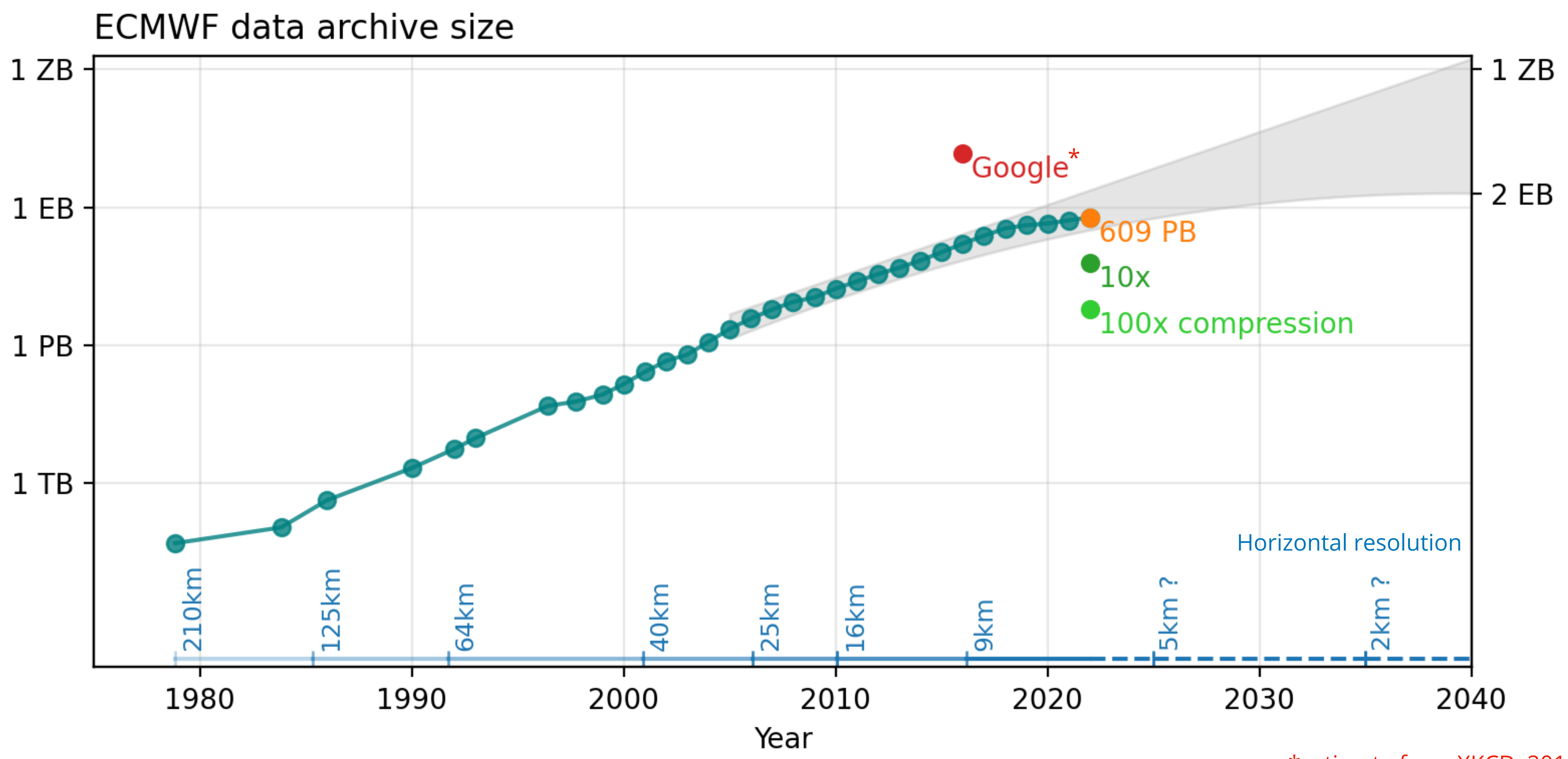
What is the real information in weather and climate data?

Milan Klöwer[1], Ayoub Fatihi, Miha Razinger[2], Juanjo Dominguez[2], Aaron Spring[3], Hauke Schulz[4], Peter Düben[2], Tim Palmer[1]

[1]Massachusetts Institute of Technology, Cambridge USA
[2]European Centre for Medium-Range Weather Forecasts, UK
[3]Max Planck Institute for Meteorology, Hamburg, Germany
[4]University of Washington, Seattle, USA

Will we enter the *Google regime*?

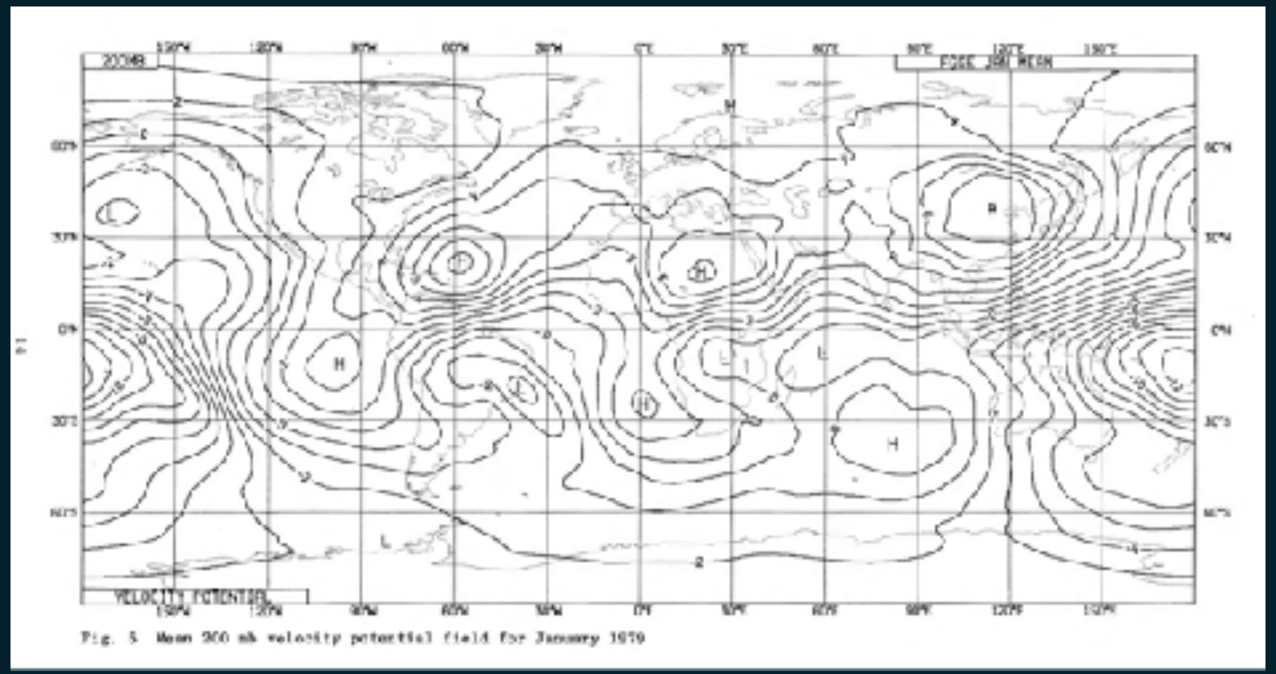


*estimate from XKCD, 2016

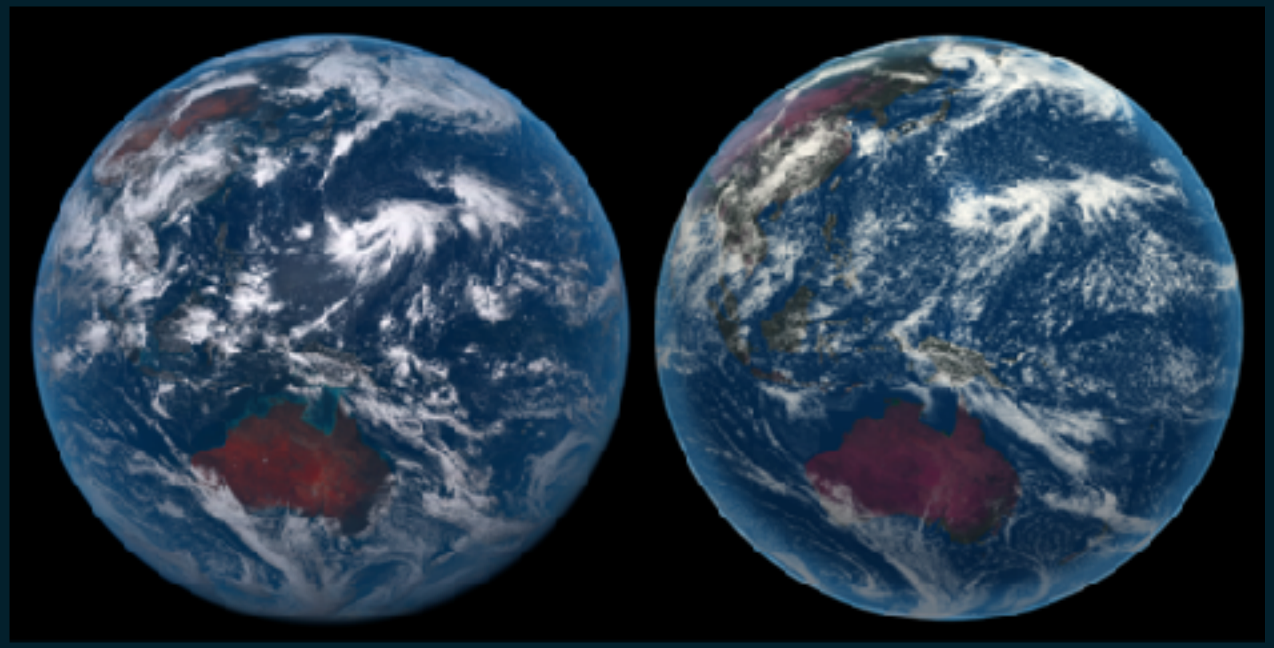
40+ years of numerical weather prediction

How much have we questioned the bitwise representation of our data?

ECMWF 1979



Digital twins, Stevens 2019



IEEE Standard

64-bit floats

32-bit floats

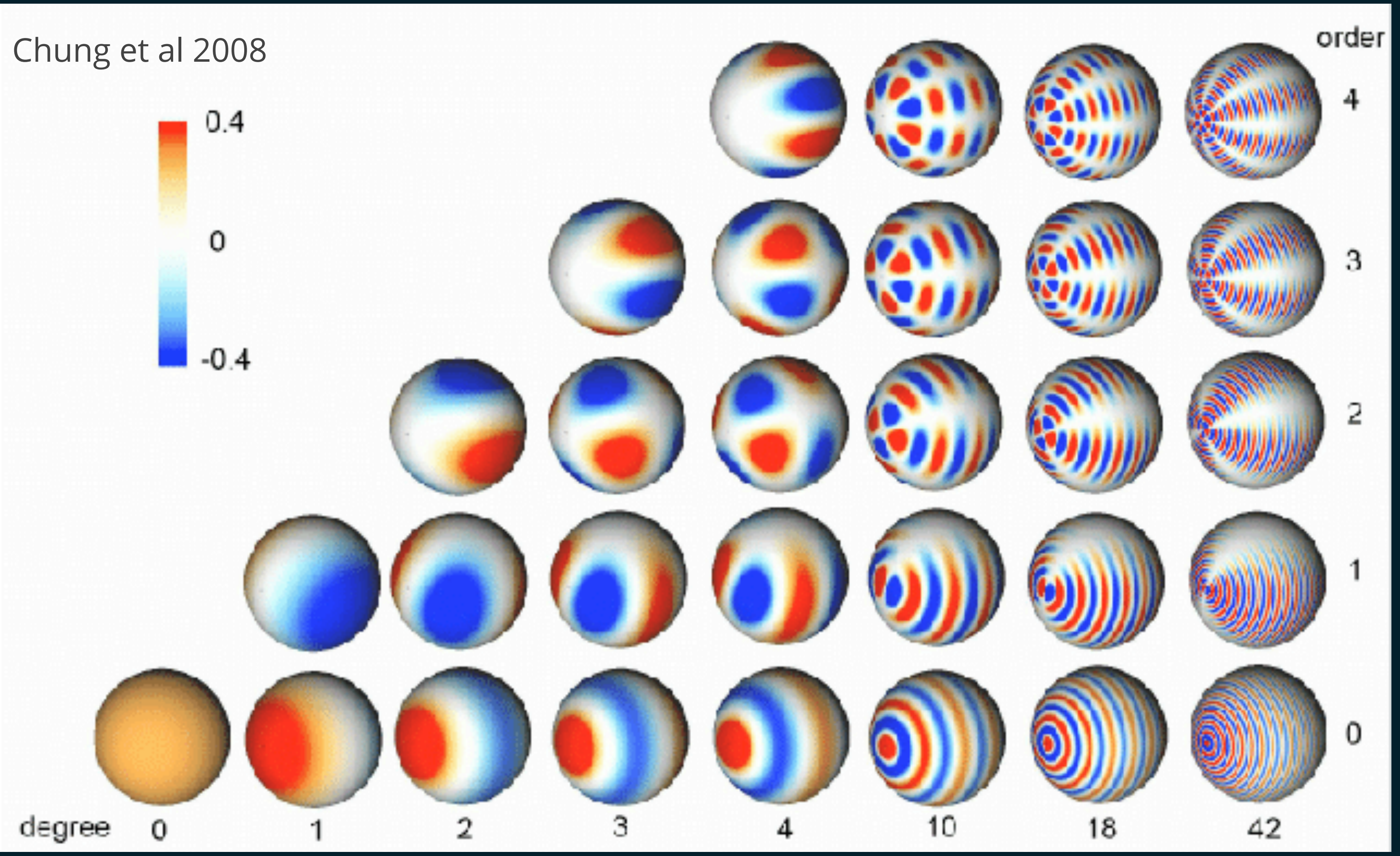
1978 1985 2021

Schools of thoughts: Data compression

School 1: Transformations
(the physical perspective)

- Spectral
- EOF
- Tensor trains
- Neural Networks

- Spatial structure ✓
- Compute expensive ✗
- Error bounds difficult (✗)
- No random access (✗)



