Widespread sampling bias in atmospheric data

Andrew Kowalski (andyk@ugr.es)
University of Granada https://orcid.org/0000-0001-9777-9708
Matilde García-Valdecasas Ojeda
University of Granada

Brief Communication

Keywords:

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1592600/v1

Abstract

Air's compressibility practically precludes unbiased atmospheric sampling because warmer air, occupying more specific volume at equal pressure, is more likely to be measured. Weighted statistics are required to compensate for the effects of sampling bias on statistical moments of state and flow variables. These effects occur from the microscale to the global scale, at which uncorrected sampling bias causes underestimation of both global warming and its statistical certainty.

Introduction

The Eulerian frame of reference is convenient for atmospheric observation and modelling but conceals prevalent sampling bias in atmospheric data. This frame examines a control volume or "point" whose mass varies with inward and outward air fluxes, accommodating immobile measurement stations and model grid cells. By contrast, a Lagrangian frame that follows a particular fluid mass is preferable for expressing physical conservation laws[1], and therefore for recognizing sampling bias. It shows that in the atmosphere it is practically impossible to practice "simple random sampling", which requires an equal sampling likelihood for every element of the population and is the key to avoiding bias and simplifying analytic computations[2].

Since Lagrangian measurement is an exceedingly difficult challenge, it is demonstrative to specify a simple, artificial dataset. Consider therefore a thermodynamic system composed of many dry air parcels, each with identical mass and pressure but thermally divided into two categories, with half the parcels warm (20 °C) and half cold (-20 °C). Clearly, the average temperature is 0 °C. However, because most of the aggregate volume is warm (by Charles's Law), a thermometer deployed randomly within the system is more likely to sample warm air than cold, despite their having equal masses. This exposes a universal issue with atmospheric sampling and reveals bias that repeated measurement does not easily amend.

Increasing the number of observations in this simple example, whether by random or gridded sampling, would yield > 53% warm samples and an overestimation of the mean temperature from arithmetic averaging by nearly 1.5 °C. Dividing the system into numerous equal-volume grid cells would produce more warm cells than cold, denser cells. Such spatially defined measurement or modelling domains are typical in atmospheric science and exemplify what statisticians call "cluster sampling"[3]. Critically, cluster populations vary with atmospheric state (particularly temperature) but are unknown *a priori*. This precludes unbiased sampling unless the instrument sample volume exceeds the scales of atmospheric variability.

Such bias affects the statistical moments of any state or flow variable. If the parcels defined above were moving vertically in a convective boundary layer, with warm/cold parcels rising/falling each at 10 cm s^{-1}, the average velocity would be zero. However, an anemometer deployed to a random point would sample rising air with >53% probability. Arithmetic averaging of measurements from numerous such anemometers within the system, or of a time series measured by a single instrument with "frozen

turbulence" advecting past (employing Taylor's hypothesis[4]), would then yield an upward average vertical velocity of 7 mm s^{-1}, again because sampling bias nudges the arithmetic average towards the characteristics of warmer air.

Failure to recognize this has confused the physical mechanisms of convective boundary-layer transport. Long-standing theory in micrometeorology, reasoned from the Eulerian perspective, has it that turbulence transports air downward against an upward heat flux, since warm updrafts are less dense than cool downdrafts[5,6]. This would imply an average velocity in the direction of the heat flux (upward), and is the basis of the "density corrections"[6] that have underpinned decades of micrometeorological research[7]. But the simple example specified above illustrates its fault: with equal mass and opposing velocities of warm and cold parcels, their momenta sum to zero, defining a null ensemble velocity. Cold Eulerian volumes are denser but also fewer – the same mass occupies less volume. Neglecting this, random spatial sampling and arithmetic averaging combine to produce an upward average velocity that is an artefact of sampling bias. In fact, neither the turbulence nor the mean flow transports air for the conditions specified. Thus, unrecognized sampling bias has led to the entangling of gas transport mechanisms, by turbulent diffusion versus by the mean flow[8].

Systematic errors such as those derived from sampling bias should be quantified when known and corrected when not negligible. If sampling bias cannot be avoided, unbiased estimates of statistical moments such as the average can be obtained via weighted calculation schemes. The Horvitz-Thompson estimator[9] was designed for this purpose, and for dry air can accurately approximate the average by weighting each sample by the number of (moles of) air molecules that it represents.

However, atmospheric composition varies, and the leverage exerted by a sample on the aggregate is determined not by its molecule count, but rather by conservation laws. For example, in thermodynamics the determinant is the heat content: when we combine equal numbers of molecules of water vapour at one temperature and dry air at another, the temperature of the mixture is closer to that of the water vapour due to its high specific heat, inferior mass notwithstanding. Generally, any isolated system of air molecules that mixes internally conserves its mass, linear momentum, and energy, yielding final values of state and flow variables that correspond to their ensemble averages, which remain constant in the absence of external forcing. Therefore, weighting factors have been derived from mixing theory[10] to negate the systematic errors that sampling bias otherwise induces.

