Supplement of Geosci. Model Dev., 10, 413–423, 2017
http://www.geosci-model-dev.net/10/413/2017/
doi:10.5194/gmd-10-413-2017-supplement

Geoscientific
Model Development

Supplement of

The compression–error trade-off for large gridded data sets

J. D. Silver and C. S. Zender

Correspondence to: Jeremy D. Silver (jeremy.silver@unimelb.edu.au)

1 Compression methods

1.1 Command-line calls used

The compression methods compared were realized using the commands listed below. For each method apart from LAY, command-line tools from the NCO bundle (Zender, 2008) were used.

1. DEFLATE: Deflate compression (level 4) with shuffle filter

```
ncks -4 -L4 in.nc out.nc
```

2. NSD2, NSD3, NSD4, NSD5: Deflate compression (level 4) with shuffle filter, and bit grooming storing 2, 3, 4, or 5 significant figures (respectively). The following yields three significant digits (NSD3).

```
ncks -4 -L4 --ppc $var=3 in.nc out.nc
```

3. LIN: Deflate compression (level 4) with shuffle filter, scalar linear packing for each variable

```
ncpdq -4 -L4 in.nc out.nc
```

4. LAY: Deflate compression (level 4) with shuffle filter, layer packing for selected dimensions

```
ncpacklayer -L4 -v $var -d $dims in.nc out.nc
```

In the above, `$var` and `$dims` are Linux/Unix shell variables giving, respectively, the name of the sole variable contained within the input file and the names of the thick dimensions chosen for layer-packing.

We note that the above omits details of how the handling of chunking of variables was controlled. This was done by repeating, for each thick dimension, the argument `--cnk_dmn $dim,1` with the shell variable `$dim` set to the name of the dimension.

1.2 Further details about ncpacklayer

The compression is performed as follows:

```
ncpacklayer -d thickdim1,thickdim2 -v var1,var2,var3 original.nc packed.nc
```

Other optional flags allow for increased verbosity (`-V`), over-writing existing output files (`-O`) and defining the DEFLATE compression level (`-L`).

The mandatory `-d` flag is followed by a comma-separated list of the thick dimensions. The optional `-v` flag is followed by a comma-separated list of variables to pack. The default is to pack all variables defined along any of the thick dimensions listed. In the output file (in this example `packed.nc`) each variable that is packed (e.g. `var1`) is replaced by a trio of variables containing the arrays of packed values, scale factors and offsets. In this example, these are termed `var1__short`, `var1__scale` and `var1__offset`, with data type unsigned short (i.e. two-byte) integer, floating-point and floating-point, respectively. Suppose the original definition of `var1` is (following output format for the command line utility `ncdump`, which is provided when the netCDF API rather than the NCO bundle):

```
    float var1(thindim1, thickdim1, thickdim2, thindim2) ;
```

then the corresponding trio will have dimensions as follows:

```
    ushort var1__short(thindim1, thickdim1, thickdim2, thindim2) ;
    float var1__scale(thickdim1, thickdim2) ;
    float var1__offset(thickdim1, thickdim2) ;
```

In other words, the scale and offset arrays have one element per thin slice. Data remain in netCDF format in this packed format and retain all their attributes. Data can be unpacked as follows:

```
ncunpacklayer packed.nc unpacked.nc
```

The `-d` and `-v` flags are not used, since this information is contained in the trios of packed arrays.

2 Datasets

The tests described above were applied to the following datasets. In each case, we have provided the full list of variables in the analysis since in some cases not all variables provided in the files were featured in the analysis. We have not gone so far as to describe each of variables listed below, since this would take up much more space and because this information can generally be found within the metadata of each data set (the availability is listed in each case).

1. ERA-Interim reanalysis data (Dee et al., 2011)

   - Filename: `ei_mnth_an_pl_15x15_90N0E90S3585E_20080901_20081201`
   - Horizontal domain: a regular latitude-longitude grid covering the globe at 1.5° resolution. Latitude dimension of length 121, longitude dimension of length 240.
   - Vertical dimension: 37 pressure levels ranging from 1000 hPa to 1 hPa
   - Time dimension: 16 six-hourly snap-shots ranging from 2008-01-09 00:00 UTC to 2008-01-12 18:00 UTC
   - Notes: Converted from GRIB format prior to the analysis.
   - Layer packing: Thick dimensions chosen to be the time and vertical level.
   - What do the variables describe: atmospheric dynamics, temperature, ozone mixing ratio, cloud properties, humidity
   - Variables: 14 variables were included in the analysis. These were: PV_GDS0_ISBL_S123, Z_GDS0_ISBL_S123, T_GDS0_ISBL_S123, U_GDS0_ISBL_S123, V_GDS0_ISBL_S123, Q_GDS0_ISBL_S123, W_GDS0_ISBL_S123, VO_GDS0_ISBL_S123, D_GDS0_ISBL_S123, R_GDS0_ISBL_S123, O3_GDS0_ISBL_S123, CLWC_GDS0_ISBL_S123, CIWC_GDS0_ISBL_S123, CC_GDS0_ISBL_S123
   - Availability: This dataset could not be distributed with the other files due to licensing restrictions but can be accessed through the ECMWF's public dataset portal (`http://apps.ecmwf.int/datasets/`), using the following set of inputs: stream = synoptic monthly means, vertical levels = pressure levels (all 37 layers), parameters = all 14 variables, dataset = interim_mnth, step = 0, version = 1, time = 00:00:00, 06:00:00, 12:00:00, 18:00:00, date = 20080901 to 20081201, grid = 1.5° × 1.5°, type = analysis, class = ERA Interim.