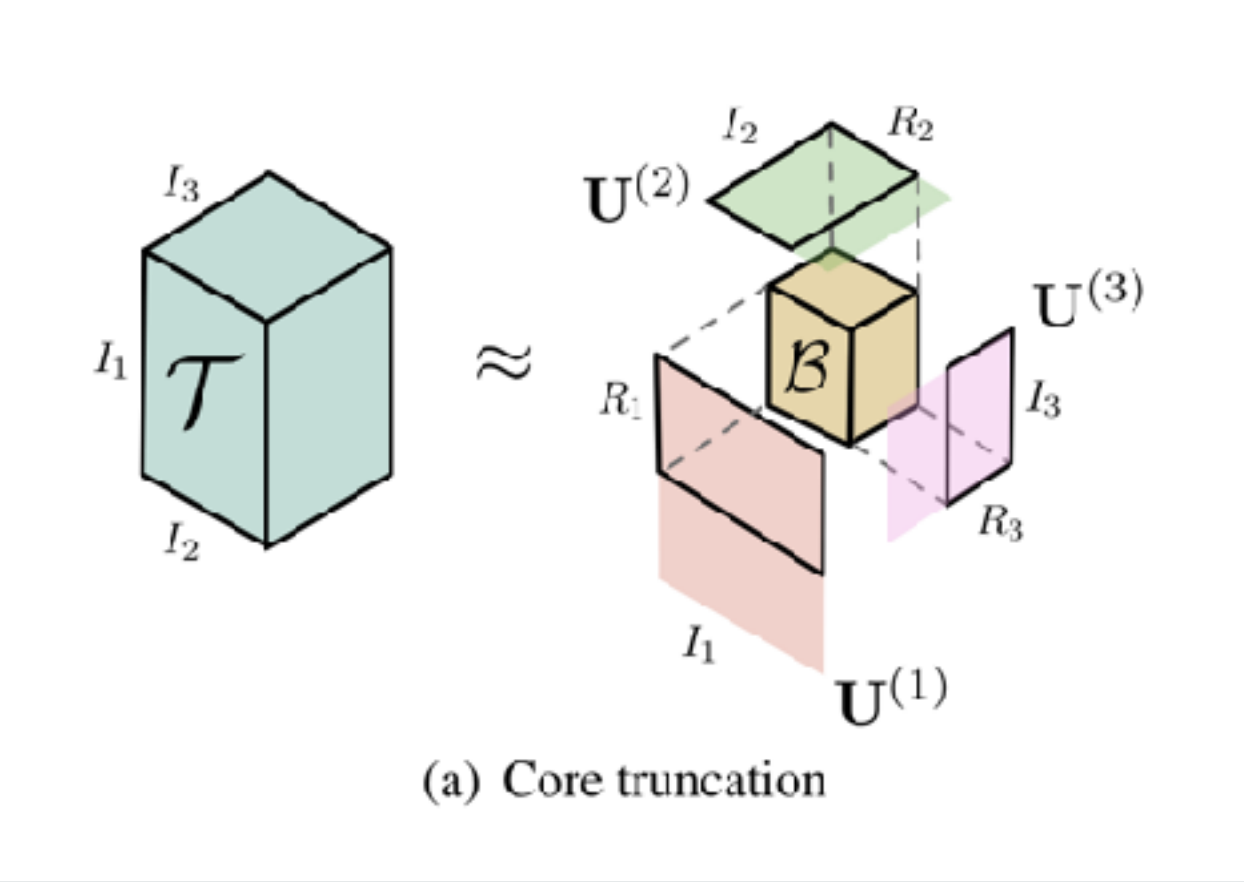
School 2: Precision and information theory
(the binary perspective)

- Bitwise encoding
- Floats; linear or logarithmic quantisation
- Entropy coding
- Lossless compression

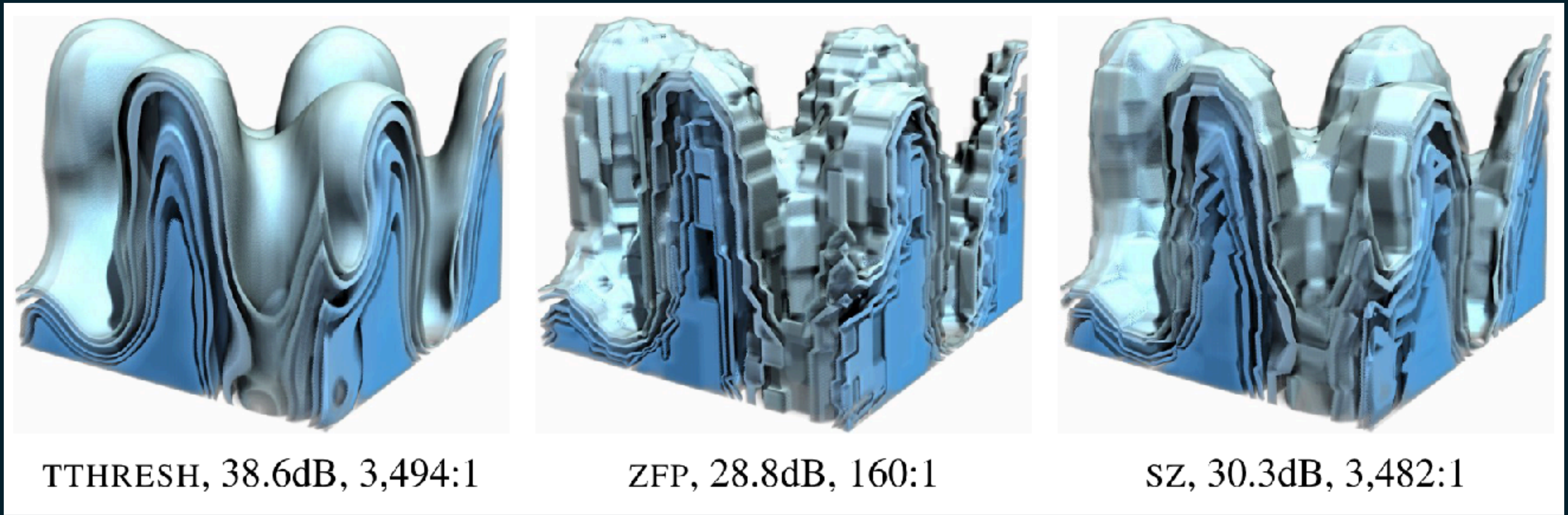
- Spatial structure (✗)
- Compute cheap ✓
- Rigid error bounds ✓
- Random access (✓)

Transformations: TensorTrains

$$T = B \cdot U^{(1)} \cdot U^{(2)} \cdot \ldots \cdot U^{(n)}$$



Ballester-Ripoll et al, 2019



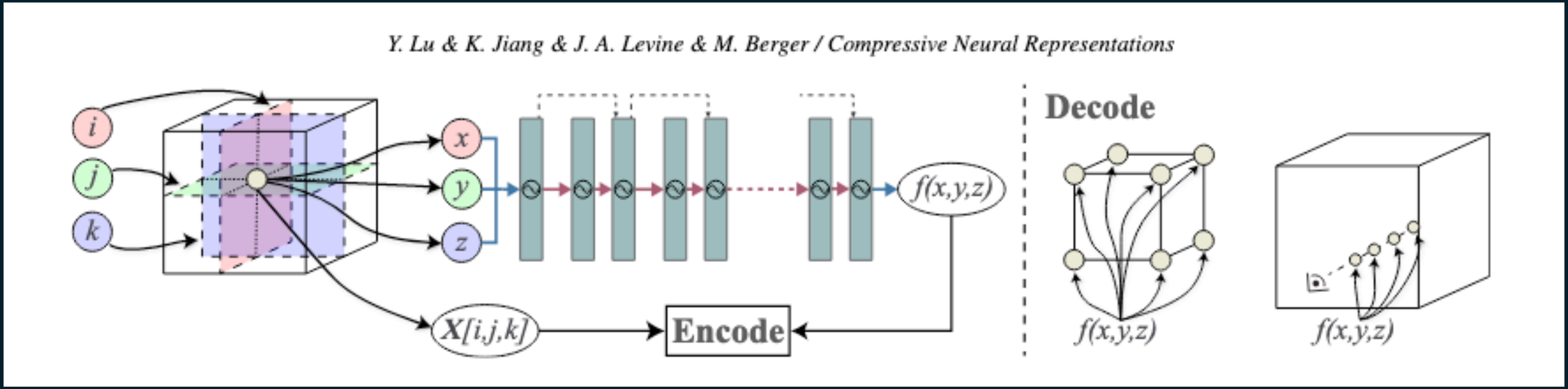
Advantages

- Approximates well smooth data
- Claim: >1000x compression possible

Disadvantages

- Expensive (de)compression
- Changing statistics?
- Unstructured grids?

Neural network compression



Advantages
- Interpolates automatically
- Unstructured grids
- Error scales with training, i.e. compute
- ~1000x compression possible

Disadvantages
- Difficult to control the error
- Expensive (de)compression
- Size impacts the error

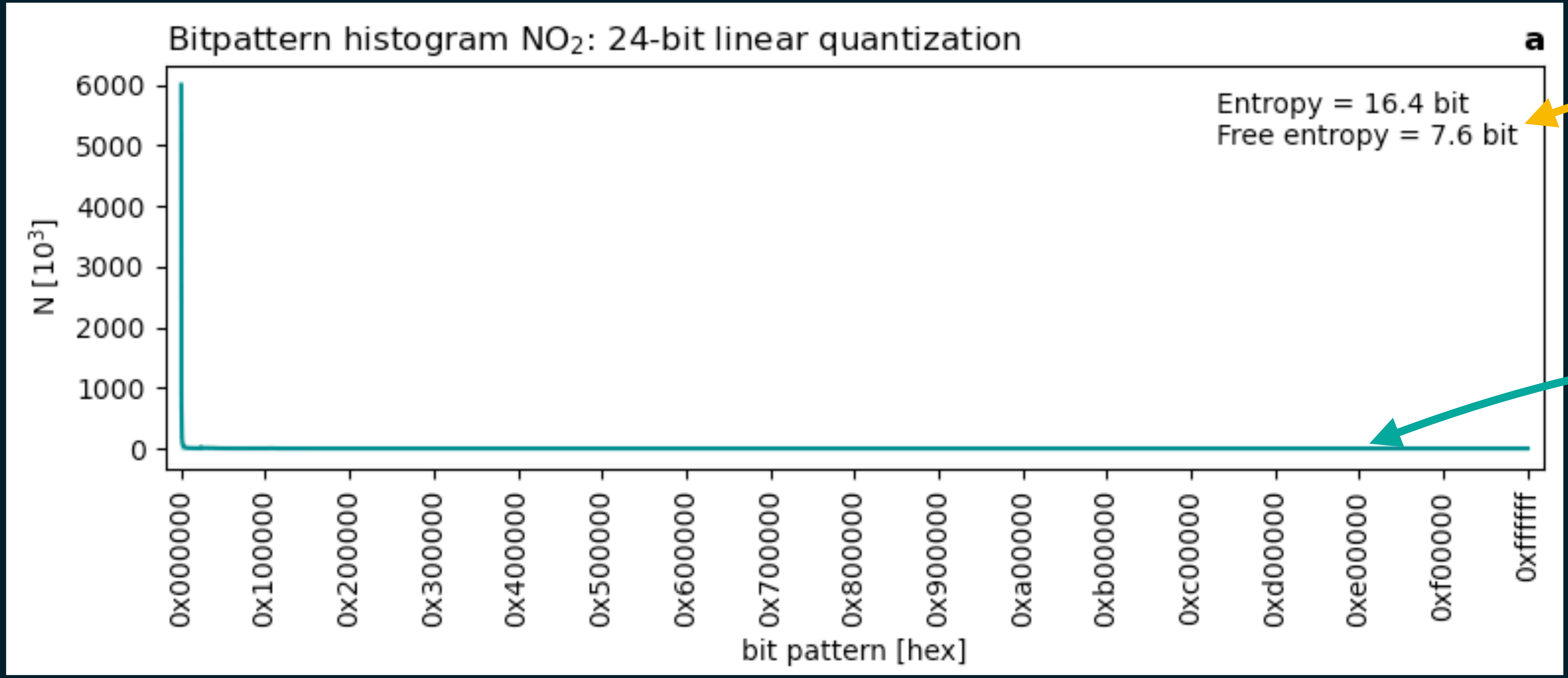
See Huang, Hoefler, 2023: *Compression of weather and climate data into neural networks*

Current compression methods: 24-bit linear quantization

split into 2^{nbits} equidistant intervals

min max

Use nbits = 24 instead of 64 bit per number



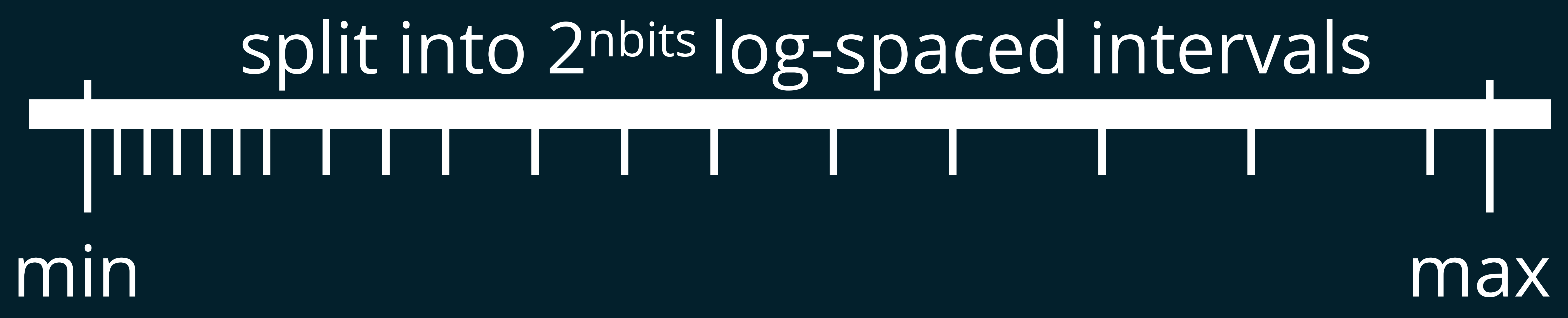
>7 bits are effectively unused

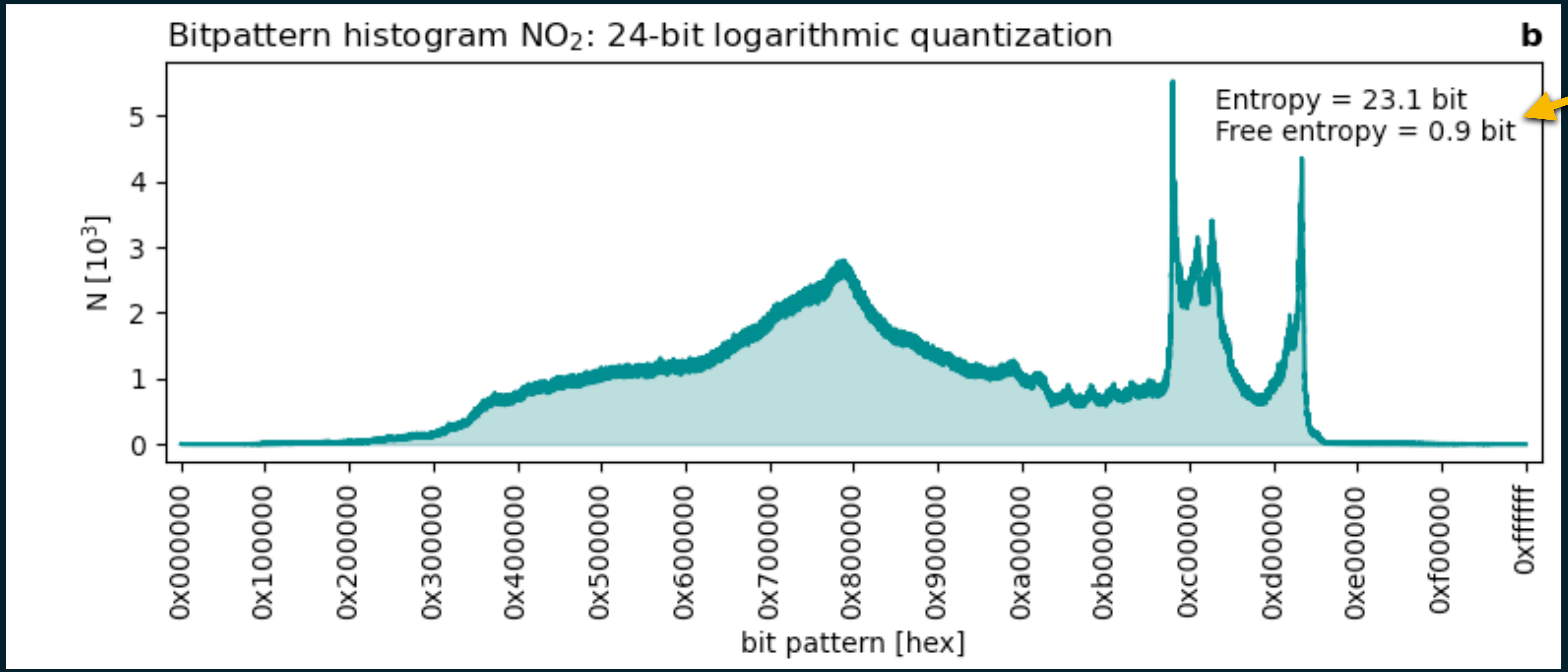
Most bit patterns very rarely used

How much information is in the first bit?

