Bit-Grooming: Shave Your Bits with Razor-sharp Precision

Charles Zender[1] and Jeremy Silver[2]

[1]Affiliation not available
[2]University of Melbourne

February 23, 2018

Abstract

Lossless compression can reduce climate data storage by 30-40%. In general, further reductions require lossy compression that also reduces precision. Fortunately, geoscientific models and measurements generate false precision (scientifically meaningless data bits) that can be eliminated without sacrificing scientifically meaningful data. We introduce Bit Grooming, a lossy compression algorithm that removes the bloat due to false- precision, those bits and bytes beyond the meaningful precision of the data. We evaluated Bit Grooming against competitors Linear Packing, Layer Packing, and GRIB2/JPEG2000.

Bit Grooming: Shave Your Bits with Razor-sharp Precision

Charlie Zender[1] <zender@uci.edu> and Jeremy D. Silver[2]

[1]Departments of Earth System Science & Computer Science, UC Irvine, [2]University of Melbourne

Why Lossy Compression?

Lossless compression can reduce climate data storage by 30-40%. In general, further reductions require lossy compression that also reduces precision. Fortunately, geoscientific models and measurements generate false precision (scientifically meaningless data bits) that can be eliminated without sacrificing scientifically meaningful data. We introduce Bit Grooming, a lossy compression algorithm that removes the bloat due to false-precision, those bits and bytes beyond the meaningful precision of the data. We evaluated Bit Grooming against competitors Linear Packing, Layer Packing, and GRIB2/JPEG2000.

Bit Grooming Algorithm

- Alternately shave (to 0) and set (to 1) least significant bits of consecutive values
- Symmetric, two-sided variant of Bit Shaving algorithm that solely zeroes bits
- Alternation eliminates artificial low-bias produced by always zeroing bits
- Implemented as bit-mask, no floating-point arithmetic (or rounding) required
- Bit Grooming preserves any requested Number of Significant Digits (NSD):

Sign[a]	Exponent[b]	Significand[c]	Decimal	Notes
0	10000000	10010010000111111011011	3.14159265	Exact π
0	10000000	10010001111010111000011	3.14000000	Three significant digits
0	10000000	10010010000000000000000	3.14062500	DSD = 2 (Decimal Rounding)
0	10000000	10010010000000000000000	3.14062500	NSD = 3 (Bit Shaving)[d]
0	10000000	10010010000111111111111	3.14160132	NSD = 3 (Bit Setting)

- Bit-Groomed data compresses well with standard lossless algorithms (DEFLATE)
- More accurate, greater range, less compression than packing (netCDF default)
- Unlike all viable competitors, BG guarantees specified precision for all data
- Preserves IEEE floating-point format—no special software required to read

Bit-Grooming Pi

Sign	Exponent	Fraction (significand)	Decimal	Notes
0	10000000	10010010000111111011011	3.14159265	Exact
0	10000000	10010010000111111011011	3.14159265	NSD = 8
0	10000000	10010010000111111011010	3.14159262	NSD = 7
0	10000000	10010010000111111011000	3.14159203	NSD = 6
0	10000000	10010010000111111000000	3.14158630	NSD = 5
0	10000000	10010010000111100000000	3.14154053	NSD = 4
0	10000000	10010010000000000000000	3.14062500	NSD = 3
0	10000000	10010010000000000000000	3.14062500	NSD = 2
0	10000000	10010000000000000000000	3.12500000	NSD = 1

Accuracy

Bit Grooming (BG) is, unlike Bit Shaving (BS), *statistically unbiased*:

			Artificial data[a]			
	BG and BS[c]		BGSP	BSSP	BGDP	BSDP
NSD[d]	ϵ^{+}_{max}	$\bar{\epsilon}^{+}$	$\bar{\epsilon}$	$\bar{\epsilon}$	$\bar{\epsilon}$	$\bar{\epsilon}$
1	0.31	0.11	4.1×10^{-4}	−0.11	4.0×10^{-4}	−0.11
2	0.39	0.14	6.8×10^{-5}	−0.14	5.5×10^{-5}	−0.14
3	0.49	0.17	1.0×10^{-6}	−0.17	-5.5×10^{-7}	−0.17
4	0.30	0.11	3.2×10^{-7}	−0.11	-6.1×10^{-6}	−0.11
5	0.37	0.13	3.1×10^{-7}	−0.13	-5.6×10^{-6}	−0.13
6	0.36	0.12	-4.4×10^{-7}	−0.12	-4.1×10^{-7}	−0.17
7	0.00	0.00	0.0	0.00	1.5×10^{-7}	−0.10

Fig 1: Bit Grooming (NSD*) compression ratio (larger is better) is intermediate between lossless (DEFLATE) and other lossy (LIN, LAY) compression:

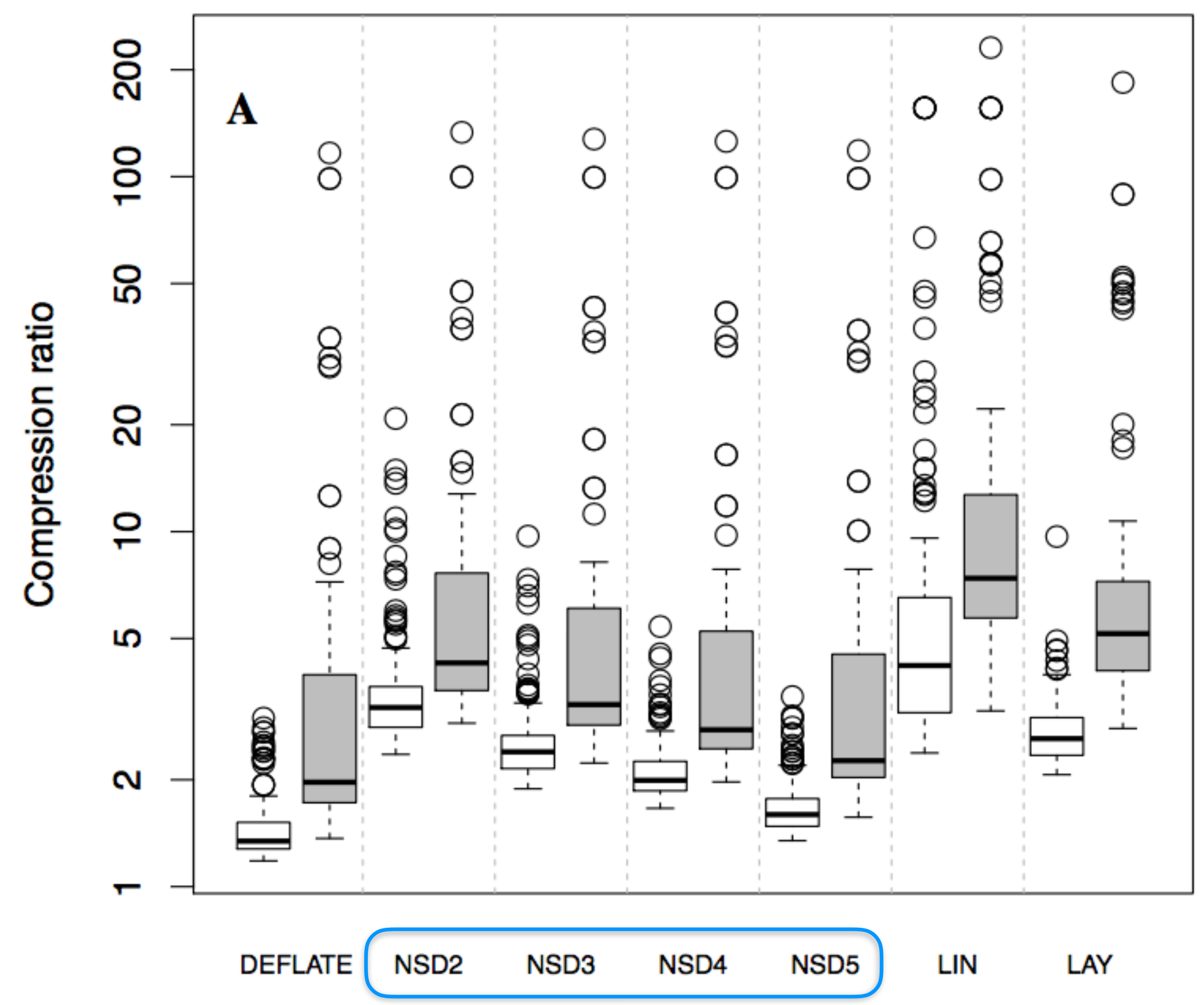

Fig. 3: Bit Grooming for NSD ~3.5 has similar trade-off between accuracy and compression to Layer Packing (LAY):