2. A limited area subset from global MOZART model output (Brasseur et al., 1998). Dimensions: 9 × 10 grid-points in the horizontal, 56 vertical levels, 172 time-points. 77 variables with these four dimensions.

   - Filename: `mozart4geos5_2011-02-01_2011-03-16.nc`
   - Horizontal domain: a limited area subset of a global domain covering Australia. The global domain appears to have 95 × 144 gridpoints, while only 9 × 10 grid-points (lon × lat) in the horizontal are covered in this file. The grid spacing is regular at 2.5° resolution in the latitude dimension and 1.895° resolution in the longitude dimension.
   - Vertical dimension: fixed pressure levels ranging with mid-points ranging from 992.5 Pa to 1.868 Pa.
   - Time dimension: 172 temporal snapshots at six-hourly resolution ranging from 2011-02-01 06:00 UTC to 2011-02-01 12:00 UTC.
   - Notes: Originally downloaded through through the web-page `http://www.acom.ucar.edu/wrf-chem/mozart.shtml`. This file contained smaller variables (other than coordinate variables) that were not included in the analysis due to their relatively small size.
   - Layer packing: Thick dimension chosen to be the vertical level.
   - What do the variables describe: volume mixing ratios of many trace gases, mass mixing ratios of some aerosol classes, atmospheric dynamics, temperature, photolytic reaction rates
   - Variables: 77 variables were included in the analysis. Their names were: BIGALD_VMR_inst, BIGALK_VMR_inst, BIGENE_VMR_inst, C10H16_VMR_inst, C2H2_VMR_inst, C2H4_VMR_inst, C2H5OH_VMR_inst, C2H6_VMR_inst, C3H6_VMR_inst, C3H8_VMR_inst, CB1_VMR_inst, CB2_VMR_inst, CH2O_VMR_inst, CH3CHO_VMR_inst, CH3CN_VMR_inst, CH3COCH3_VMR_inst, CH3COCHO_VMR_inst, CH3COOH_VMR_inst, CH3COOOH_VMR_inst, CH3O2_VMR_inst, CH3OH_VMR_inst, CH3OOH_VMR_inst, CH4_VMR_inst, CO_VMR_inst, CRESOL_VMR_inst, DMS_VMR_inst, DUST1, DUST2, DUST3, DUST4, GLYALD_VMR_inst, H2O, H2O2_VMR_inst, HCN_VMR_inst, HCOOH_VMR_inst, HNO3_VMR_inst, HO2NO2_VMR_inst, HO2_VMR_inst, HYAC_VMR_inst, HYDRALD_VMR_inst,

  ISOPNO3_VMR_inst, ISOP_VMR_inst, MACR_VMR_inst, MEK_VMR_inst, MPAN_VMR_inst, MVK_VMR_inst, N2O5_VMR_inst, N2O_VMR_inst, NH3_VMR_inst, NH4NO3_VMR_inst, NH4_VMR_inst, NO2_VMR_inst, NO3_VMR_inst, NOX, NOY, NO_VMR_inst, O3_VMR_inst, OC1_VMR_inst, OC2_VMR_inst, OH_VMR_inst, ONITR_VMR_inst, ONIT_VMR_inst, PAN_VMR_inst, Q, SA1_VMR_inst, SA2_VMR_inst, SA3_VMR_inst, SA4_VMR_inst, SO2_VMR_inst, SO4_VMR_inst, SOA_VMR_inst, T, TOLUENE_VMR_inst, U, V, jno2_rcon_inst, jo1d_rcon_inst

  - Availability: available online at `https://figshare.com/projects/Layer_Packing_Tests/14480`