~0 bit. Information <= Entropy.

Alternative: Logarithmic quantization





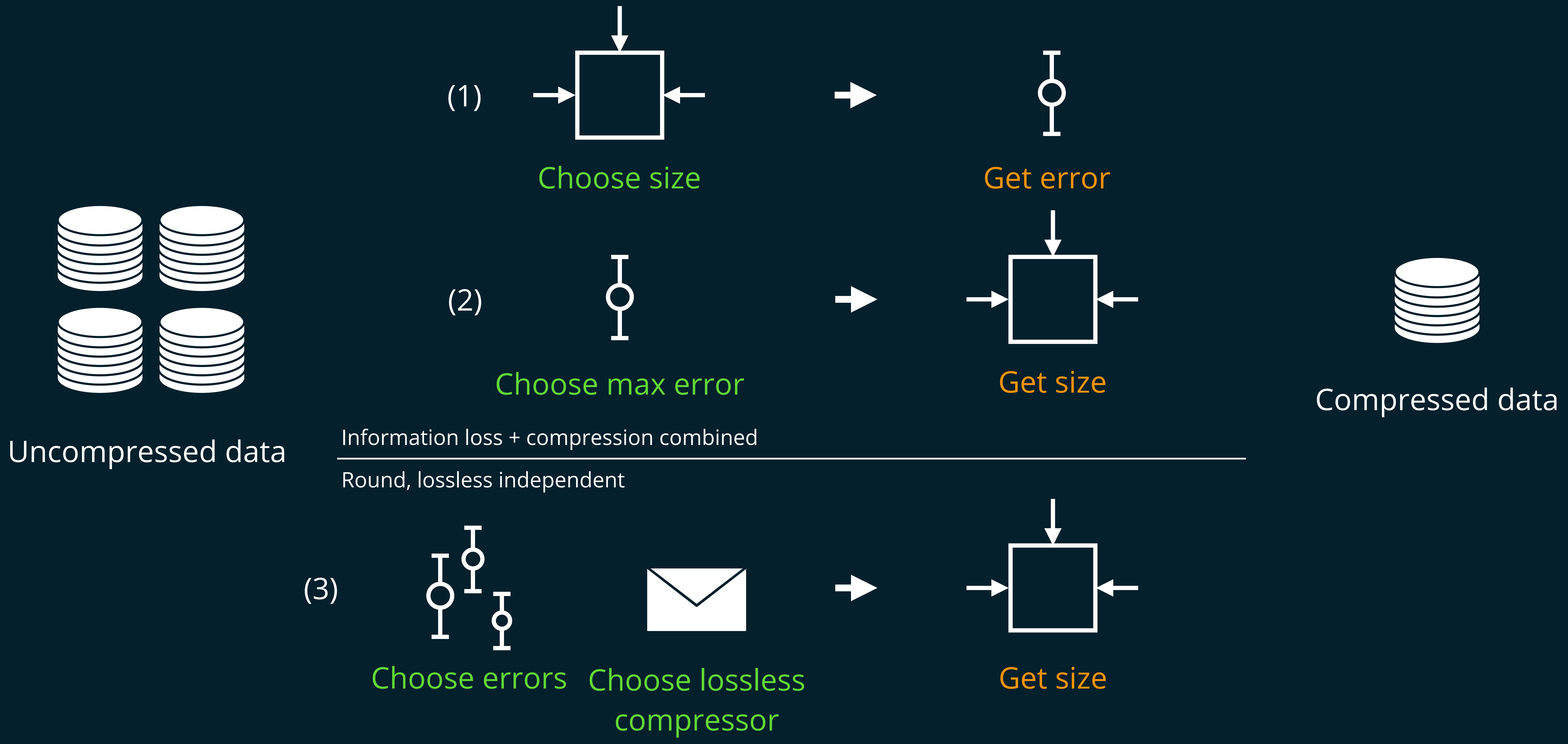
Most bits are effectively used

Floats are also approximately log-distributed

How much entropy is in the first bit?

~1bit

Data compression control knobs



Data compression: What do we want?

Case 1: Reanalysis data

Important
- Small
- Decompression speed
- Portability
- Random access

Less relevant
- Compression speed

Case 2: Research simulation

Important
- (De)compression speed
- Random access

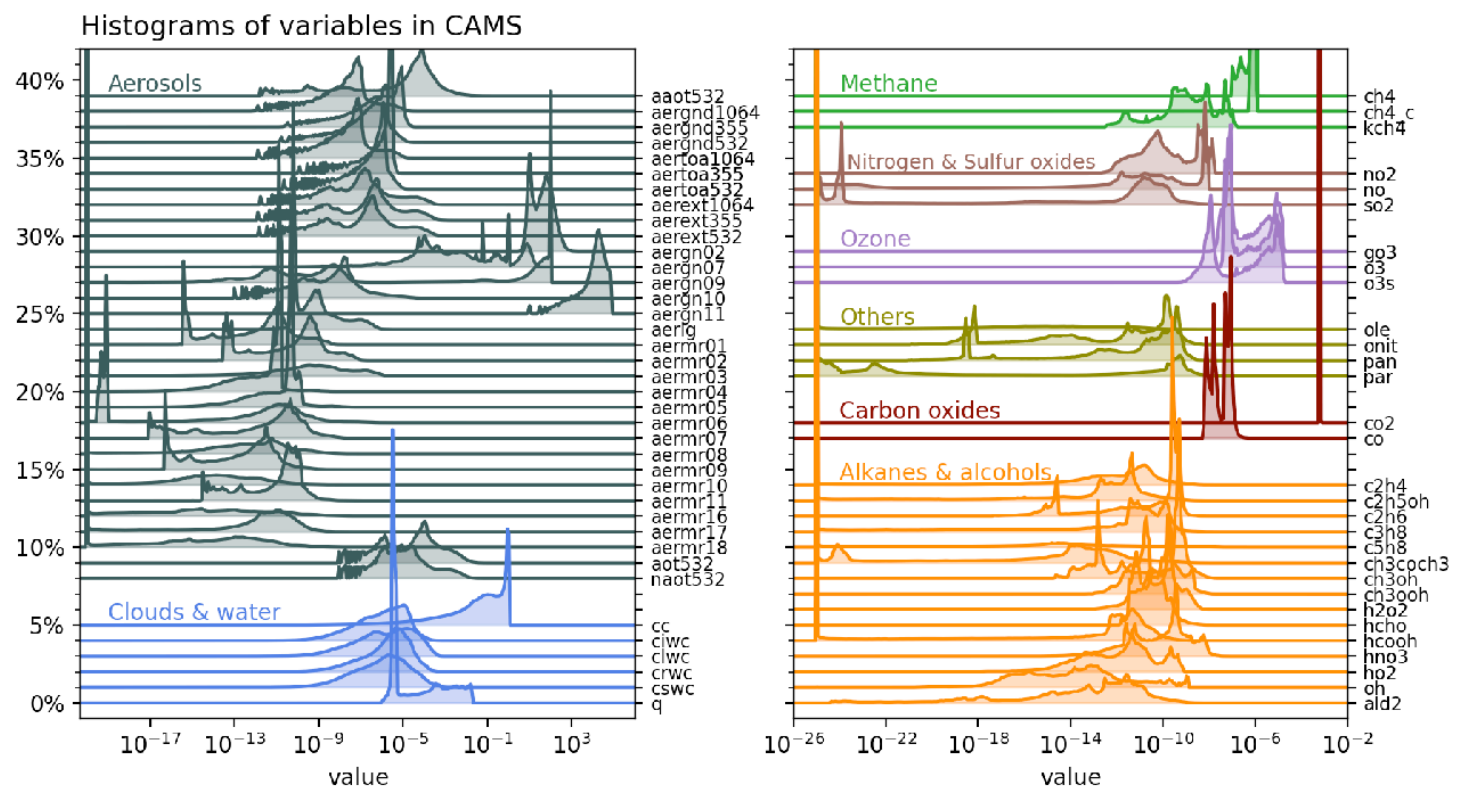
Less relevant
- Size

Case 3: Long storage

Important
- Size

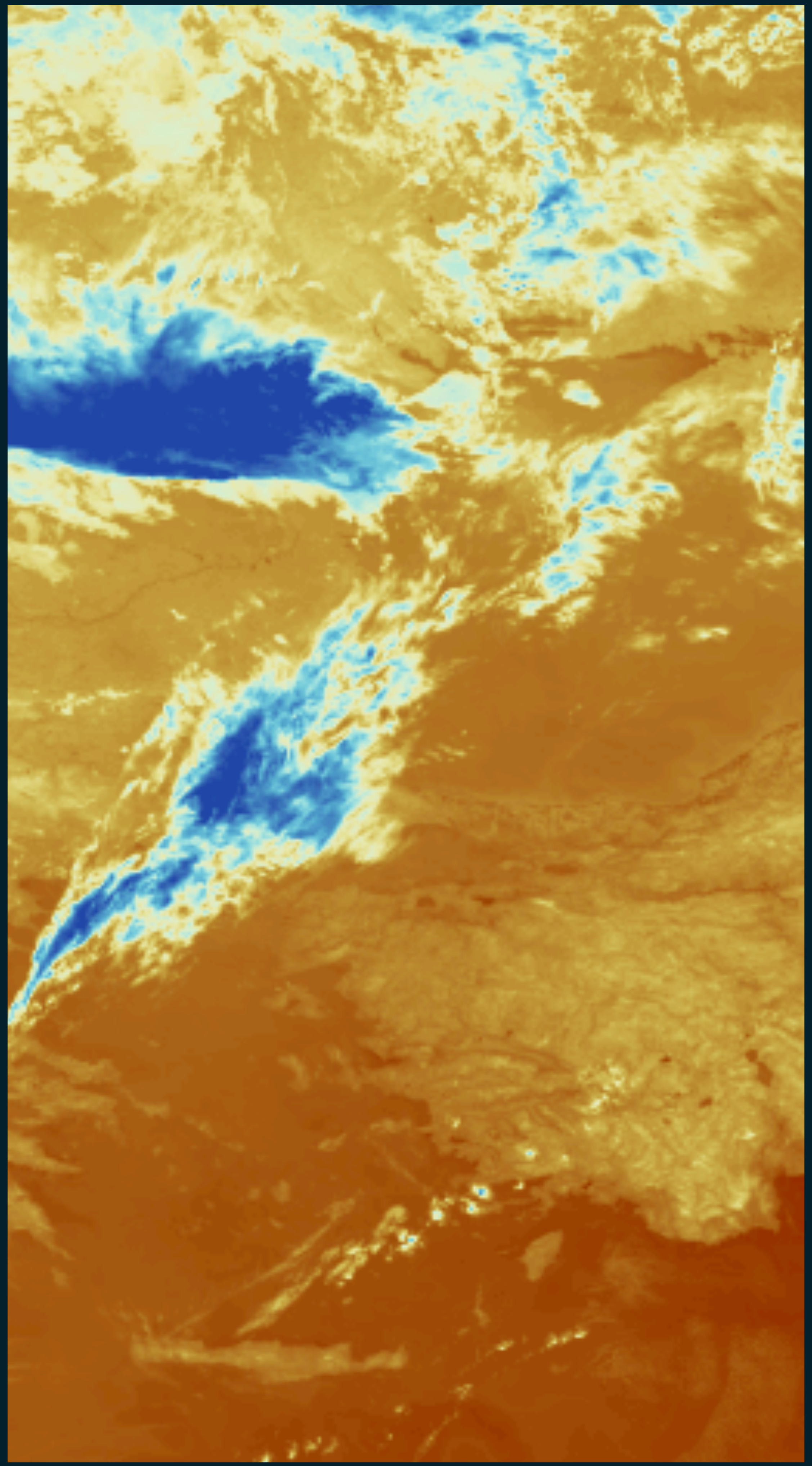
Less relevant
- Portability
- (De)compression speed
- Random acccess

What do we compress?

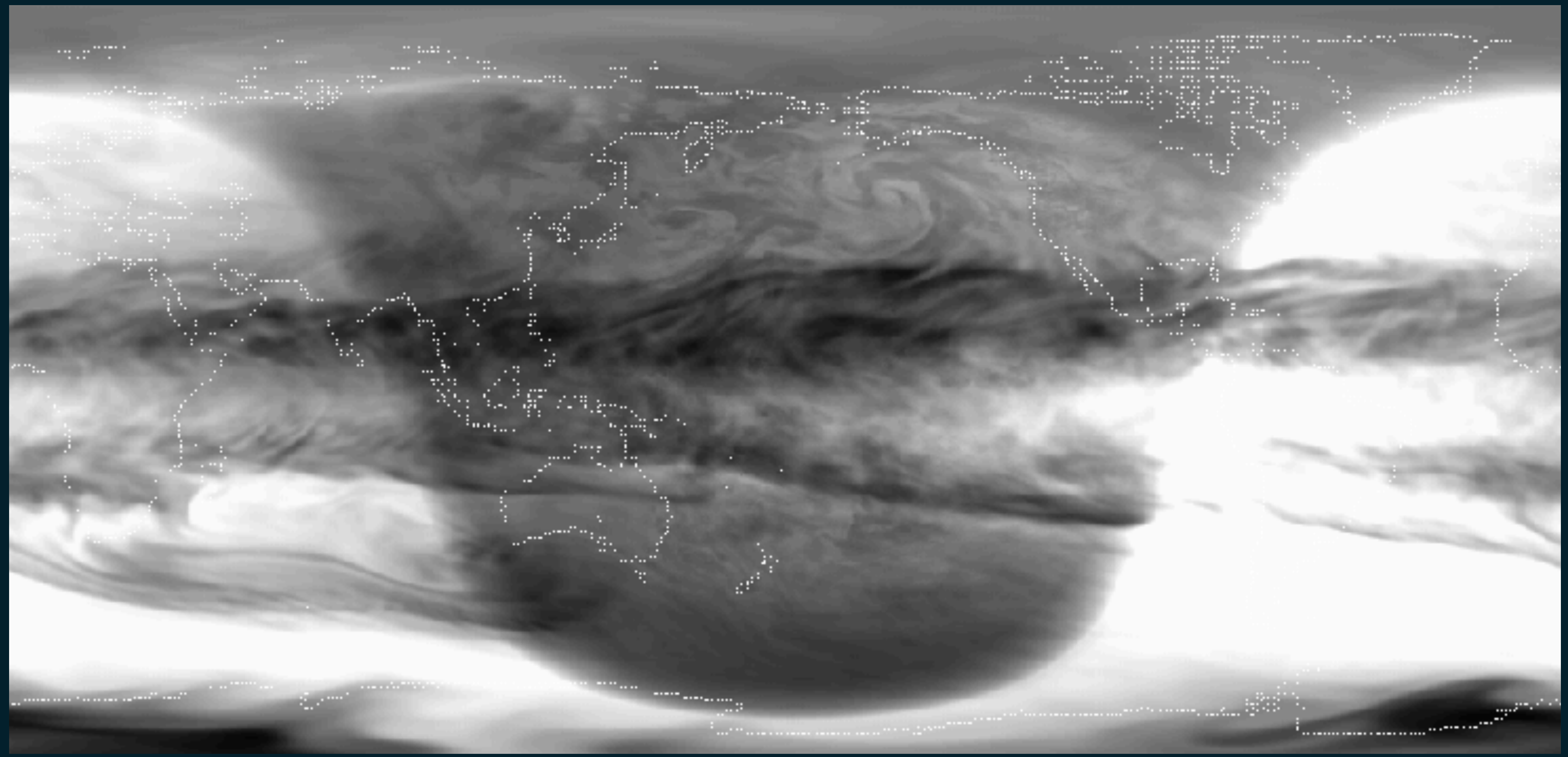


- **Many** different variables
- Varying uncertainties
- Linear or log-distributed
- Possibly many zeros