The effects of sampling bias span the full range of atmospheric scales. We have calculated the magnitude of sampling-bias induced errors regarding global-scale warming, comparing trends in the National Center for Environmental Prediction (NCEP)-National Center for Atmospheric Research (NCAR) reanalysis[11] temperatures averaged arithmetically (T_a) versus weighting (T_w) according to mixing theory (Fig. 1). The artefacts of bias towards the characteristics of warm air on T_a include (a) the overestimation of the global-average surface air temperature by *ca*. 0.4 °C; (b) the underestimation of the rate of temperature increase by about 9%, since the tropics have warmed less rapidly than the poles[12]; and (c)

the underestimation of the variance explained by the linear trend. The difference between trends in T_w and T_a is of a magnitude similar to the differences between ensemble members of the HadCRUT4 data set for the period 1979-2010[13]. Thus, uncorrected sampling bias causes meaningful underestimation of both the magnitude and certainty of global warming. (Similar conclusions were reached using the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis data (ERA5)[14,15]; see Supplementary Information.)

Unless unbiased sampling strategies can be devised, micro- to global-scale atmospheric science requires weighted calculation of averages and other statistical moments, in order to avoid systematic error.

Methods

Averaging in space and time

Global data from NCEP-NCAR reanalysis for surface air temperature (T in Kelvin), surface pressure (p in Pa), and relative humidity (U as a percentage) were downloaded with a time resolution of six hours. Data are structured according to a 2.5° X 2.5° grid (144 X 73 grid points), selecting the years 1950–2021 that coincide with available ERA5 data to enable comparison. Climate Data Operators (CDO) software was used to facilitate calculations on grid points.

For each grid point i and time j, the saturation vapour pressure (e_s in Pa) was estimated as a function of T using Eq. (10) of Bolton (1980)[16], and the saturation mixing ratio (r_s, dimensionless) calculated as

$$r_{s_{i,j}} = 0.622 \frac{e_{s_{i,j}}}{p_{i,j} - e_{s_{i,j}}} \quad (1)$$

The mixing ratio (r, dimensionless) is then

$$r_{i,j} = \frac{U_{i,j}}{100} r_{s_{i,j}} \quad (2)$$

and the vapour pressure (e in Pa)

$$e_{i,j} = \frac{p_{i,j} r_{i,j}}{r_{i,j} + 0.622} \quad (3)$$

Dry air (ρ_d) and water vapour (ρ_v) densities come from the ideal gas law as

$$\rho_{d_{i,j}} = \frac{p_{i,j} - e_{i,j}}{R_d T_{i,j}} \quad (4)$$

and

$$\rho_{v_{i,j}} = \frac{e_{i,j}}{R_v T_{i,j}} \quad (5)$$

where R_d (287.05 J kg^{-1} K^{-1}) and R_v (461.51 J kg^{-1} K^{-1}) are the particular gas constants for dry air and water vapour.

Sample masses of dry air (m_d) and water vapour (m_v) were calculated as the product of density (ρ_d and ρ_v for dry and water vapour, respectively) and volume (V), where grid cell V is the product of the height (2m) and the area output by the CDO 'gridarea' command.

Then, yearly arithmetic averages were calculated by weighting each data point by the grid cell volume, as

$$T_a = \frac{\sum_{i=1,\,j=1}^{n,\,m} V_j T_{i,j}}{\sum_{i=1,\,j=1}^{n,\,m} V_j} \quad (6)$$

with n being the total number of points in time in a year (1460, or 1464 for leap years) and m the total number of grid points for the entire globe (144 x 73).

In the same way, annual mixing-theory weighted averages were calculated by weighting each data point by the heat content, as

$$T_w = \frac{\sum_{i=1,\,j=1}^{n,\,m} \left(c_{pd}\, m_{d_{i,j}} + c_{pv}\, m_{v_{i,j}} \right) T_{i,j}}{\sum_{i=1,\,j=1}^{n,\,m} \left(c_{pd}\, m_{d_{i,j}} + c_{pv}\, m_{v_{i,j}} \right)} \quad (7)$$

where c_{pd} (1005 J kg^{-1} K^{-1}) and c_{pv} (1850 J kg^{-1} K^{-1}) are the specific heat at constant pressure for dry air and water vapour, respectively.

Estimation of the global annual trends

Global annual trends from 1950 to 2021 were estimated using linear least-square regression.

Data Availability

Data related to the paper can be downloaded from following websites:

NCEP database:

https://www.psl.noaa.gov/data/gridded/data.ncep.reanalysis.surface.html

ERA5 database:

https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-preliminary-back-extension?tab=form

https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form

Code availability

MATLAB scripts for calculating averages are available on request from M.G-V.O.

Declarations

Acknowledgements

This paper was financed by the Spanish Ministry of Science and Innovation project INTEGRATYON3 (PID2020-117825GB-C21), the FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (project P20_00035), and the Andalusian regional government project ICAERSA (P18-RT-3629), including European Union ERDF funds. We thank Ana López-Ballesteros for suggestions that improved the clarity of the manuscript.

