

DATA COMPRESSION

Advancing data compression via noise detection

Compressing scientific data is essential to save on storage space, but doing so effectively while ensuring that the conclusions from the data are not affected remains a challenging task. A recent paper proposes a new method to identify numerical noise from floating-point atmospheric data, which can lead to a more effective compression.

Dorit M. Hammerling and Allison H. Baker

Checking the weather forecast with an application (or app) on one's phone is a critical part of many morning routines. For example, when making the decision of whether to wear pants or shorts on a sunny day in Colorado, if the app reports that the day's high will be 50 degrees Fahrenheit (°F), most people will potentially opt for pants. But what if the app reports 50.456783957642 °F for the high? Probably still pants. For most of us, twelve extra decimal places of information are unlikely to influence this decision. Weather (and climate) forecasts come from complex model simulation codes, and while simulation codes are an important tool in scientific discovery, one must put the information that they output into context. If the weather app reported that today's high would be 50.456783957642 °F, a few issues would come to mind. First, do we need that much precision for the data to be useful? The precision probably depends on how we will use the data. Second, are all these digits meaningful? Do they contain real information? Because computers store and operate on data in predetermined sizes, generally 8 bytes (double-precision) or 4 bytes (single precision), for many simulation codes a lot of that information is either not necessary for its intended use (meaning, the weather app) or the trailing digits may just be numerical noise (called false precision by Zender¹). The bottom line is that models often use a higher precision than what is physically important (for instance, storing temperature as 64 bits), and precision does not necessarily equal practically useful information. Writing in *Nature Computational Science*, Milan Klöwer and colleagues² propose a method that automatically determines noise in large atmospheric data, leading the way to advance compression of large scientific datasets from numerical models.

There is a subtle difference between the notion of noise and the precision that we may want in a variable in practice. Sometimes, we have more precision than what is needed, and we could drop more

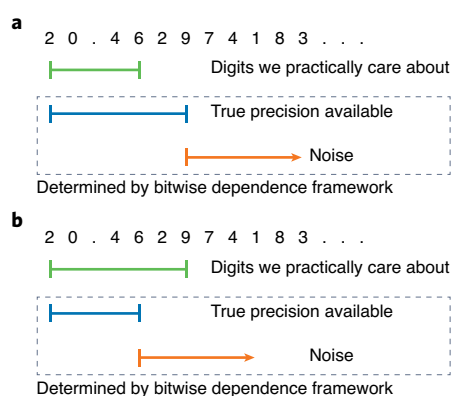


Fig. 1 | The notion of noise and precision.

a, b. The figure illustrates that, sometimes, we have more precision than what is needed (**a**) and we could drop more bits than just those representing noise, while we may also have less precision than desired (**b**) and should recognize which digits are noisy. The approach developed by Milan Klöwer and colleagues² provides a means to automatically distinguish true precision from noise by assessing bitwise dependence across adjacent grid cells.

trailing bits than just those representing noise (Fig. 1a). The opposite scenario can also occur, where the noise exceeds our desired level of precision (Fig. 1b). Regardless of the scenario, assessing which bits represent noise is important since using valuable storage space for noise is undesirable. Storage is a particularly critical issue for Earth system models (ESMs), which have benefited tremendously from advances in computing power in recent decades and, consequently, are capable of producing enormous volumes of data.

Data compression has long been popular for image, audio, and video; comparatively, compressing scientific simulation data has only garnered interest recently. While lossless compression exactly preserves the original data, lossy compression does not, which allows for much higher compression rates and subsequently smaller storage requirements. This trade-off is particularly

important for floating-point simulation data, which often appear random after the first several significant digits, particularly for double-precision computations. This randomness, also known as noise, largely results from magnified roundoff errors incurred by the floating-point arithmetic in the simulation, as well as by model truncation and discretization errors. Lossless compressors are, in fact, ineffective at compressing random data³, as they rely on finding patterns in the data that can be represented with less information. Lossy compressors, on the other hand, can be quite effective for such data as they have the flexibility to not exactly reproduce the (random) data. For example, truncation is a simplistic form of lossy compression that essentially 'chops-off' or 'zeros-out' some of the least significant bits. By changing the noisy bits to zeros, data become much more likely to have detectable patterns (assuming spatial correlation), meaning that their size can be effectively reduced by a lossless compressor. Here, the lossy compression (meaning, truncation) can be thought of as a preconditioning step that enables an effective application of a lossless method. But in practice, how do we determine what digits are noise in order to more effectively compress the data while ensuring that we do not affect the scientific conclusions drawn from the data?