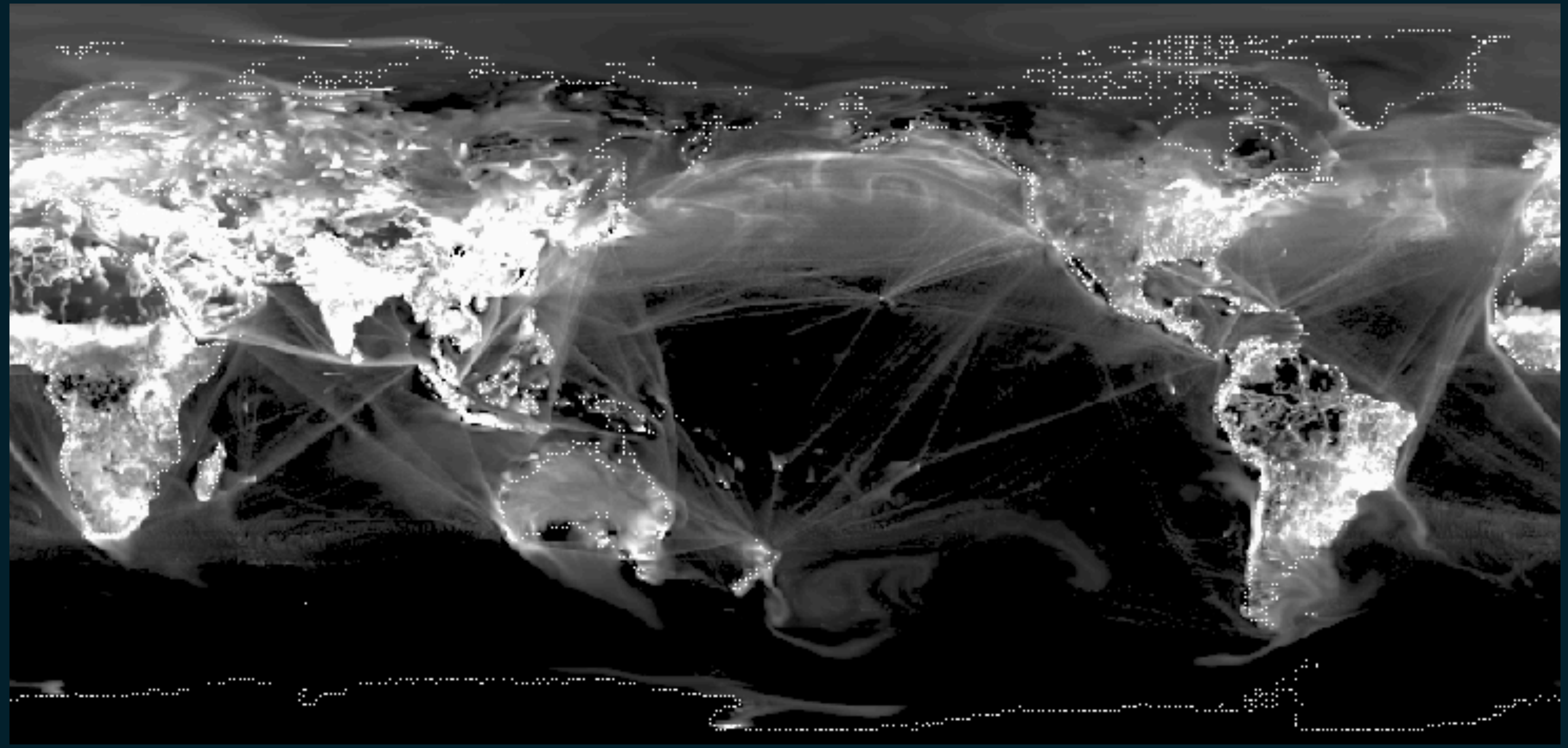
What do we compress?



Brightness temperature



NO_2 in the stratosphere



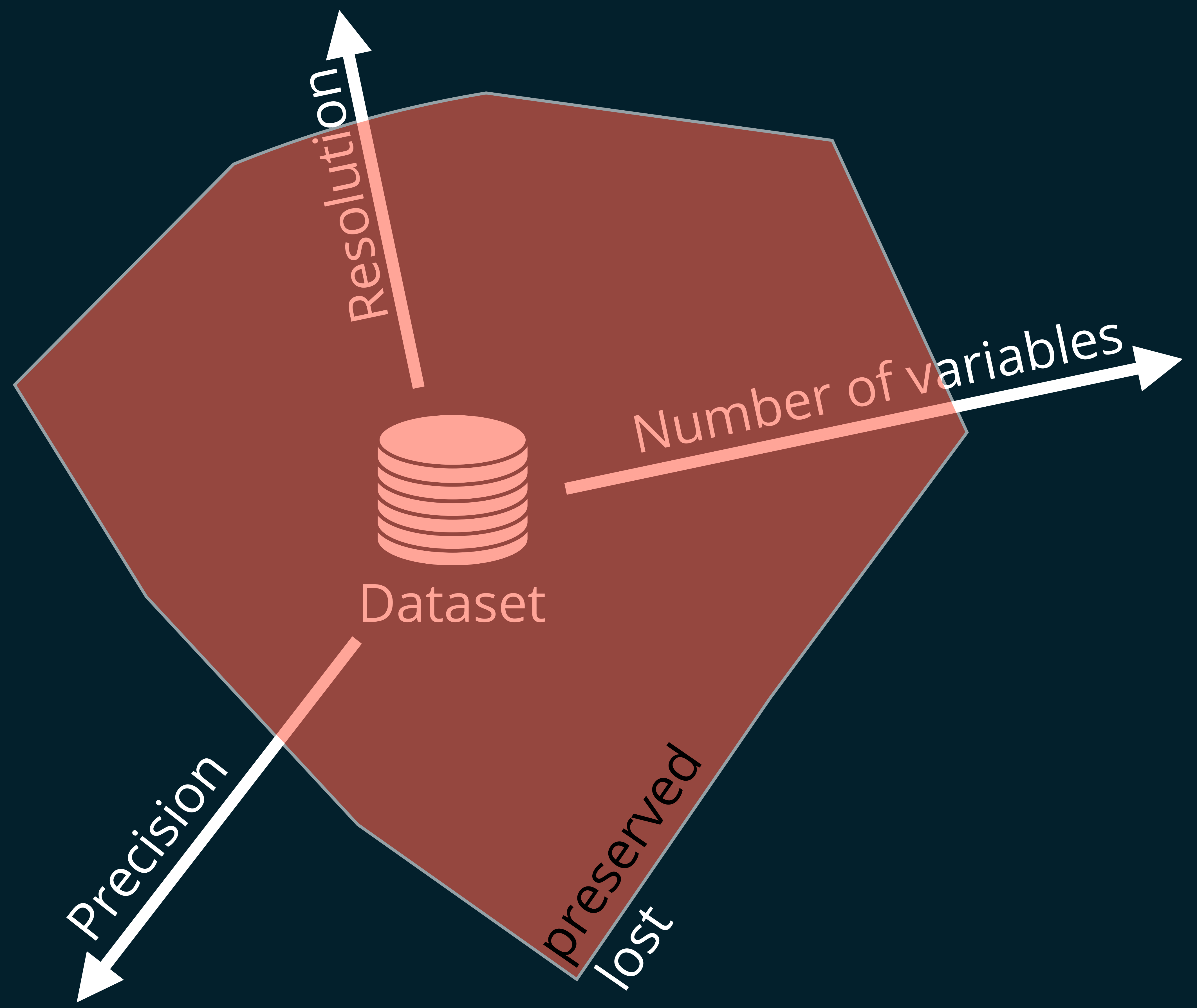
NO_2 at the surface

- **Many** different variables
- Varying uncertainties
- Linear or log-distributed
- Possibly many zeros

- Smoothness varies spatially
- Strong gradients
- Point sources

- Unstructured grids
- Spectral coefficients
- Masked data

What information is there in a dataset?



The *real information problem* of lossy data compression

$$F\left(\text{Smallest subset of Dataset}\right) = \text{Some answer}$$

Any question researchers ask

Qualitatively the same compared to full dataset

What compression error is okay?

Real!? ↓   False!? ↓

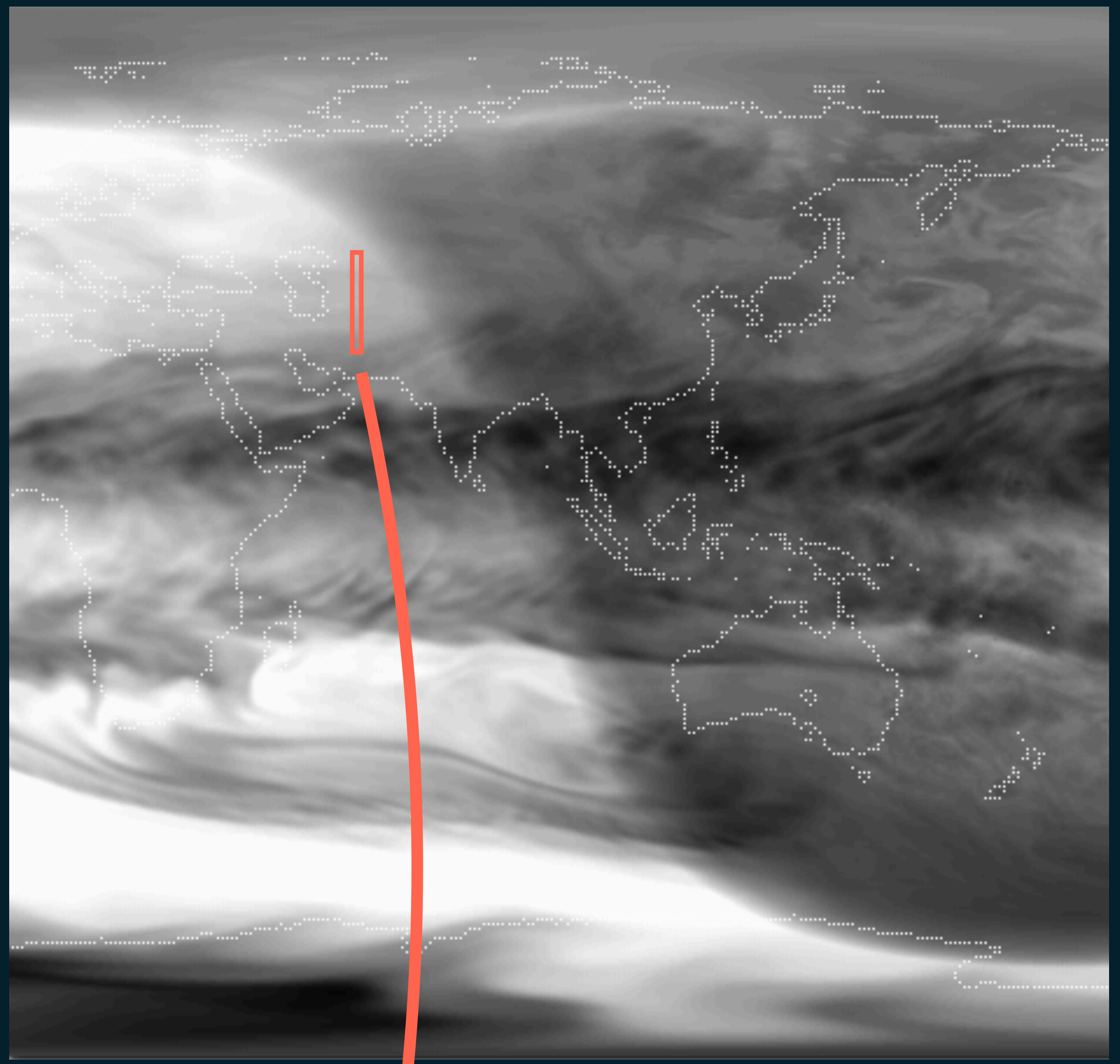
303.25 *unit* + uncertainty ⟶ **303.3 *unit* ?**
303 *unit* ?
300 *unit* ?

The acceptable error depends on

- Uncertainty, changes with:
- Variable and unit
- Application
- Model and its resolution
- …

Problem: What is the uncertainty in data and how can it be estimated if unknown?

What is real and false information in data?



Problem: What is the uncertainty in data and how can it be estimated if unknown?

Possible solution:
Find an objective way to separate real and false information!

```
0.050386034
0.050390966
0.05040059
0.050441727
0.05046302
0.05046855
0.050488267
0.05050127
0.050520953
0.05052939
0.050532646
```

Encoded in bits

Real!? False!?

```
00111101010011100110000110010110
00111101010011100110011011000010
00111101010011100111000011011001
00111101010011101001101111111100
00111101010011101011001001010000
00111101010011101011100000011100
00111101010011101100110011001001
00111101010011101101101001101011
```

Mutual information of adjacent grid points

…11010000000110000011000010000110000000100…

0 is mostly followed by a 0; 1 either remains 1 or switches back to 0.