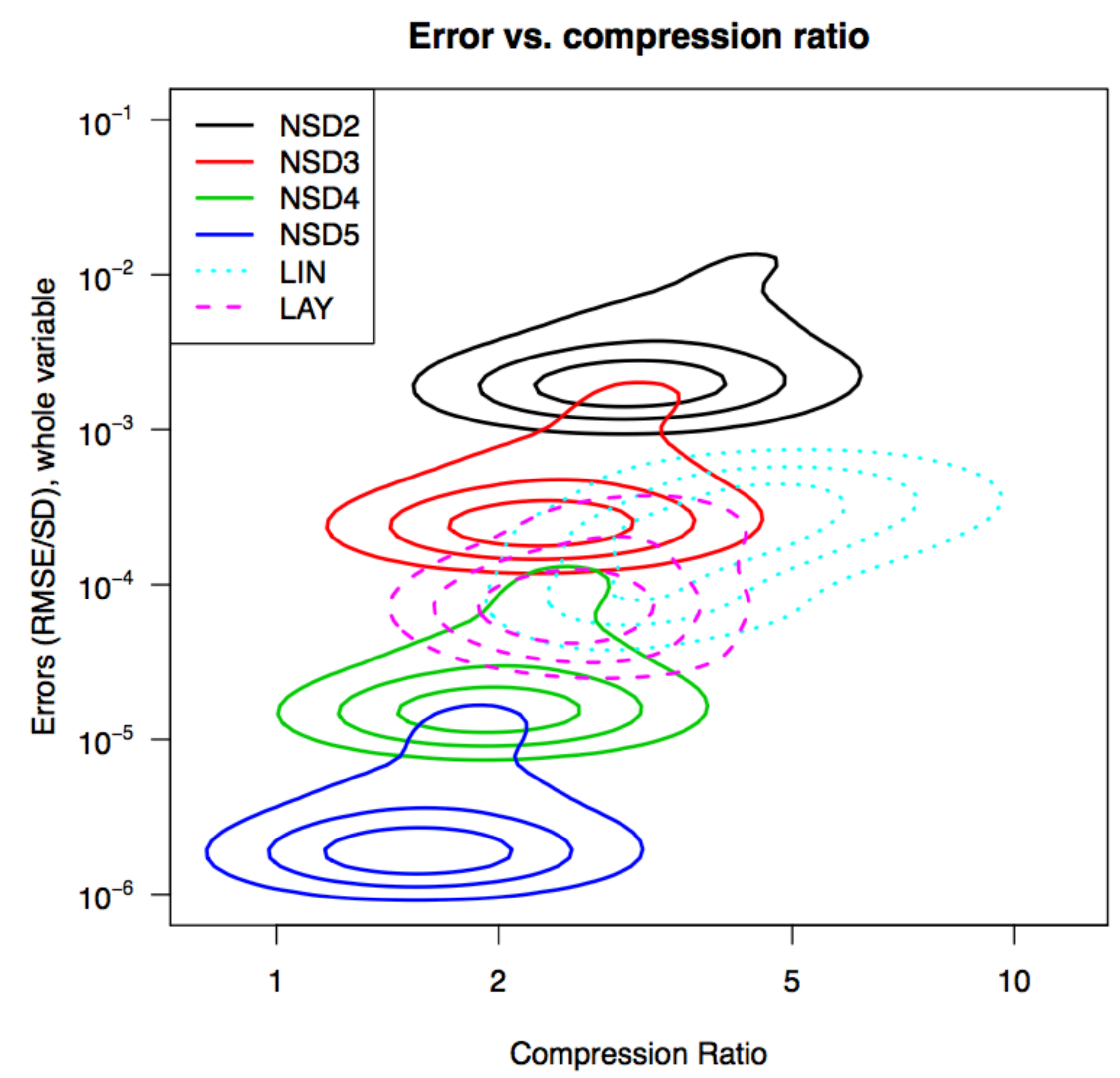

Fig 2: Bit Grooming (NSD*) mean error (smaller is better) is tunable, smaller than linear packing (LIN). NSD = 3.5 is rough equivalent of layer packing (LAY):

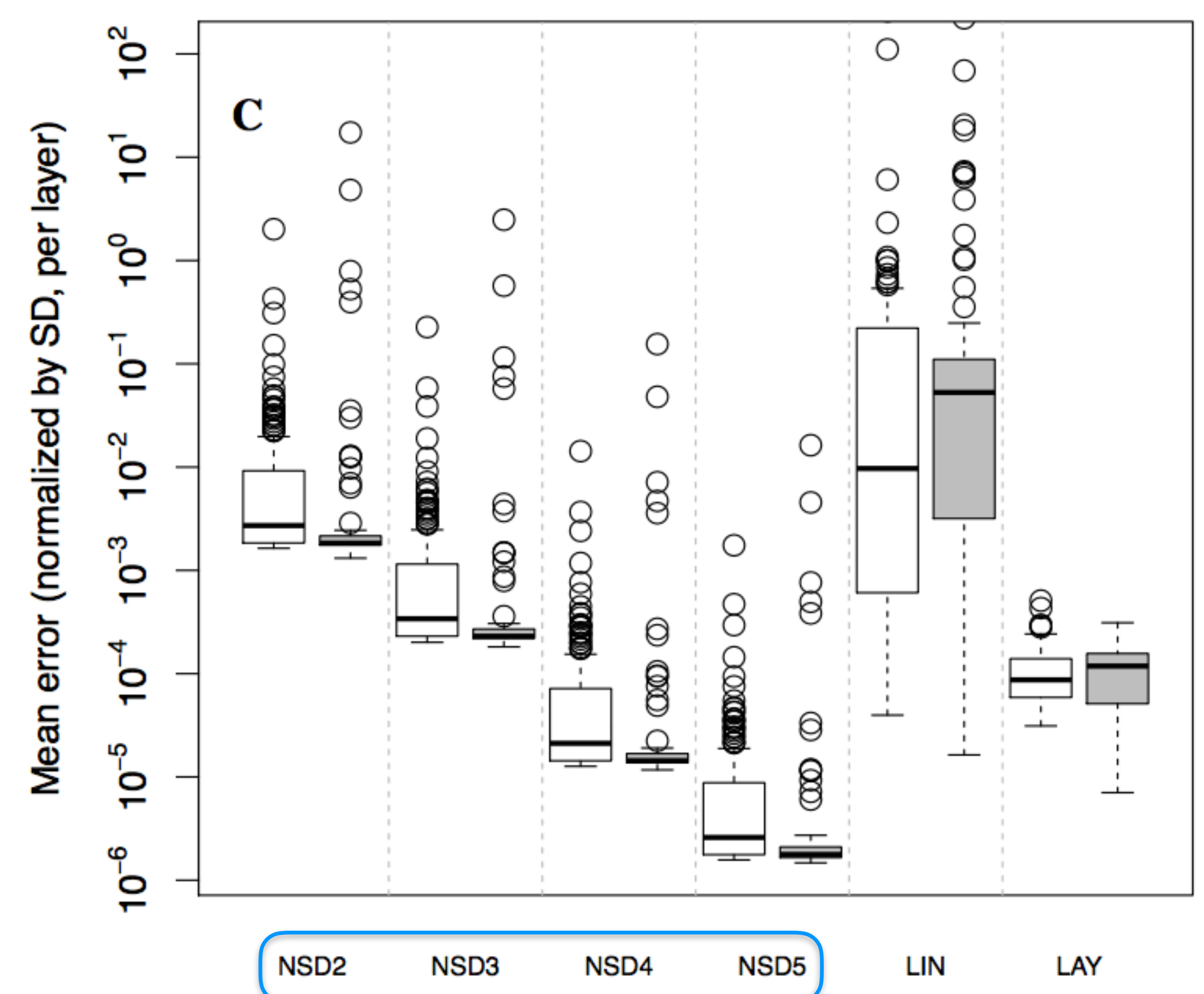

Fig. 4: Bit Grooming compression ratio (smaller is better) for NSD = 3 roughly 40% better than default lossless (DEFLATE) compression:

Row	Fmt	LLC	Qnt	Rng	NSD	Size	CR	Method
A	N3	-	-	10^{37}	~7	839.6	100.0	Uncompressed
B	N3	BZ1	-	10^{37}	~7	581.8	69.3	Bzip2
C	N3	BZ9	-	10^{37}	~7	580.8	69.2	Bzip2
D	N7	-	-	10^{37}	~7	823.2	98.1	Uncompressed
E	N7	DF1	-	10^{37}	~7	503.7	60.0	DEFLATE
F	N7	DF9	-	10^{37}	~7	491.3	58.5	DEFLATE
G	N7	-	LP	10^{5}	~1–4	413.4	49.2	Linear Packing
H	N7	DF1	LP	10^{5}	~1–4	162.6	19.4	Linear Packing
I	N7	DF1	BG	10^{37}	~7	503.6	60.0	Bit Grooming
J	N7	DF1	BG	10^{37}	6	485.0	57.8	Bit Grooming
K	N7	DF1	BG	10^{37}	5	427.6	50.9	Bit Grooming
L	N7	DF1	BG	10^{37}	4	346.2	41.2	Bit Grooming
M	N7	DF1	BG	10^{37}	3	289.6	34.5	Bit Grooming
N	N7	DF1	BG	10^{37}	2	229.2	27.3	Bit Grooming
O	N7	DF1	BG	10^{37}	1	161.4	19.2	Bit Grooming

Conclusions

How does Bit Grooming perform?

Bit Grooming is statistically unbiased, applies to all floating point numbers, and is easy to use. Bit-Grooming reduces ACME data storage requirements by 40-80%. We compared Bit Grooming to competitors Linear Packing, Layer Packing, and GRIB2/JPEG2000. The other compression methods can have better compression ratios, yet Bit Grooming is the most accurate, usable, and portable.

Why don't we Bit Groom already?

We're lazy. Bit Grooming provides flexible and well-balanced solutions to the trade-offs among compression, accuracy, and usability required by lossy compression. Users could reduce their long term storage costs, and show leadership in the elimination of false precision, by adopting Bit Grooming.

Implementation

netCDF Operators (NCO) produce Bit Groomed datasets with —ppc option:

```
ncks -7 —ppc default=5 in.nc out.nc # 5 sig. digits
ncks -7 —ppc p,w,z=5 —ppc q,RH=4 —ppc T,u,v=3 in.nc out.nc
ncks -7 —ppc default=5#q,RH=4#T,u,v=3 in.nc out.nc # Same
```

Bit-Groomed data is IEEE format, requires no special software to read!

References

Algorithm details and error analysis:
Zender, C. S. (2016), Bit Grooming: Statistically accurate precision-preserving quantization with compression, evaluated in the netCDF Operators (NCO, v4.4.8+), Geosci. Model Dev., 9, 3199-3211, doi:10.5194/gmd-9-3199-2016.

Intercomparison with other lossy compression algorithms:
Silver, J. D. and C. S. Zender (2017), The compression-error trade-off for large gridded datasets, Geosci. Model Dev., 10, 413-423, doi:10.5194/gmd-10-413-2017.

Software documentation:
`http://nco.sf.net/nco.html#ppc`

Support

DOE ACME DE-SC0012998, NASA ACCESS NNX14AH55A, NSF AGS-1541031.