Authors Information

Affiliations

Departamento de Física Aplicada, Universidad de Granada, Granada, Spain

Andrew S. Kowalski & Matilde García-Valdecasas Ojeda

Instituto Interuniversitario de Investigación del Sistema Tierra en Andalucía (IISTA), Granada, Spain

Andrew S. Kowalski & Matilde García-Valdecasas Ojeda

Contributions

A.S.K. conceived the analysis and wrote the initial manuscript; M.G-V.O. acquired climatological data, conducted the analysis, produced the results, wrote the methods section and supplementary information, and contributed to interpreting results and improving the paper.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Holton, J. R. *An introduction to dynamic meteorology* (Academic Press, New York, 1992).
2. Ramachandran, K. M. & Tsokos, C.P. *Mathematical Statistics with Applications* (Elsevier Academic Press, London, 2009).
3. Särndal, E. -K., Swensson, B. & Wretman, J. *Model Assisted Survey Sampling* (Springer-Verlag, New York, 1992).
4. Stull, R. B. *An introduction to boundary layer meteorology* (Kluwer, Dordrecht, 1988).
5. Priestley, C. H. B. & Swinbank, W. C., *Proceedings of the Royal Society of London*, **189**, 543–561 (1947).
6. Webb, E. K., Pearman, G. I. & Leuning, R., Q. J. R. Meteorol. Soc., **106**, 85–100 (1980).
7. Lee, X. & Massman, W. J., Bound.-Layer Meteorol., **139**, 37–59 (2011).
8. Kowalski, A. S., Serrano-Ortiz, P., Miranda-García, G. & Fratini, G., Bound.-Layer Meteorol., **179**, 347–367 (2021).
9. Horvitz, D. G. & Thompson, D. J., J. Am. Stat. Assoc, **47**, 663–685 (1952).
10. Kowalski, A. S., J. Atmos. Sci., **69**, 1750–1757 (2012).
11. Kalnay, E. et al., Bull. Amer. Meteor. Soc.; **77**: 437–472 (1996).
12. Taylor, P. C., Cai, M., Hu, A., Meehl, J., Washington, W. & Zhang. G. J., J. Clim., **26**, 7023–7043 (2013).
13. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D., J. Geophys. Res., **117**, 1984–2012 (2012).
14. Hersbach, H. et al., Q J R Meteorol Soc. **146**, 1999–2049 (2020).
15. Bell, B. et al., Q J R Meteorol Soc., **147**, 4186–4227 (2021).
16. Bolton, D., Mon. Weather Rev., **108**, 1046–1053 (1980).

Figures

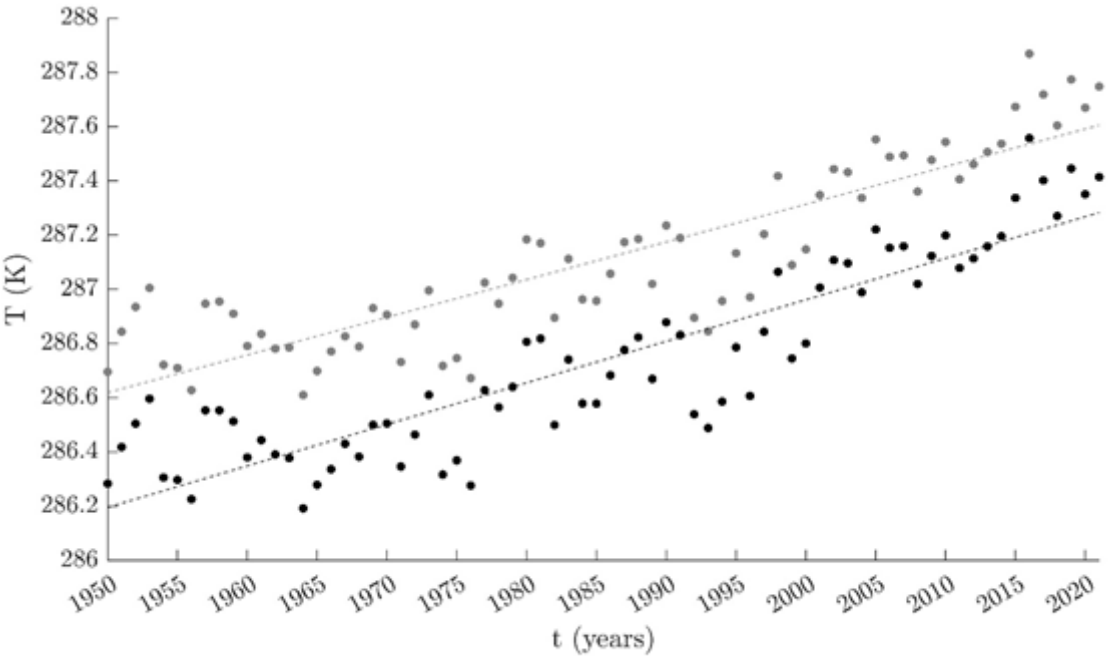

Figure 1

Annual surface air temperatures from NCEP-NCAR reanalysis, globally averaged, arithmetically (T_a) versus using mixing-theory weighting (T_w). Corresponding regression lines have slopes (R^2 values) of 0.139 °C per decade (0.78) for T_a, versus 0.153 °C per decade (0.81) for T_w.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf