatments, consisting of data refinement processes necessary for the definition and production of complete series of data, usable on a global scale.

Paradoxically, the same Institutions confirm the presence of “fragility” of the climatological datasets through their continuous commitment to carry out scientific research to make the series more truthful and to be able to quantify their intrinsic uncertainty.
This aspect raises considerable questions regarding the accuracy and restitution of climate forecasting models "today so celebrated as absolute and incontestable truth", since the very input climate datasets that have allowed the generation of climate models, inevitably incorporate inaccuracies and errors that are difficult to identify in environmental parameters. Climate models that show, with extreme certainty, how anthropogenic forcings, including killer CO2, are the only ones responsible, from 1900 to today, for the increase of 1.5° C in global temperature and that if not severely contrasted could lead to increases of 4° C by the end of 2100, downgrading or completely suppressing the effect of natural forcings, such as variations in solar and astronomical cycles.
It is therefore necessary and proper to realize that the climate datasets, which have constrained the climate forecasting models, are scientifically imprecise and misleading.
Precisely for this reason, we proceeded to analyze and mathematically evaluate the discrepancies present between the climatological curves of the correct graph proposed by Lenssen et al. (2024), in which the 95% confidence intervals of the anomalies of the annual global temperature present in the datasets proposed by NASA GISTEMP, NOAA, HadCRUT and Berkeley Earth and by Lenssen et al. (2019) itself with the adjustments made on the GISTEMP dataset are shown.
Through a data digitization process based on automated visual analysis techniques, it was possible to extract and define the pair of coordinates that characterizes the mathematical function of the individual proposed curves. The definition of the wave functions of the curves thus generated has allowed the detailed mathematical analysis of the existing discrepancies and has allowed the construction of a new graph in which these discrepancies can be better highlighted, through their emphasis produced with the data of the scaled years, represented on the abscissa axis and those relating to the uncertainties in °C, on the ordinate axis. Substantially, through the "linear transformation" we proceeded to convert the coordinates extracted from the original curves into the real values of the axes (years ⇄ X axis and uncertainties in °C ⇄ on the Y axis).
Specifically, this representation was used to show the trend of uncertainties on global temperature anomalies (in °C) over time, from 1880 to 2020, by calculating the average value for each year through the different data series represented in the Lenssen graph.



The analysis of the discrepancies was carried out through an initial grouping of the coordinates scaled by year, of which the average of the uncertainties was calculated for each year. This step served to obtain a single representative value for each year that represents the red node of the curve.
The graph reflects scientific and technological progress in understanding global temperatures, but also the inherent difficulties of comparing different estimates. The main discrepancies are evident in peak periods (such as the 1940s) and transition phases.
Specifically, from the main observations it is evident that the highest uncertainty is that of the early years (1880-1920), where the average uncertainty values are higher at the beginning of the period, probably due to the poor availability of global data, rather advanced mathematical treatments of homogenization and extension of the data and less advanced methodologies for estimating temperature anomalies.
Relevant is the peak around the 1940s, during the Second World War, where there is a clear increase in the average uncertainty right during the conflict period (1940-1945), which is very pronounced in the representations of the HadCRUT5, Berkeley Earth datasets. This could have been due to the difficulty of sampling global temperature data during the Second World War.
Since the 1960s, uncertainties have shown a relatively stable trend, probably due to the implementation and replacement of obsolete climate observation stations.
However, after 2000 a sharp increase in the trend of the uncertainty of the anomalies is observed, which can be attributed to the integration of satellite and terrestrial data and to new mathematical interpolation methods for the representation of temperature data over the entire Earth's surface. However, these variations are not uniformly detected for all datasets, still denoting a divergence between the mathematical methods used.
In summary, although all the curves show a reduction in uncertainty over time, especially since the 1960s, with a more uniform trend in recent decades, probably reflecting an evolution and improvement in data collection and analysis, significant divergences are noted between 1880 and 1920, a significant peak between 1940 and 1945 (Second World War), a convergence of uncertainty values from 1960 onwards, suggesting greater coherence between the analysis methods used

and a sudden change in the uncertainty trend, after 2000, for the reasons mentioned above.

Importance of data consistency and homogeneity

Climatology is a young science that has acquired a significant role and a more systematic structure mainly during the twentieth century.

Despite the dynamic technological and methodological progress of the last decade, climatology bases its evaluations on datasets collected over a time span of about 200 years, with uncalibrated and dissimilar instrumentation, which has sampled temperature data over time with shortcomings both in the quality and temporal continuity of the data, and in the mathematical homogenization treatment used for its cataloguing.

In fact, the main disharmonies detected in the datasets relate to the inhomogeneity and the relative brevity of the available time series, which in fact denature the reliability of the definition of correct and truthful predictive models, which being defined through mathematical artifices and simulations that "chase" real models that do not take into account the uncertainties and indeterminacies of the data, represent unrealistic climate trends.