Joint probability matrix: $\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = p_{rs} = \begin{pmatrix} 0.6 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$

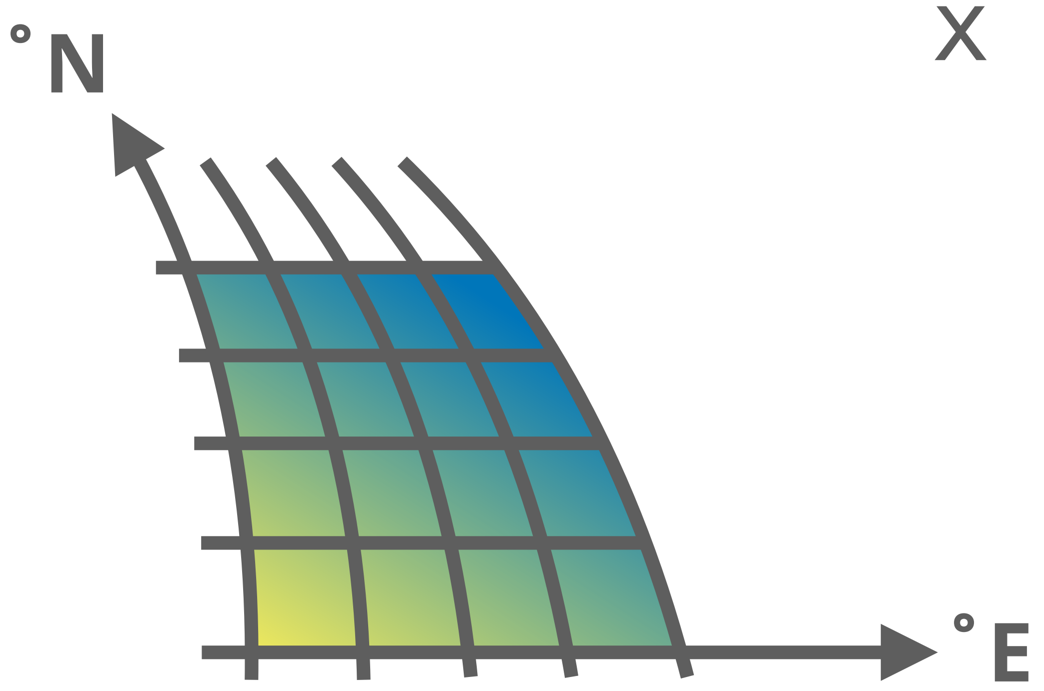
Mutual information:
What does one bit tells about the next?

$$M = \sum_r \sum_s p_{rs} \log_2 \left(\frac{p_{rs}}{p_{r=r} p_{s=s}} \right) = 0.2 \text{ bit}$$

Bitwise real information content

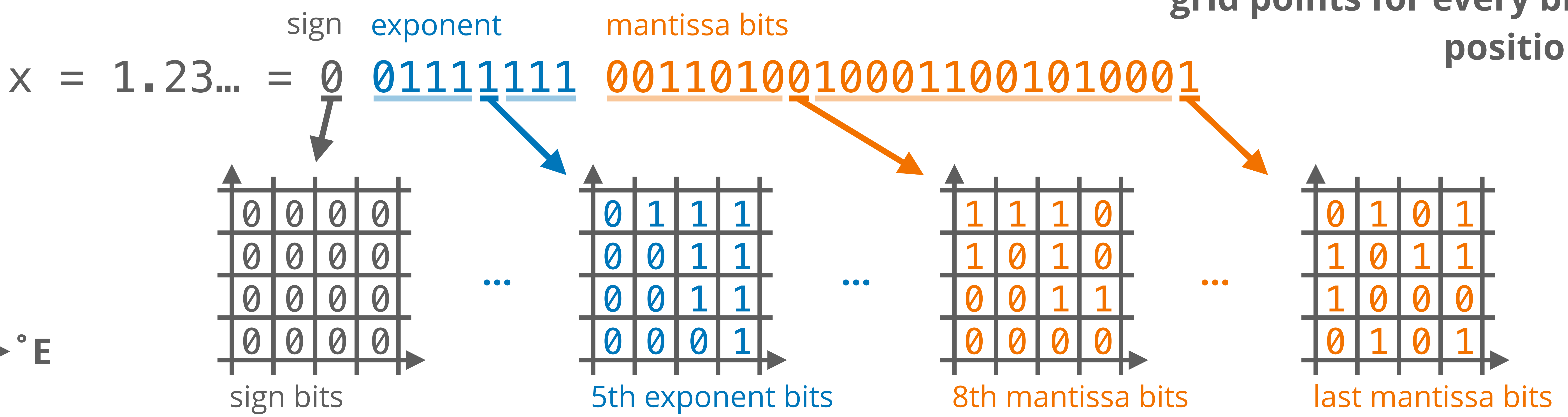
defined here as the **mutual bitwise information in adjacent grid points**

1 Gridded data



2 Data as binary floating-point numbers

3 Analyse bits in adjacent grid points for every bit position



4 The mutual information M between bits in adjacent grid points

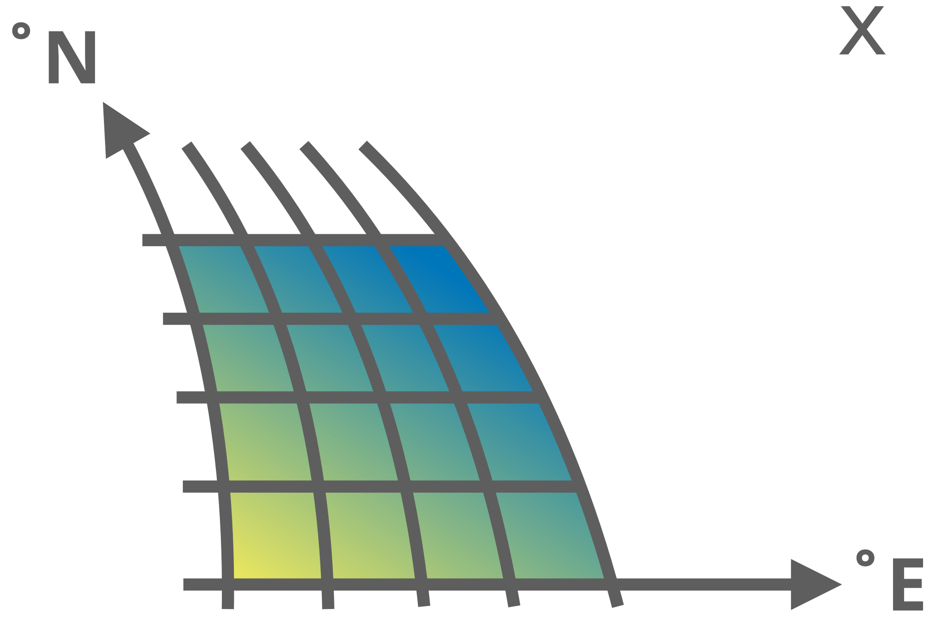
M = 0 bits
If all bits are **identical**

M ≈ 1 bit
If 0 is **certainly adjacent** to 0; 0 and 1 occur equally frequent

M > 0 bit
If 0 is **likely adjacent** to 0 and 1 is likely adjacent to 1

M = 0 bit
If adjacent bits are **independent**

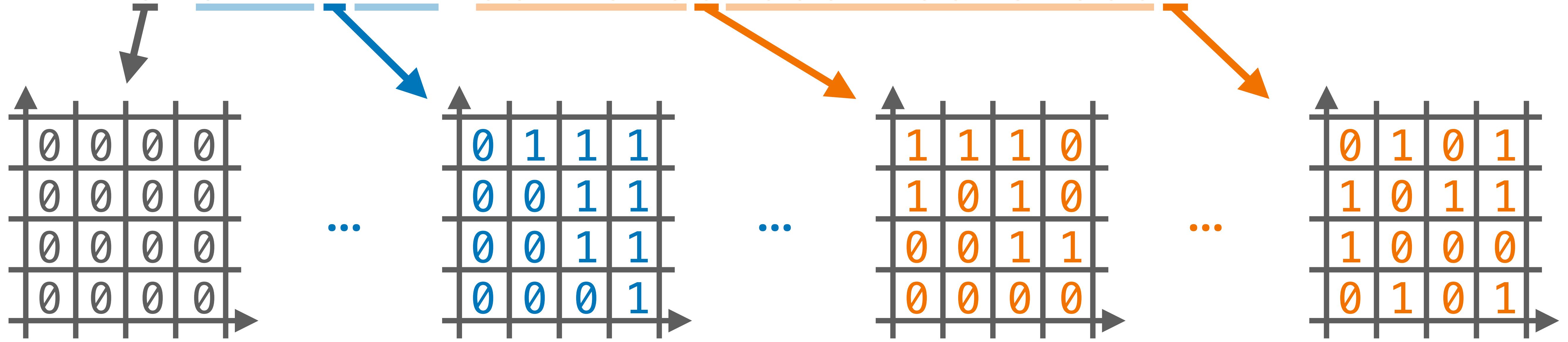
1 Gridded data



2 Data as binary floating-point numbers

sign exponent mantissa bits

x = 1.23… = 0 01111111 00110100100011001010001



3 Analyse bits in adjacent grid points for every bit position

4 The mutual information M between bits in adjacent grid points

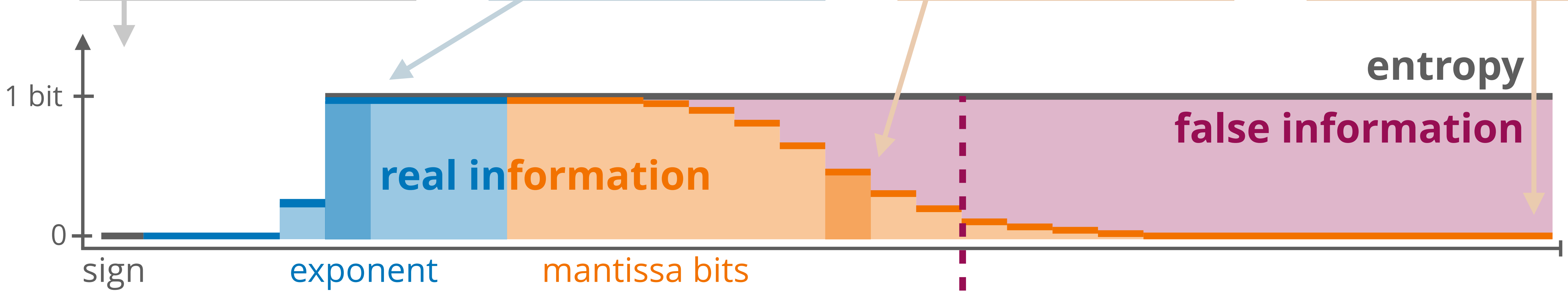
M = 0 bits
If all bits are **identical**

M ≈ 1 bit
If 0 is **certainly adjacent** to 0; 0 and 1 occur equally frequent

M > 0 bit
If 0 is **likely adjacent** to 0 and 1 is likely adjacent to 1

M = 0 bit
If adjacent bits are **independent**

5 Bitwise real information is the mutual information between bits



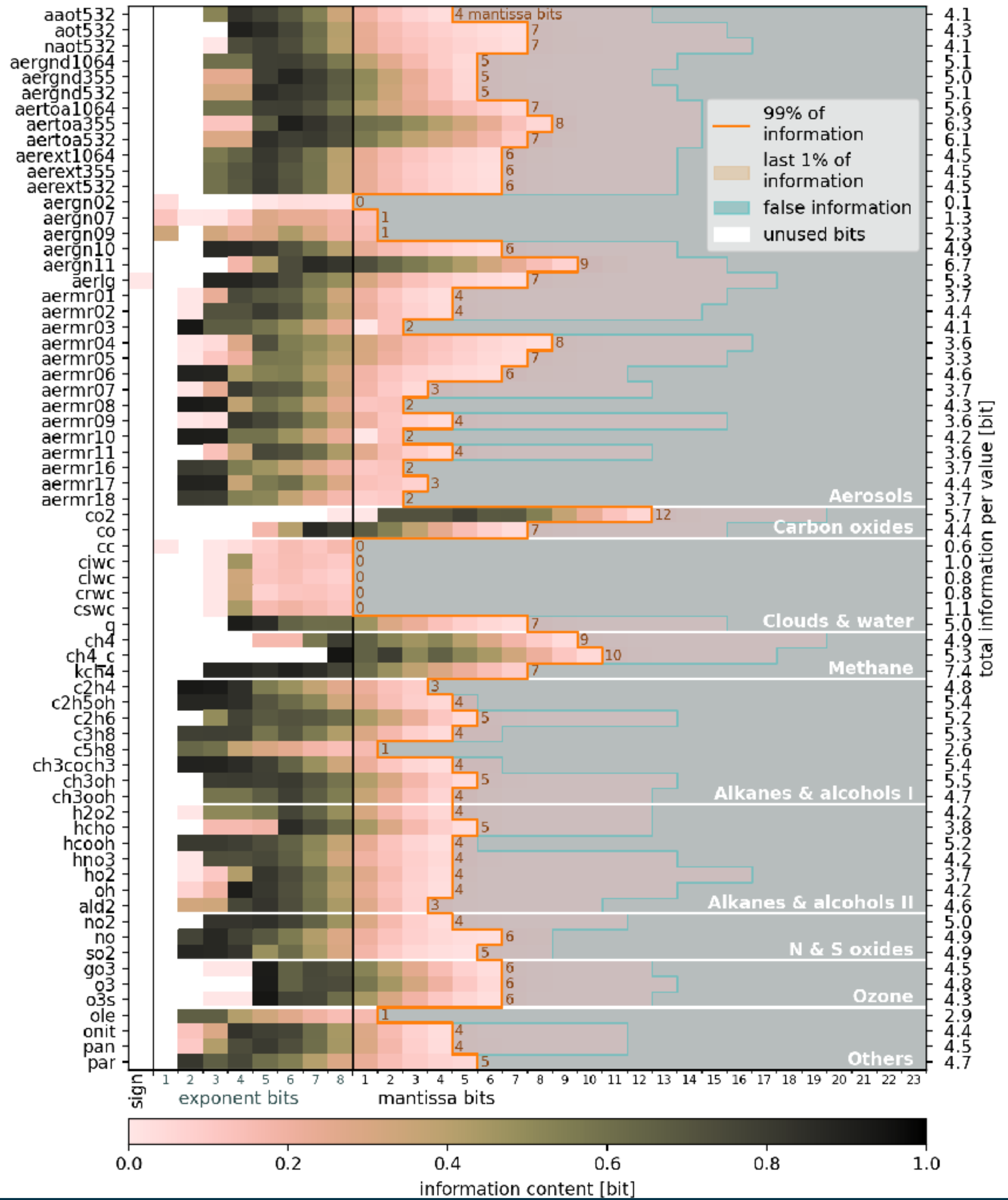
6 Rounding to remove the false information

x ≈ 0 01111111 00110100100000000000000

Retain bits with **>99% of real information** in total

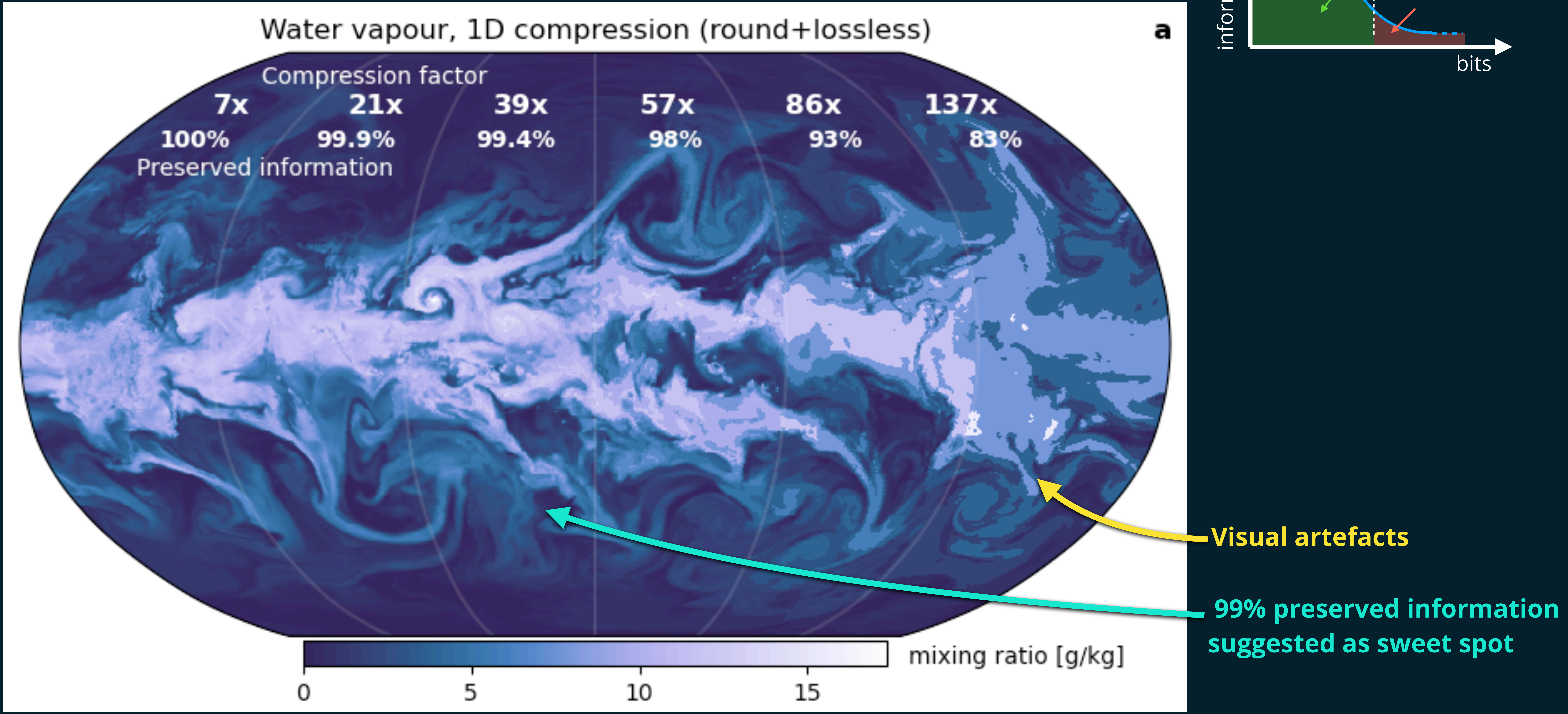
Remove false information by rounding trailing bits to 0

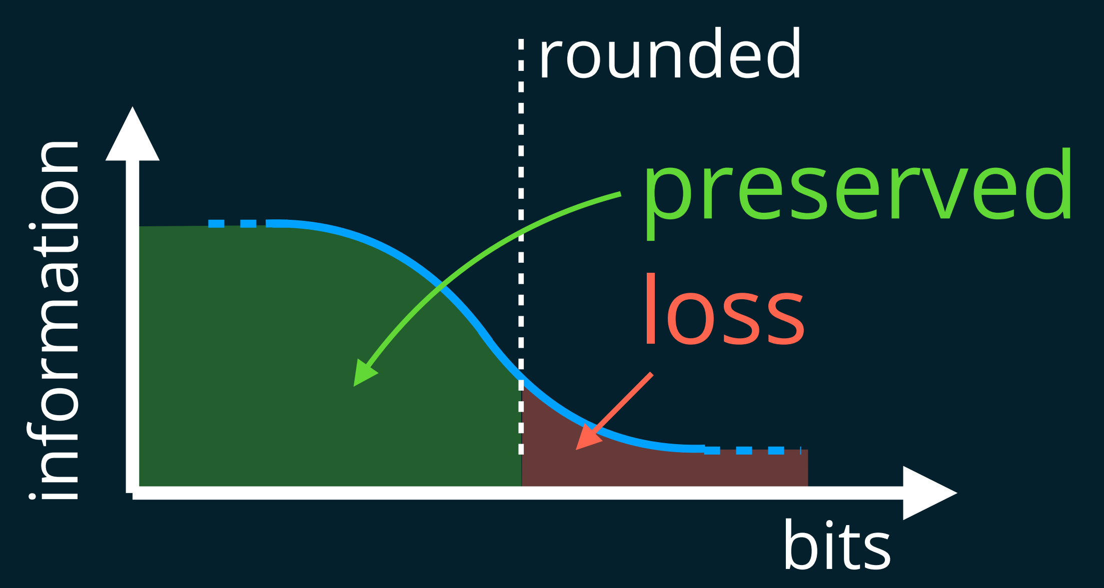
Bitwise
real information content



- Every variable requires a different precision
- Many bits do not contain real information
- Preserve only the bits with real information

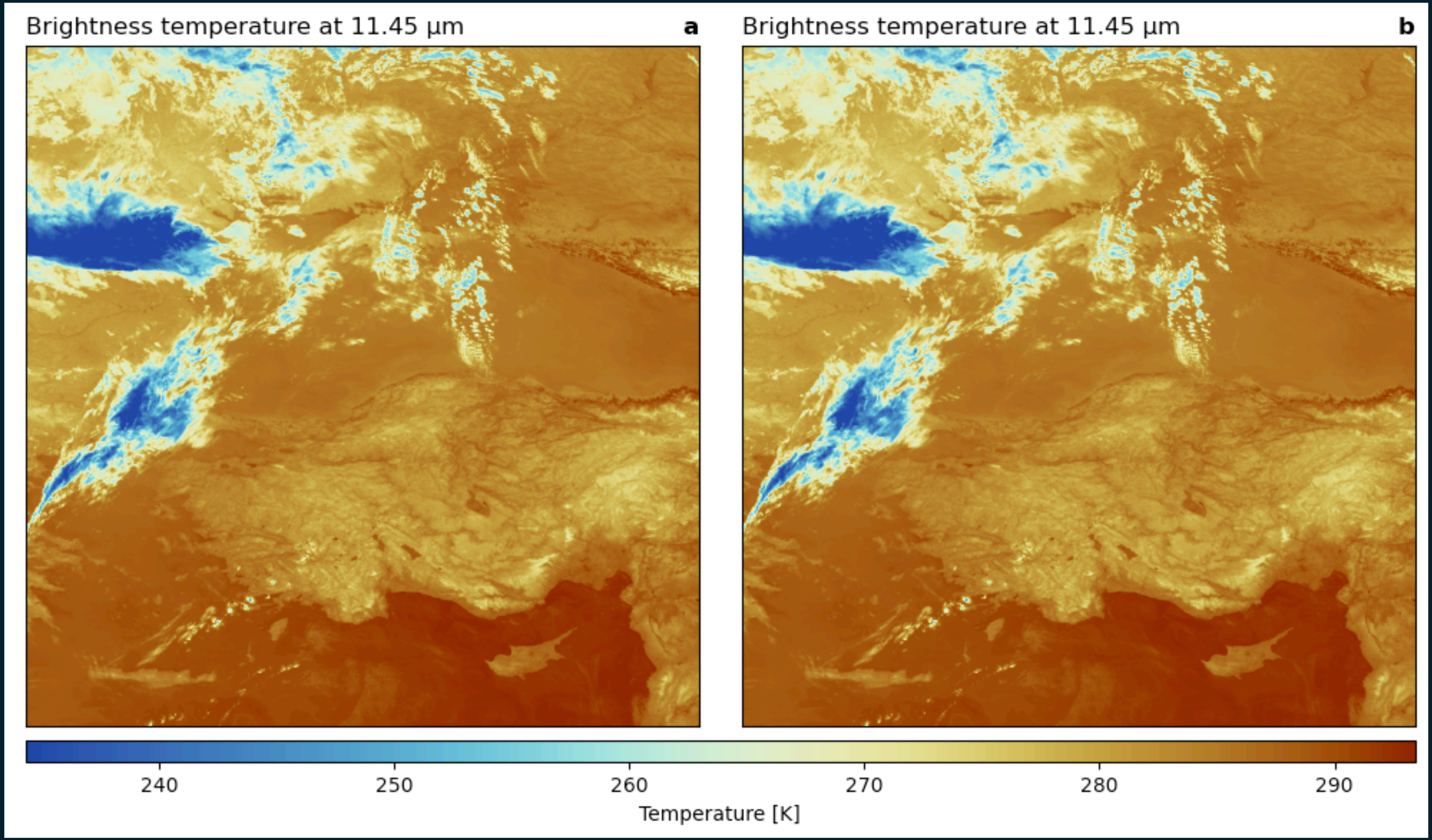
Information-preserving compression





Information-preserving compression

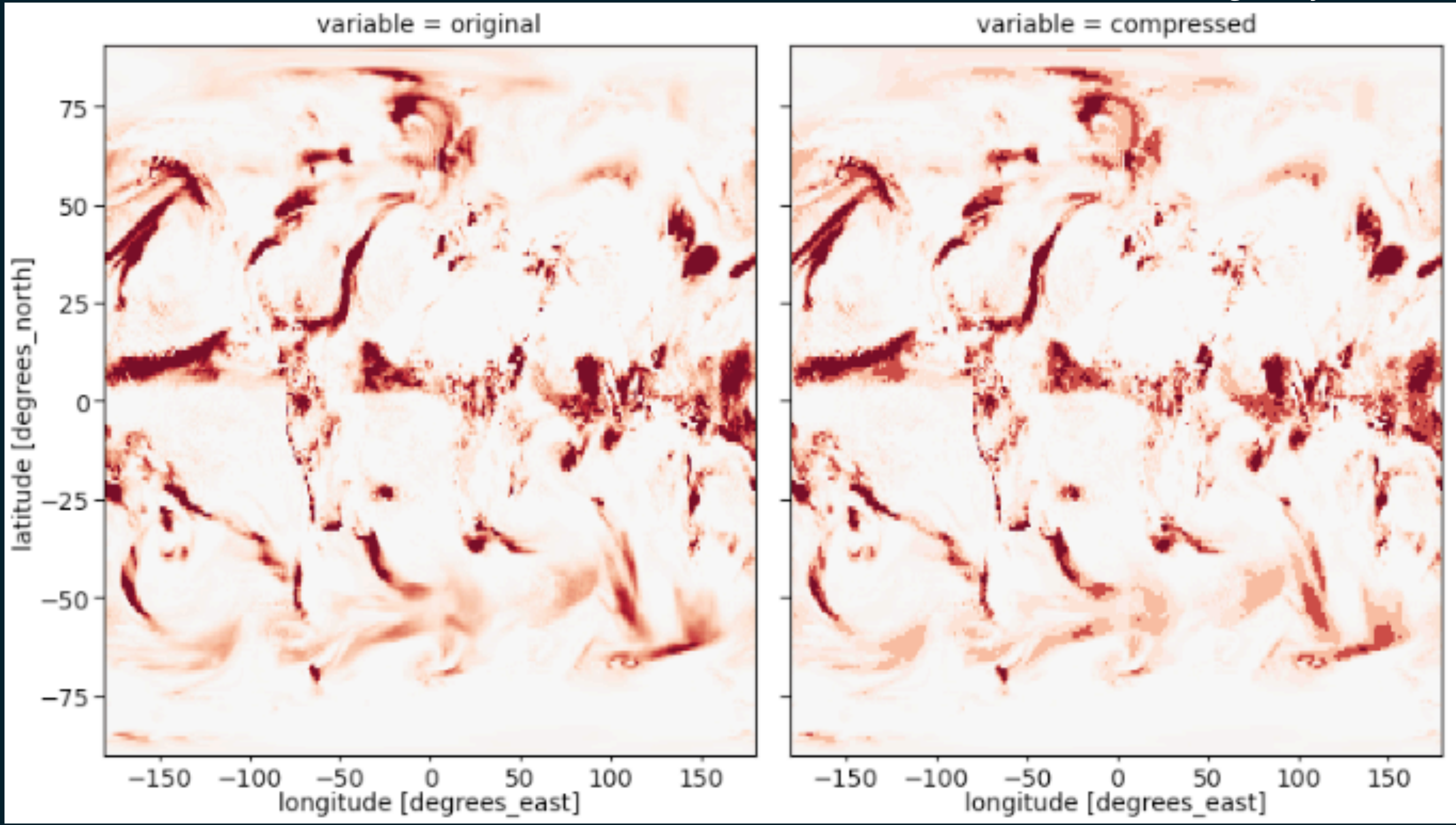
One is 15-20x smaller than the other, which has been compressed?



The ugly?

Precipitation

Aengenheyster 2022

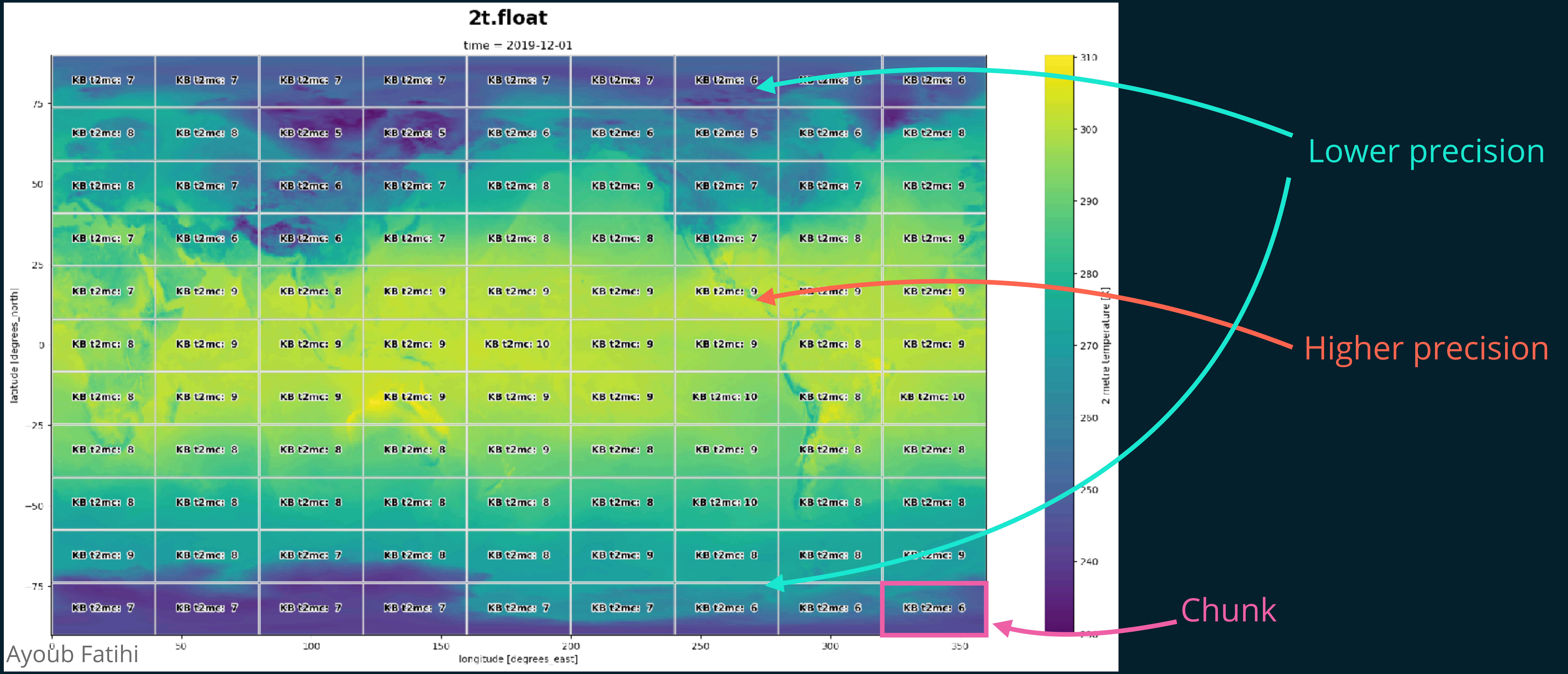


Problem: Both smooth and rough data

Solution: Independent information analysis and rounding by region

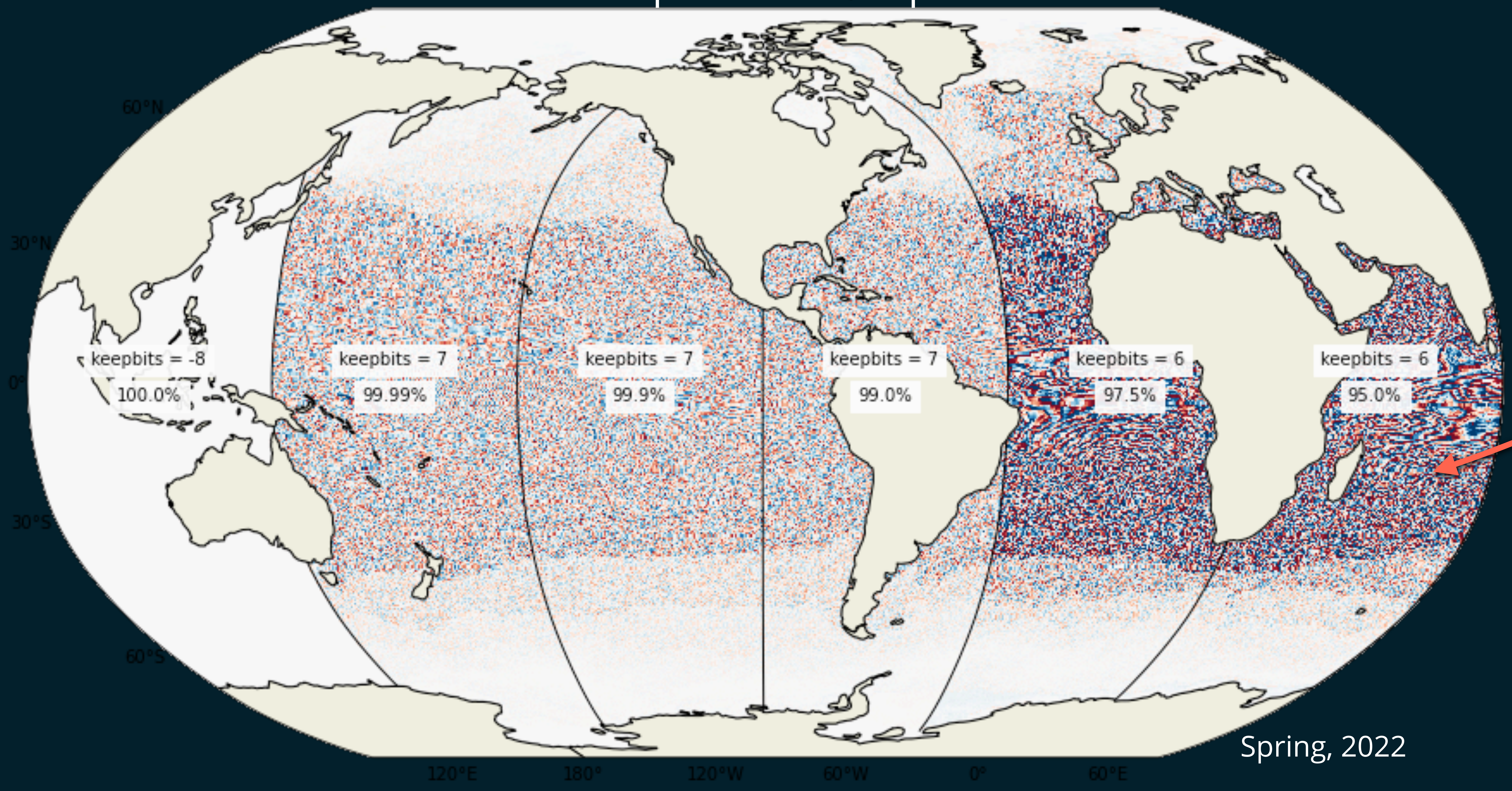
Variable-precision compression

Higher precision in one region, lower in another?



The ugly 2?

Sea surface temperature compression error



Compression error higher in tropics

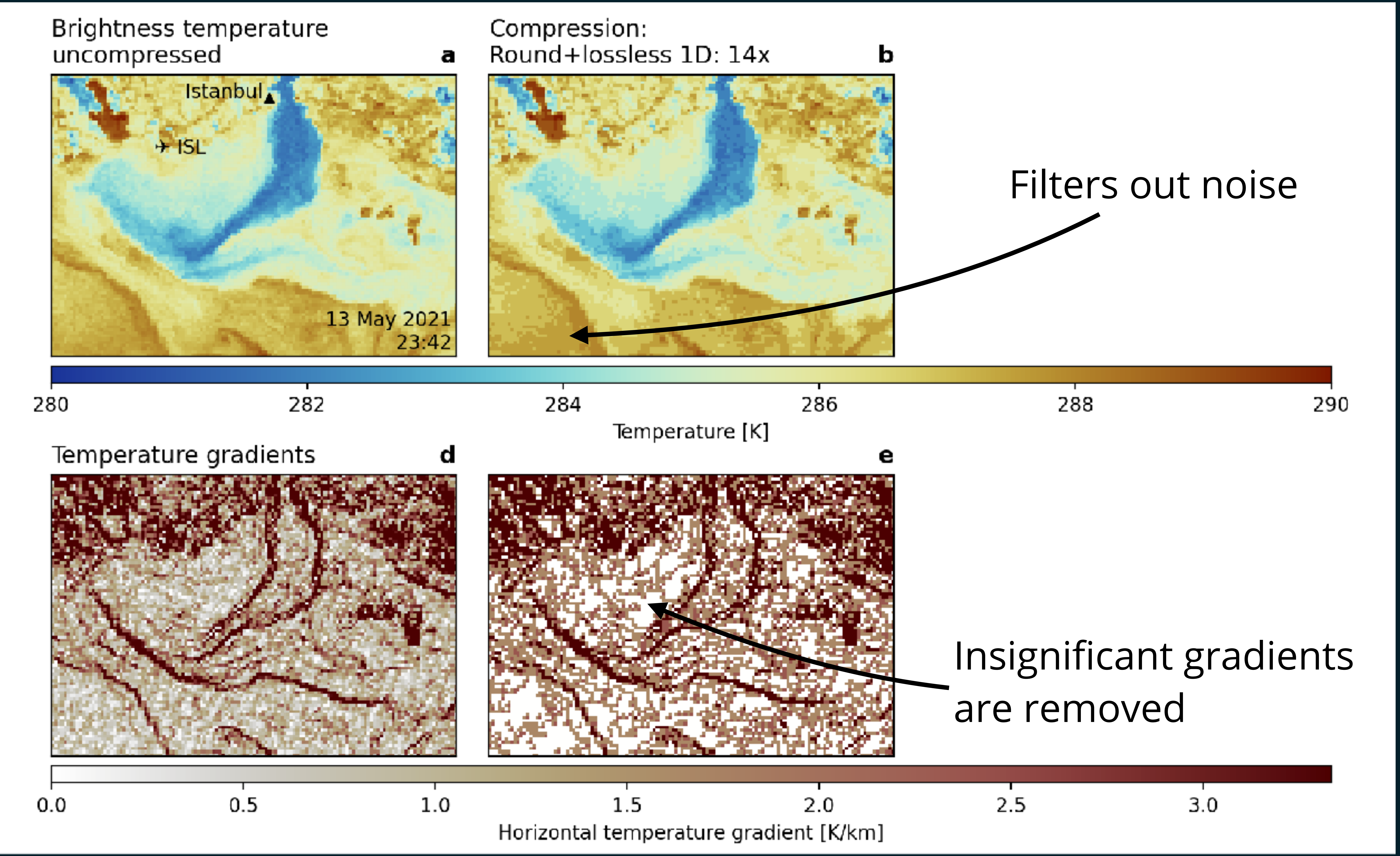
Spring, 2022

Problem: Absolute error vs log encoding

Solution: Adapt encoding to data or variable keep bits

The bad? What about gradients!?

Filtering out insignificant gradients

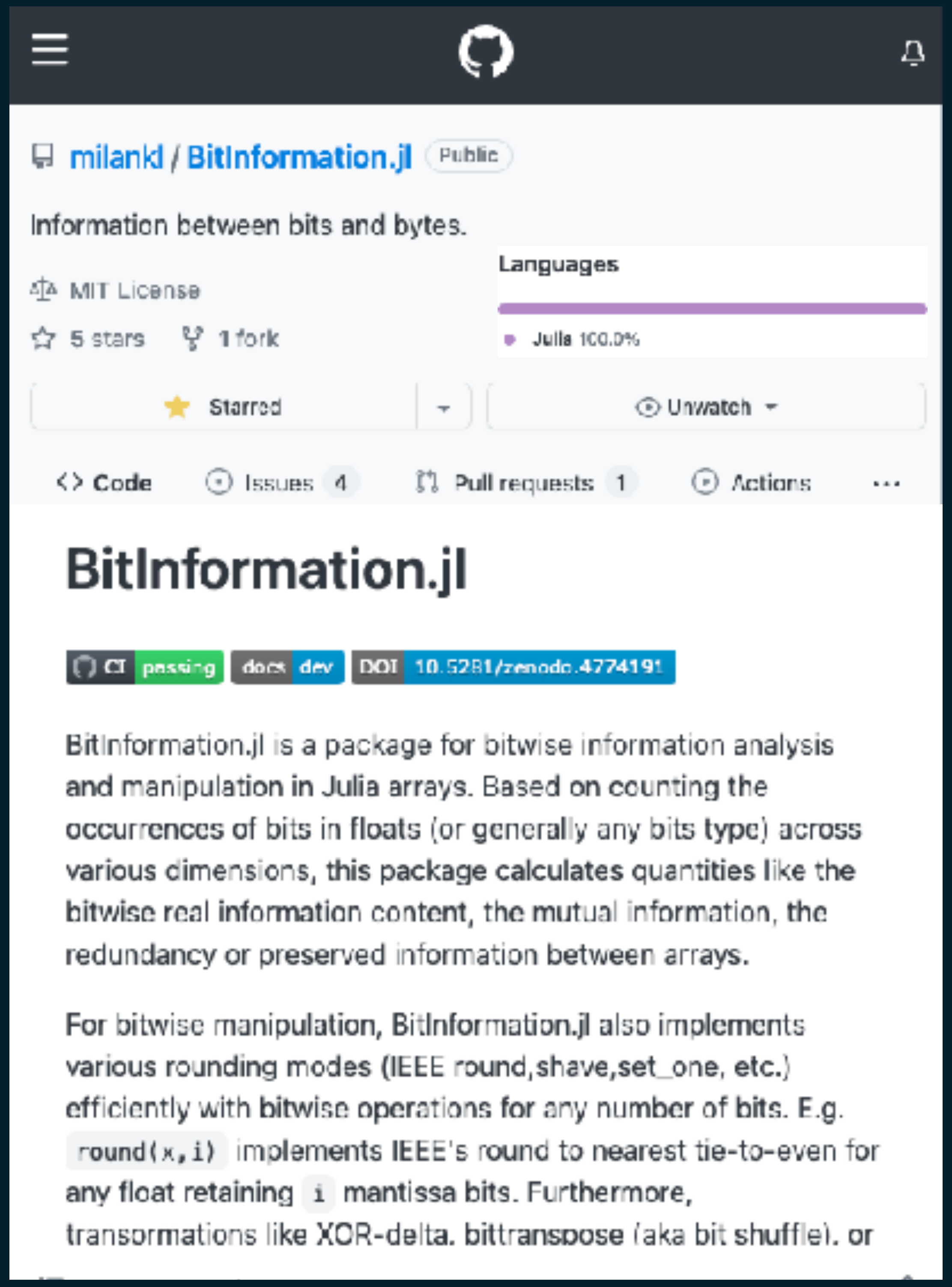


Workflow

1. Information analysis
   - Only once offline
   - Yields acceptable error bounds
2. Rounding
   - Very fast
   - Variable precision possible
3. Lossless compression
   - Any codec can be used
   - Flexible size/performance trade-off
   - Chunking possible