3. Model output from the Weather Research and Forecasting (WRF) model (Skamarock et al., 2005).
   - Filename: `wrfout_d03_2013-01-24_07:00:00`
   - Horizontal domain: A limited area domain over the city of Sydney and surrounding areas (Australia), including a portion over the sea. A Lambert Conformal map projection was used and the horizontal resolution was 1 km in the east-west and north-south dimensions. There were 165 grid-points in the east-west dimension and 140 in the north-south dimension.
   - Vertical dimension: 32 levels using a terrain-following, hydrostatic pressure coordinate from the surface to 5 hPa.
   - Time dimension: A single time snaps-hot (2013-01-24 07:00 UTC)
   - Notes: Model output from simulations by J. Silver. This file contained smaller variables (other than coordinate variables) that were not included in the analysis due to their relatively small size.
   - Layer packing: Thick dimension chosen to be the vertical level.
   - What do the variables describe: atmospheric dynamics, temperature, cloud properties, humidity
   - Variables: 20 variables were included in the analysis. Their names were: U, V, W, PH, PHB, T, P, PB, QVAPOR, QCLOUD, QRAIN, QICE, QSNOW, QNICE, QNSNOW, QNRAIN, QNDROP, TKE_PBL, EL_PBL, CLDFRA
   - Availability: available online at `https://figshare.com/projects/Layer_Packing_Tests/14480`
4. MERRA reanalysis product (Rienecker et al., 2011).
   - Filename: `MERRA300.prod.assim.inst3_3d_asm_Cp.20130601.nc`
   - Horizontal domain: a regular latitude-longitude grid covering the globe at 1.25° resolution. Latitude dimension of length 144, longitude dimension of length 288.
   - Vertical dimension: 37 pressure levels ranging from 1000 hPa to 0.1 hPa
   - Time dimension: 8 temporal snapshots at three-hourly frequency, ranging from 2013-06-01 00:00 UTC to 2013-06-01 21:00 UTC
   - Notes: This file contained smaller variables (other than coordinate variables) that were not included in the analysis due to their relatively small size.
   - Layer packing: Thick dimension chosen to be the vertical level.
   - What do the variables describe: atmospheric dynamics, temperature, cloud properties, humidity, ozone mixing ratio
   - Variables: 11 variables were included in the analysis. Their names were: EPV, H, O3, OMEGA, QI, QL, QV, RH, T, U, V
   - Availability: available online at `https://figshare.com/projects/Layer_Packing_Tests/14480`
5. Output of the mineral Dust Entrainment And Deposition (DEAD) model (Zender et al., 2003).
   - Filename: `dstmch90_clm.nc`
   - Horizontal domain: a regular latitude-longitude grid covering the globe at 1.875° resolution in the longitude dimension and 1.904° resolution in the latitude dimension. Latitude dimension of length 94, longitude dimension of length 192.
   - Vertical dimension: a hybrid vertical coordinate system with 28 levels ranging from 1000 hPa to 2.7 hPa.
   - Time dimension: one time-point
   - Notes: This file contained smaller variables (other than coordinate variables) that were not included in the analysis due to their relatively small size.

- Layer packing: Thick dimension chosen to be the vertical level.
- What do the variables describe: atmospheric dynamics, temperature, cloud properties, humidity, mass and mass flux rates for dust (either total or in size different bins)
- Variables: 15 variables were included in the analysis. Their names were: U, V, T, Q, RELHUM, CLOUD, CWAT, DSTQ, DSTQ01, DSTQ02, DSTQ03, DSTQ04, DSTSSPCP, DSTSSEVP, DSTSSDRY
- Availability: available at the DEAD model homepage (`http://dust.ess.uci.edu/dead/`) and also at `https://figshare.com/projects/Layer_Packing_Tests/14480`

6. Model output from the coupled numerical weather prediction and chemistry transport model CAM-SE (Dennis et al., 2012).

- Filename: `famipc5_ne30_v0.3_00003.cam.h0.1979-01-L5.nc`
- Horizontal domain: A non-rectangular cube-sphere mesh, ordered as a single array of 48602
- Vertical dimension: a hybrid vertical coordinate system with 30 levels ranging from 992 hPa to 3.6 hPa.
- Time dimension: Only a single time-point is represented
- Notes: This file contained smaller variables (other than coordinate variables) that were not included in the analysis due to their relatively small size.
- What do the variables describe: aerosol and trace-gas concentrations, atmospheric dynamics, temperature, cloud properties
- Variables: 118 variables were included in the analysis. Their names were: AQRAIN, AQSNOW, AREI, AREL, AWNC, AWNI, CCN3, CLDICE, CLDLIQ, CLOUD, DCQ, DMS, DTCOND, DTV, FICE, FREQI, FREQL, FREQR, FREQS, H2O2, H2SO4, ICIMR, ICWMR, IWC, LIQCLDF, NUMICE, NUMLIQ, OMEGA, OMEGAT, Q, QRL, QRS, RELHUM, SO2, SO2_XFRC, SOAG, T, U, UU, V, VD01, VQ, VT, VU, VV, Vbc_a1, Vdst_a1, Vdst_a3, V ncl_a1, Vncl_a2, Vncl_a3, Vpom_a1, Vso4_a1, Vso4_a2, Vso4_a3, Vsoa_a1, Vsoa_a2, WSUB, XPH_LWC, Z3, bc_a1, bc_a1_2, bc_a1_XFRC, bc_c1, dgnd_a01, dgnd_a02, dgnd_a03, dgnumwet1, dgnumwet2, dgnumwet3, dgnw_a01, dgnw_a02, dgnw_a03, dst_a1, dst_a1_2, dst_a3, dst_a3_2, dst_c1, dst_c3, ncl_a1, ncl_a1_2, ncl _a2, ncl_a2_2, ncl_a3, ncl_a3_2, ncl_c1, ncl_c2, ncl_c3, num_a1, num_a2, num_a3, num_c1, num_c2, num_c3, pom_a1, pom_a1_2, pom_a1_XFRC, pom_c1, so4_a1, so4_a1_2, so4_a1_XFRC, so4_a2, so4_a2_2, so4_a2_XFRC, so4_a3, so4_a3_2, so4_c1, so4_c2, so4_c3, soa_a1, soa_a1_2, soa_a2, soa_a2_2, soa_c1, soa_c2, w at_a1, wat_a2, wat_a3
- Availability: available online at `https://figshare.com/projects/Layer_Packing_Tests/14480`

As described in the manuscript (under the heading "Complexity statistics"), the variables were classified as "sparse" or "dense". Sparse variables were highly compressible, which was often due their non-trivial components being limited to a fraction of the data array. Sparse variables were chosen to be those satisfying any one of the following conditions: the compression ratio is greater than 5.0 using DEFLATE, the fraction of values equal to the most common value in the entire variable is greater than 0.2, and the fraction of hyperslices where all values were identical is great than 0.2. The breakdown among the different categories is given in Table 1.

CompRatio > 5	globalMaxP > 0.2	propUniform > 0.2	# vars
T	T	T	19
T	T	F	0
T	F	T	0
T	F	F	0
F	T	T	16
F	T	F	39
F	F	T	0
F	F	F	181

Table 1: Number of variables fitting different "sparsity" criteria. Abbreviations: CompRatio = compression ratio using DEFLATE (level 4), globalMaxP = the fraction of values equal to the most common value in the entire variable, propUniform = the fraction of hyperslices where all values were identical, # vars = number of variables.

3 Normalization errors

Figure 1C of the main manuscript shows the distribution of normalized errors for the six lossy compression methods. In the main manuscript (under the heading “Error and compression metrics”), four different methods are described for normalizing the root mean-squared error. These were:

A. calculating the RMSE and standard deviation (SD) separately for each thin slice, and averaging the ratio RMSE/SD across thin slices;

B. taking the average across the per-slice RMSE and SD values, and then taking the ratio of these averages – that is, mean(RMSE)/mean(SD);

C. the same as A, except normalizing by the per-slice mean rather than the per-slice SD;

D. the same as B, except normalizing by the average of the per-slice means.

Figure 1C of the main manuscript shows the distribution of errors for normalization method A, and this is repeated in panel A of Figure 1 (this document); similarly, the distribution of methods B, C and D appear in their respectively-named panels of the same figure.

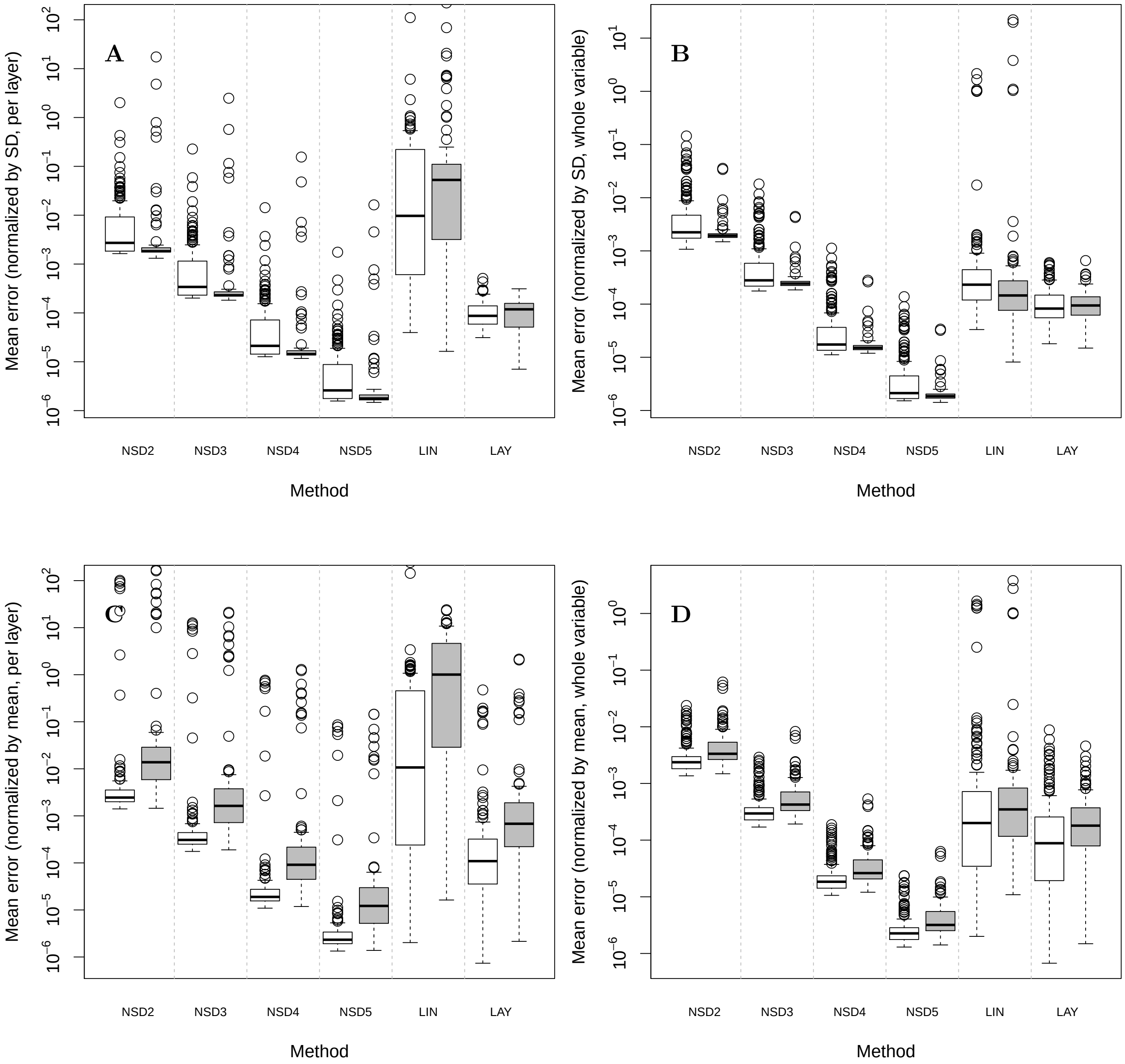


Figure 1: Distribution of errors with different normalization methods, plotted separately by dense variables and sparse variables (white and grey boxes, respectively). Top-left: normalized by the per-layer standard deviation. Top-right: normalized by the average of the per-layer standard deviations. Bottom-left: normalized by the per-layer mean. Bottom-right: normalized by the whole-variable mean.

4 Entropy and compression for reduced-precision fields

For the reduced-precision fields, we assessed relationship between the normalized entropy of the data field (NEDF) and the compression ratios. Figure 2 shows the compression ratios relative to the *uncompressed file sizes* whereas Figure 3 displays the compression ratios relative to the *DEFLATE-compressed file sizes.* Figure 2 presents the NEDF for each variable whereas Figure 3 plots the reduction in the NEDF due to the lossy filters.

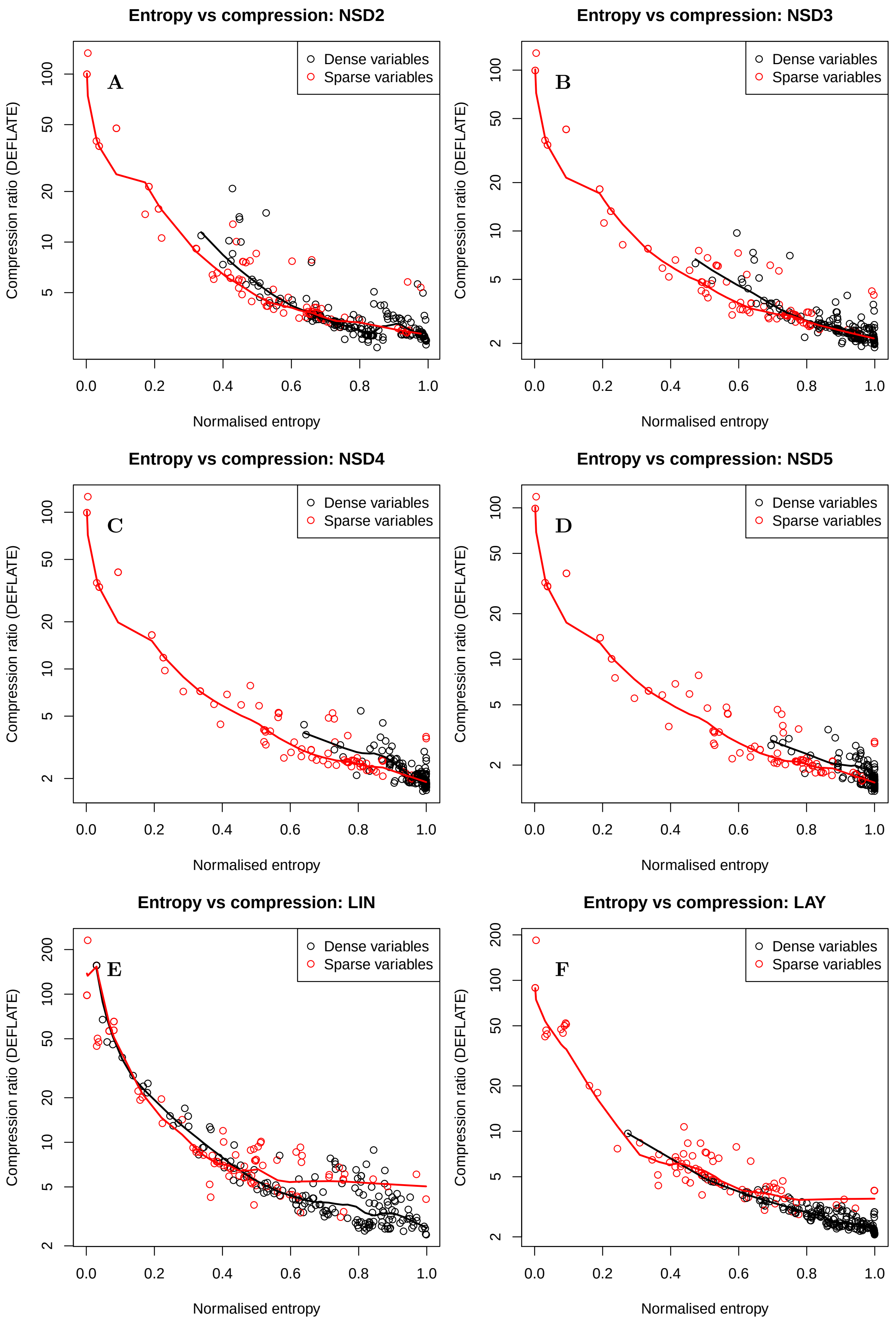


Figure 2: Compression ratios for each of the lossy compression methods compared to the respective normalized entropy of each variable's data field; this accounts for quantization of the data field.

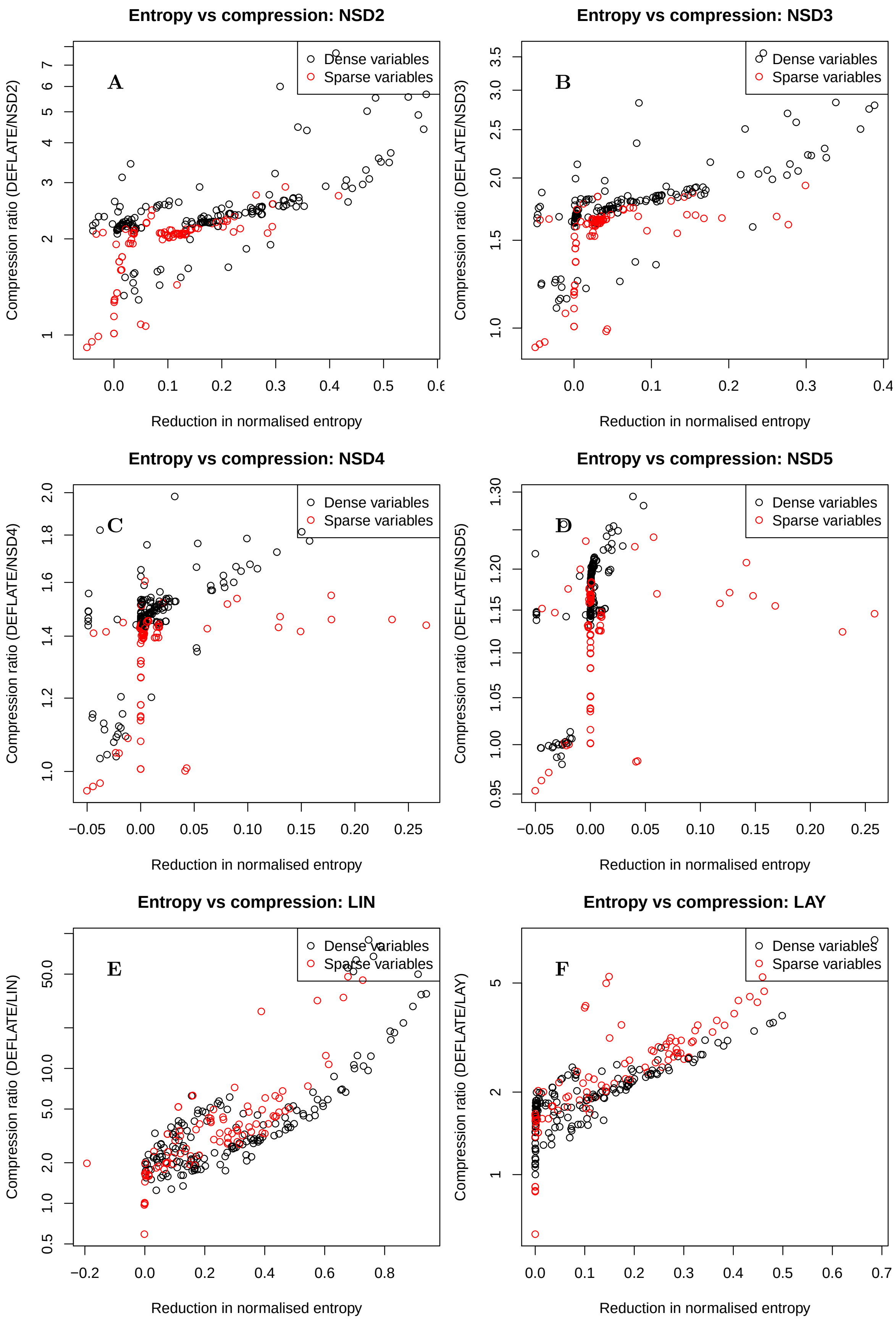
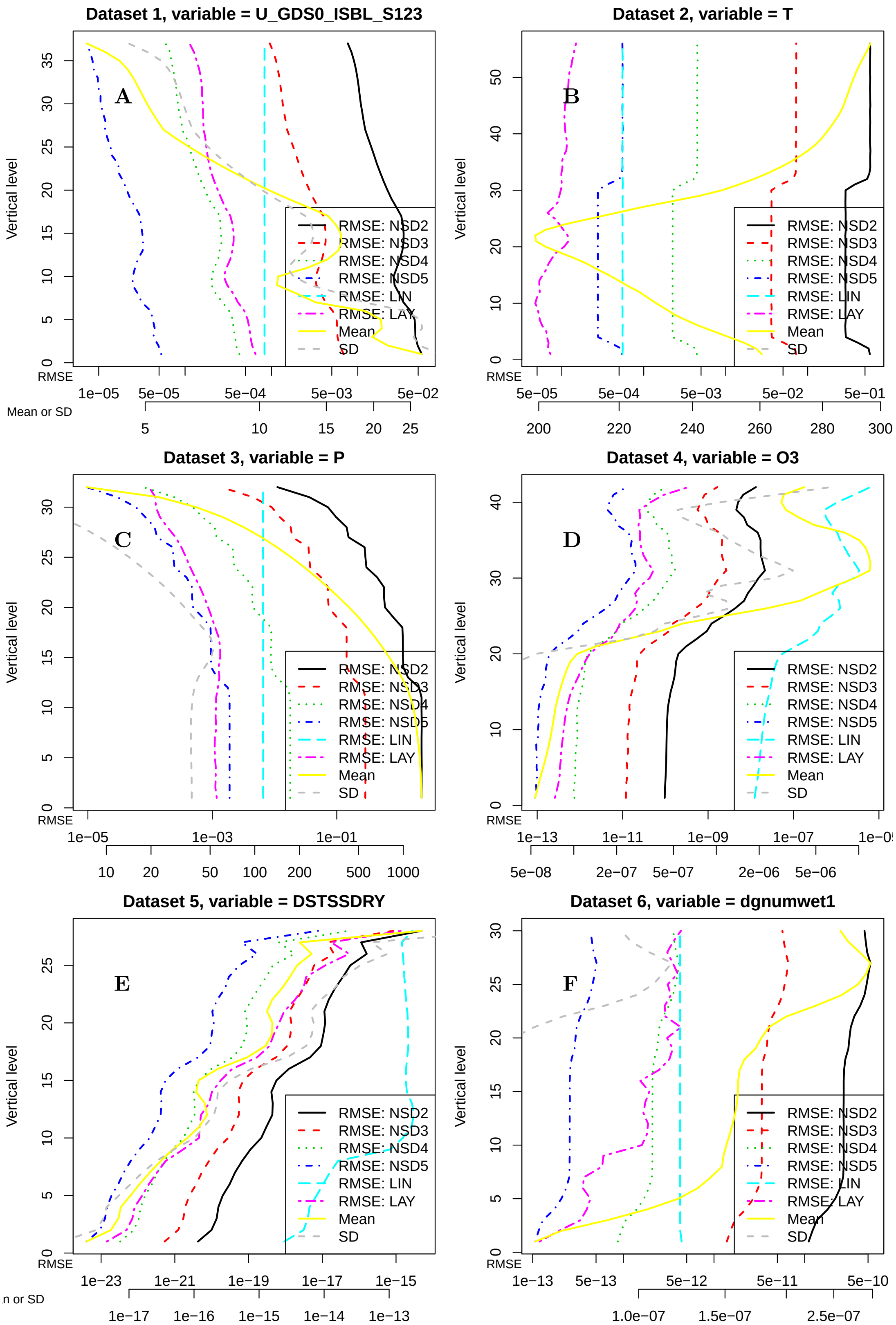


Figure 3: Compression ratios *relative to DEFLATE* for each of the lossy compression methods compared to the reduction in the normalized entropy due to the lossy compression.

5 Vertical profiles of errors for selected variables

Figures 4 and 5 illustrate, for six selected variables among the 255 considered, vertical profiles of the RMSE for each of the lossy compression methods. Figure 4 shows the *absolute* RMSE whereas Figure 5 displays the RMSE normalized by the corresponding per-level standard deviation. The six variables presented are:

1. U_GDS0_ISBL_S123: East-west wind velocity (units = m s^{-}1)
2. T: temperature (units = K)
3. P: perturbation pressure (units = Pa)
4. O3: Ozone mixing ratio (units = Kg$\cdot$Kg^{-1})
5. DSTSSDRY: Total dust tendency due to settling and turbulence (units = kg kg^{-1} second^{-1})
6. dgnumwet1: Aerosol mode wet diameter (units = m)



Dataset 1, variable = U_GDS0_ISBL_S123

Dataset 2, variable = T

Dataset 3, variable = P

Dataset 4, variable = O3

Dataset 5, variable = DSTSSDRY

Dataset 6, variable = dgnumwet1

Figure 4: Errors from the six lossy compression methods are shown as a function of vertical level for six variables (one from each dataset included). Also shown are the corresponding mean (of the absolute values) and standard deviation for the given variable. The errors were not normalized. Note that two scales are shown on the horizontal axis (at the bottom of each panel), the upper of which pertains to the errors and the lower scale to the mean and standard deviation. Also note the logarithmic scale on the x-axis.