It is therefore necessary to accept and become aware of the limitations that still exist today on climate time series and of the need, and if we resort to statistical analysis of the data, to have long and continuous time series, detected with identical instruments, which are essential to be able to define truthful predictive analysis models of climate change trends.

Inaccurate or poorly managed data can lead and have led to misinterpretations of change trends and the definition of altered models of climate change, negatively influencing the formulation of public policies and the global response to this crisis.

The homogenization techniques used have in fact manipulated, mathematically and statistically, the collected data, in order to be able to correct the inconsistencies and errors defined by the representation of the raw data, due precisely to changes in the techniques and technologies of detection, to the different location of the sampling stations and to other environmental and anthropic factors. Specifically, the application of such corrective interventions, which include interpolation for temperature estimates in unmonitored areas, gridding for the standardization of the measurements in a uniform matrix, and the correction of anomalies to adjust the data to known or suspected variations, lead to the definition of artificial and inadequate datasets.

Given the complexity and variety of homogenization techniques used, it is essential to proceed through a critical and continuous evaluation of methodologies capable of identifying potential sources of error, the analysis and development of international standards for sampling, processing and cataloging of climate data.

Only through a careful assessment of the data acquisition process and a standardization of homogenization practices will it be possible to improve the accuracy of climate projections and, consequently, the responsiveness of predictive models.

Trust in climate data is of crucial importance not only for the scientific community but also for citizens and policy makers, who base climate change mitigation and adaptation policies on such data.

The enormous intrinsic responsibility of such data leads us to affirm how fundamental it is that the datasets made available by the Agencies reflect, with the utmost precision, the possible real climatic conditions of the entire planet.

Errors or inaccuracies in datasets inevitably lead to inaccurate or incorrect conclusions, which in turn lead to the formulation of policies that are ineffective or contrary to the long-term interests of society.

One example among all, explanatory but not exhaustive of the different inaccuracies, concerns the errors in evaluating the "speed of variation of climate change", which could accelerate the governance actions to be undertaken to counteract the change in speed, with unjustified alarmism and economic overloads for the community that are certainly not justifiable, through investments in resources and technologies that may not be necessary, diverting funds and attention from more pressing problems.

To address these challenges, it is essential that the international community collaborates to improve data sampling techniques, data management and the mathematical analysis that leads to the formulation of climate datasets.

Processes that include the adoption of international standards for the calibration of measuring instruments, the development of shared protocols for data processing and investment in advanced technologies for environmental monitoring.

Only through such concerted efforts will it be possible to achieve the precision and coherence needed to guide humanity towards informed and effective responses to what is really happening with regard to climate change, one of the greatest dilemmas of our time.

Conclusions

In light of the arguments presented, it emerges with extreme transparency how current climate models and related forecasts, however sophisticated they are, present future scenarios that are not fully truthful and consistent with the real natural trends of the entire planet Earth.

The inherent uncertainties in climate data sets, coupled with the challenges posed by different Agencies to homogenize them, place significant limits on the ability to accurately represent long-term global climate dynamics.

Therefore, the construction of reliable predictive models is intrinsically linked to the availability of extended and homogeneous time series, which can offer a solid basis for statistical analyses and for the definition and understanding of climate trends, datasets that in fact are not and cannot be available yet.

The use of mathematical treatments to homogenize historically inhomogeneous raw data introduces

distortions of natural trends, leading to misleading interpretations of observed climate phenomena defined on series of information lacking in spatial and temporal coverage. Such deficits must be carefully evaluated in order to avoid representing a distorted picture of current and future climate reality.

It is essential, therefore, that the scientific community becomes aware of these critical issues and that it works towards greater transparency and greater rigor in the sampling, management and interpretation of climate data. The call is to join forces to develop community standards and shared protocols for the treatment of climate data, thus promoting more reliable climate models that reflect as faithfully as possible the complexity and intrinsic variability of the Earth's climate system.

Bibliography

1. Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., & Jones, P. D. (2006). Uncertainty estimates in observed regional and global temperature changes: a new data set from 1850. Journal of Geophysical Research: Atmospheres, 111(D12).
2. Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. Reviews of Geophysics, 48(4), RG4004.
3. Hansen, J., Sato, M., & Ruedy, R. (2012). Perception of climate change. Proceedings of the National Academy of Sciences, 109(37), E2415-E2423.
4. Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset. Journal of Geophysical Research: Atmospheres, 117(D8).
5. Peterson, T. C., & Vose, R. S. (1997). An overview of the Global Historical Climatology Network temperature database. Bulletin of the American Meteorological Society, 78(12), 2837-2849.
6. Thorne, P. W., Parker, D. E., Titchner, H., Rayner, N. A., & McCarthy, M. (2011). Uncertainties in climate trends: lessons from upper air temperature records. Bulletin of the American Meteorological Society, 92(10), 1417-1422.
7. Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J., et al. (2012). Benchmarking homogenization algorithms for monthly data. Climate of the Past, 8(1), 89-115.