Providing a possible answer to this question is the key contribution of the work by Milan Klöwer and colleagues². The authors approach this issue by assessing bitwise dependence to adjacent grid points (in space, time, or other dimensions). The key idea is that stronger dependence to adjacent grid points implies higher real information content, while lack of dependence equates to noise, or equivalently, false information. This procedure is conducted sequentially for the positions in the mantissa, evaluating dependence across adjacent cells for the same bit position. This setup leads to an automatic determination of precision, that is, the separation of real information from noise, through the data

itself, without the need for user intervention or assessment. This automatic determination of precision is a particularly attractive aspect in the context of large numerical models, such as ESMs, that contain hundreds of variables, where obtaining input from users on matters of required precision can be a notoriously daunting task (at present, acceptable levels of compression are typically detected via a costly brute-force approach that evaluates various precisions⁴). Once the true precision is automatically determined, the bits representing noise are rounded to zero. This step corresponds to the preconditioning step mentioned above that then enables a standard lossless compressor to be applied more efficiently to the remaining bits (representing real information), since the noise elimination results in more detectable patterns for the compressor.

This ingenious approach works well for all of the examples shown in the manuscript, and will likely work well in most scenarios involving weather prediction, where the variables exhibit continuity and smoothness. It is, however, important to be aware, as the authors also mention, that the key underlying assumption of dependence (or lack thereof) as a sensical means to assess precision might not hold in other contexts.

There are certainly other applications where a lack of dependence is legitimate and doesn't correspond to noise (or false information), such as data resulting from independent Poisson processes. One would want to be careful about using the proposed framework in such a context.

An important challenge when applying lossy compression to scientific simulation data is determining the degree of information loss that is desired (or acceptable). To this end, the proposed bitwise-correlation framework is an exciting new contribution to the lossy compression literature, particularly for weather and climate data, and likely for many other application areas involving large numerical models. It will be interesting to assess this framework with a broader array of applications and by paying attention to more subtle features, such as the preservation of extreme values⁵ and a set of derived quantities⁶, which can sometimes highlight compression-induced artifacts not apparent through simple standard compression metrics (for instance, root mean square error). Further, an exciting avenue to explore is whether the bitwise correlation characteristics identified by this framework can be used to simplify and guide the choice of lossy compressors (and their parameters).

If so, this advance would be particularly beneficial to climate and weather applications where there may be hundreds of variables with differing characteristics that require careful individual assessment. 

Dorit M. Hammerling¹  and Allison H. Baker²

¹Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO, USA. ²Computational and Information Systems Laboratory, National Center for Atmospheric Research, Boulder, CO, USA.

e-mail: hammerling@mines.edu

Published online: 25 November 2021
<https://doi.org/10.1038/s43588-021-00167-z>

References

1. Zender, C. S. *Geosci. Mod. Dev.* **9**, 3199–3211 (2016).
2. Klöwer, M., Razinger, M., Dominguez, J. J., Düben, P. D. & Palmer, T. N. *Nat. Comput. Sci.* <https://doi.org/10.1038/s43588-021-00156-2> (2021).
3. Lindstrom, P. Error distributions of lossy floating-point compressors. In *JSM Proc.* 2574–2589 (2017).
4. Lindstrom, P. & Isenbarg, M. *IEEE Trans. Vis. Comput. Graph.* **12**, 1245–1250 (2006).
5. Baker, A. H. et al. *Geosci. Mod. Dev.* **9**, 4381–4403 (2016).
6. Poppick, A. J. et al. *Comput. Geosci.* **145**, 104599 (2020).

Competing interests

The authors declare no competing interests.