Implementations

Julia: BitInformation.jl

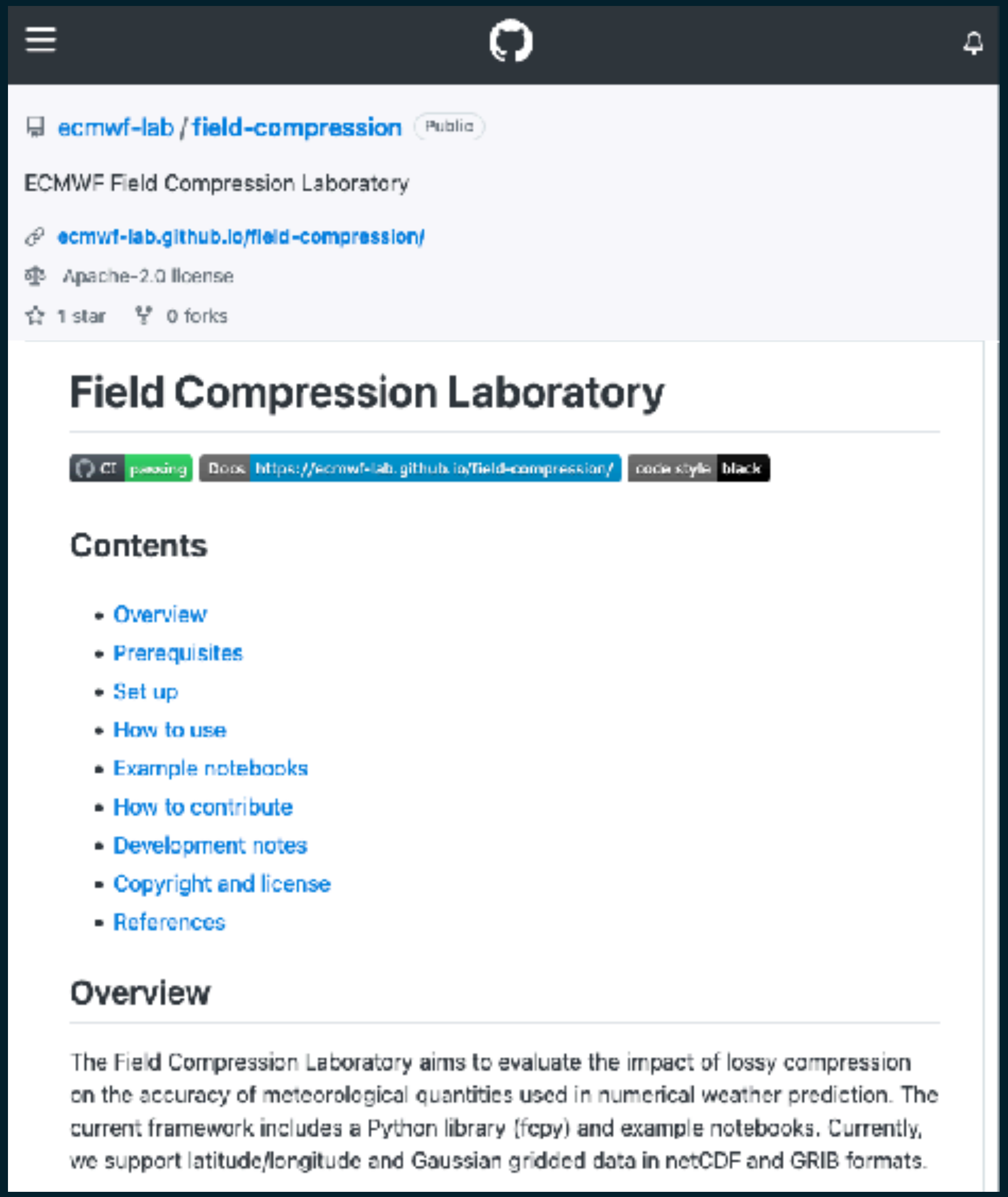


Milan Klöwer

Python & xarray: xbitinfo

Schulz, Spring

Python, netCDF+GRIB

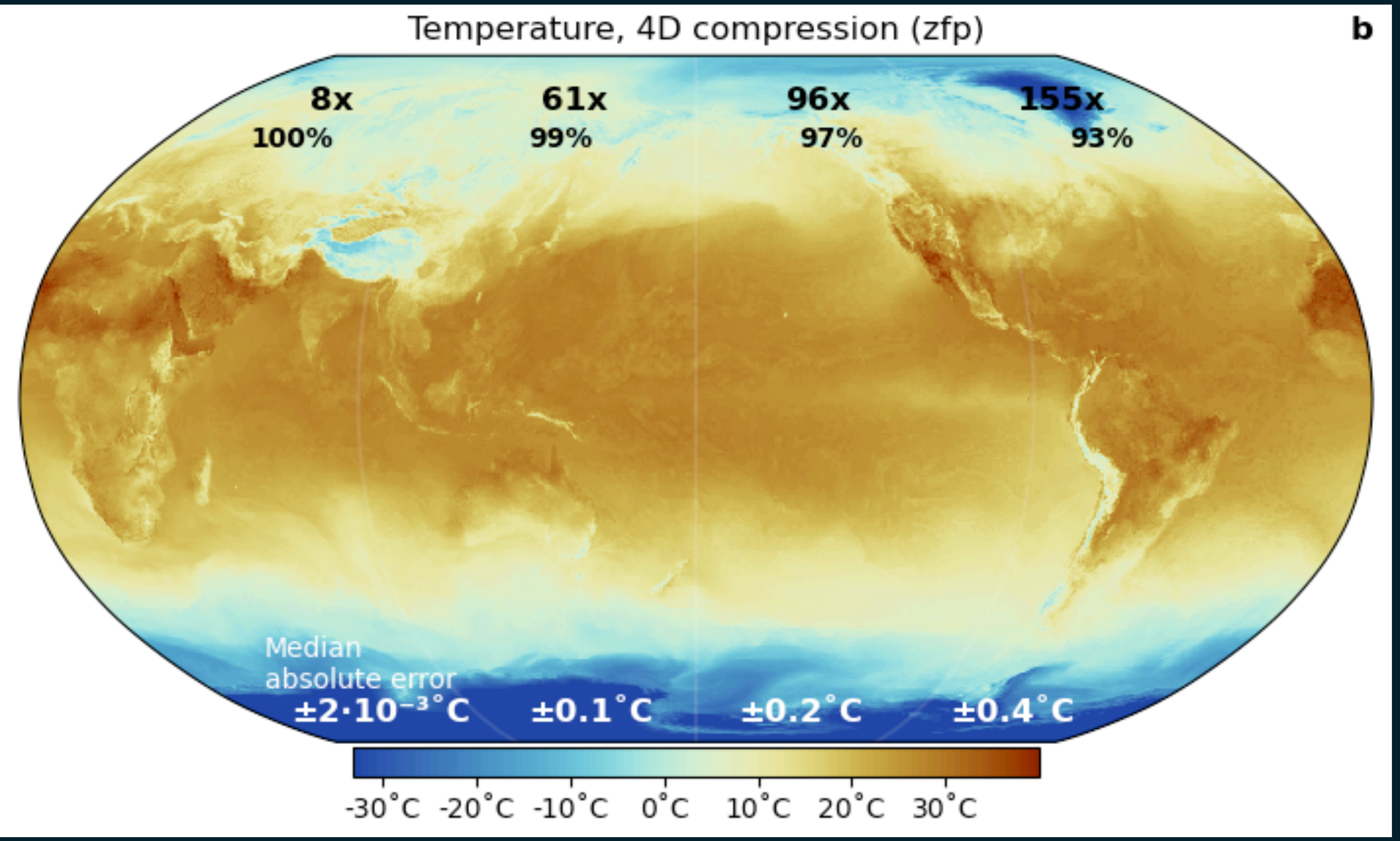


David Meyer

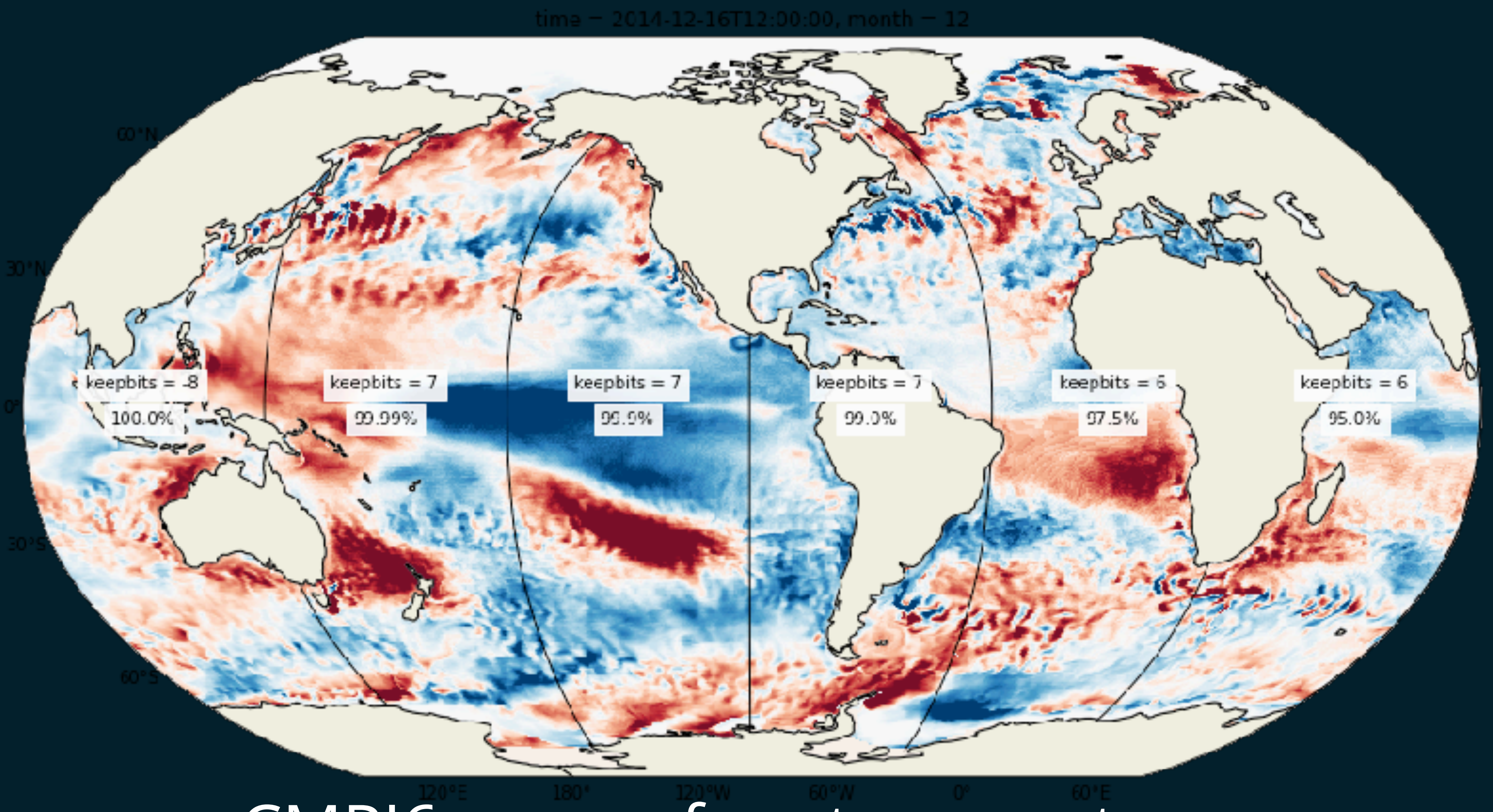
netCDF+HDF: NCO

Zender et al

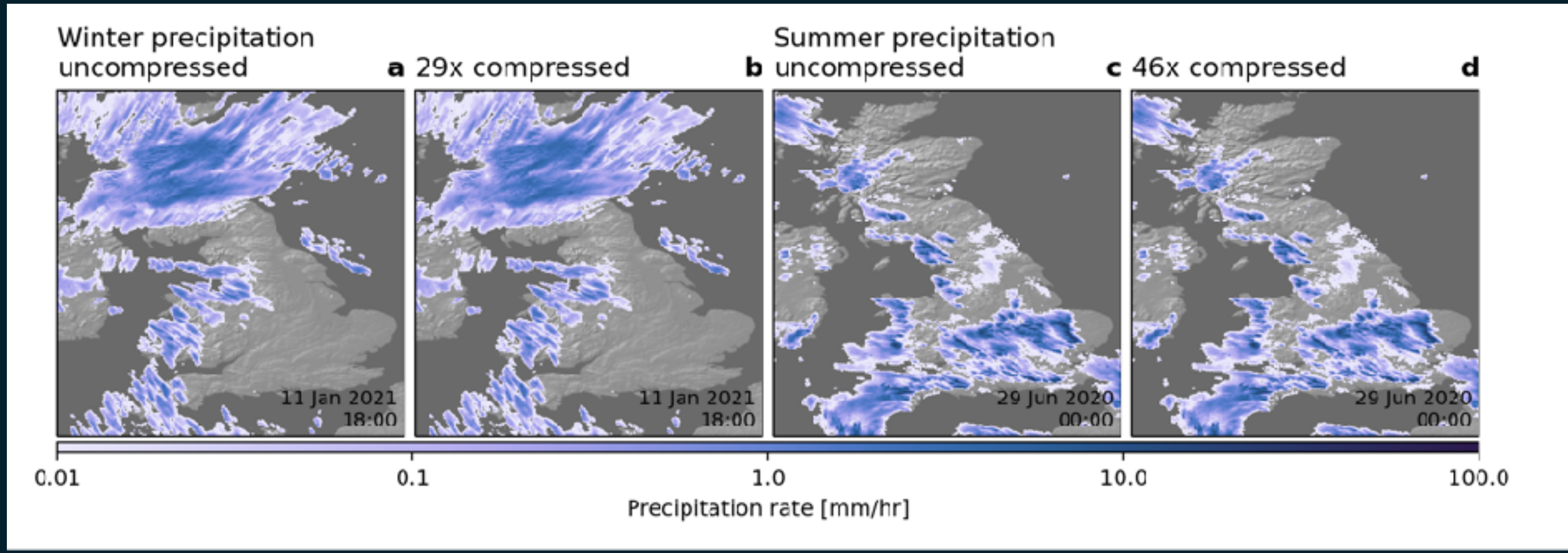
Application examples



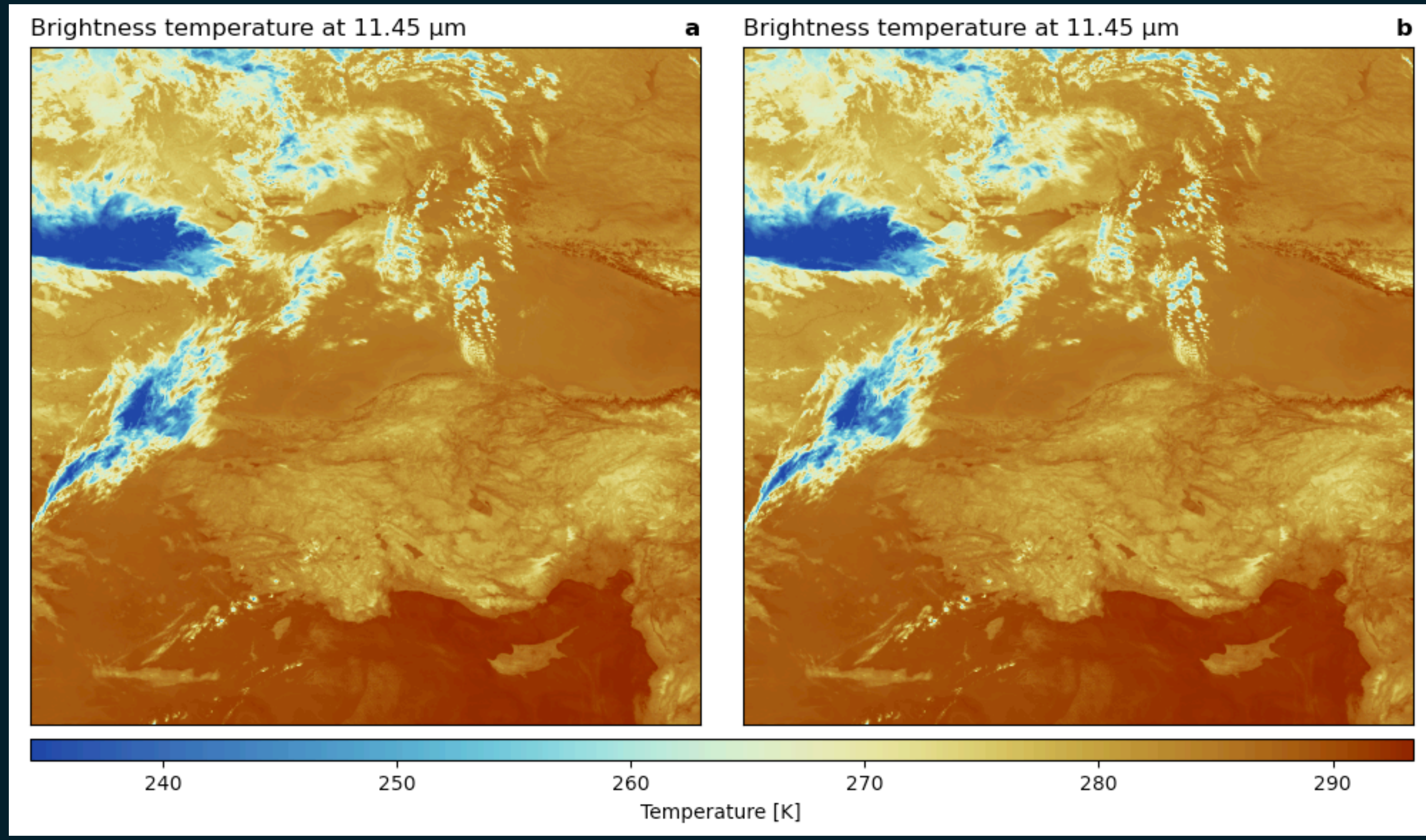
ECMWF’s ensemble temperature forecast



CMPI6 sea surface temperatures and ICON-Ocean model output



Precipitation from radar data



Brightness temperature from satellites