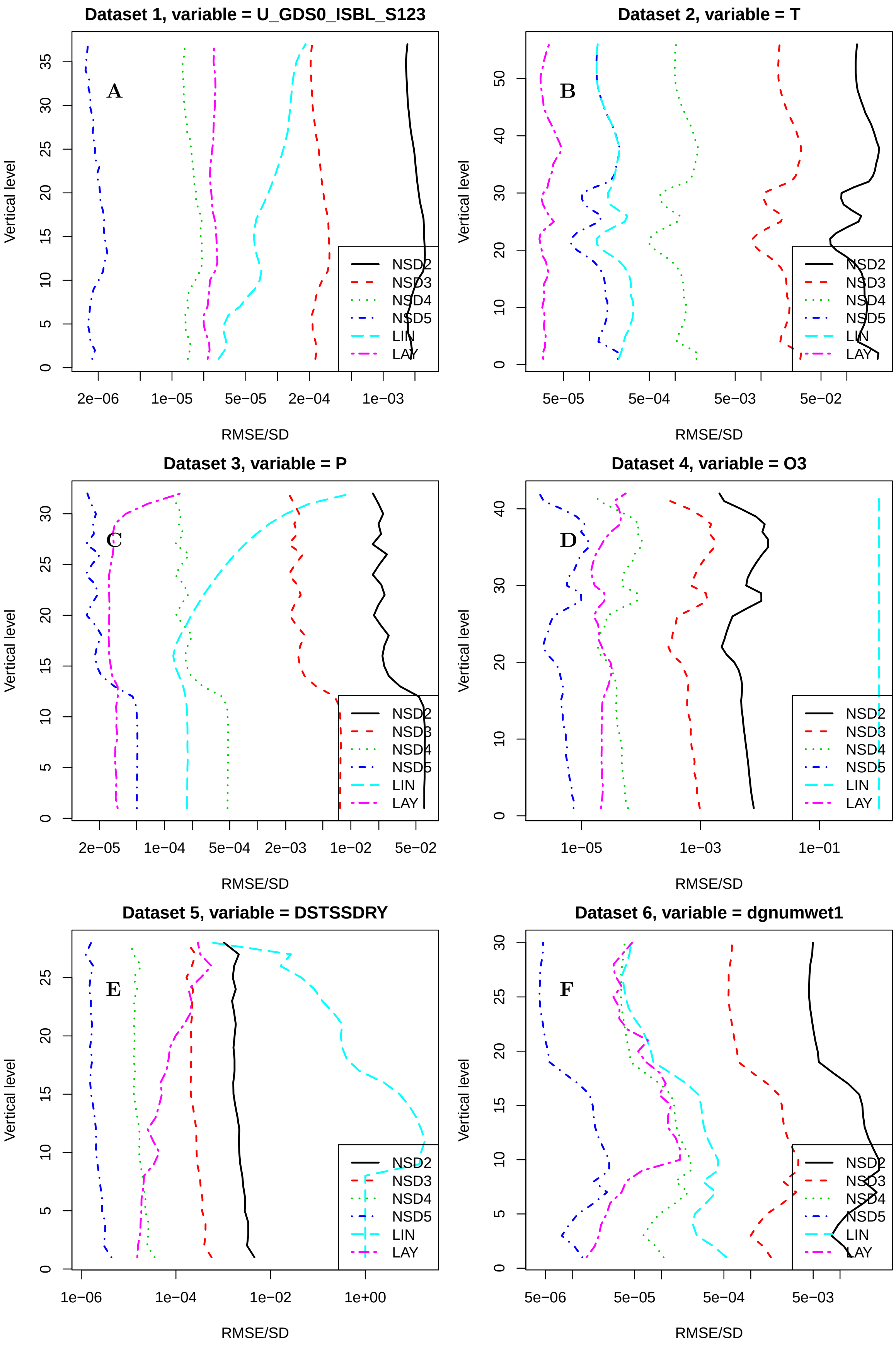


Figure 5: Relative errors (normalizing by the per-layer standard deviation) from the six lossy compression methods are shown as a function of vertical level for six variables (the same variables as shown in Figure 4). Note the logarithmic scale on the x-axis.

6 Details of the complexity statistics calculated

As described in the manuscript, a range of statistics were calculated for every variable in the analysis. The following presents details of each of these. The statistics calculated were:

1. the normalized entropy of the floating point array,
2. the normalized entropy of the exponent array,
3. the normalized entropy of the mantissa array,
4. the fraction of values equal to the mode (i.e. the most common value in the hyperslice),
5. statistics representing the decay rate of singular values,

   This was calculated by calculating the singular-value decomposition of the two-dimensional slice, then finding the points at which the cumulative sum exceeded 0.5, 0.75, 0.9 or 0.95 times the total sum of the singular values; this was then represented as the fraction of the total number of singular values at which these points were reached (i.e. this yielded four separate statistics).

   Similar to the above, except searching for the fraction of the singular value beyond which the singular values fall below or 0.05, 0.1, 0.25, or 0.5 times the largest singular value (i.e. this also yields four statistics).

6. the spatial autocorrelation at fixed separation distances,

   This was calculated by estimating, for each separation distance up to 0.66 of the smaller array dimension in the two-dimensional slice, the correlation between a random sample of points separated by this distance (calculating distances by the Cartesian distance metric in terms of grid-spacing, rather than physical space). This then formed a scatter-plot of correlation versus distance, through which a locally-weighted scatter-plot smoother (LOWESS) curve was fitted (Cleveland, 1981). The points at which this curve fell below 0.95, 0.9, 0.75 or 0.5 were noted and these were represented as the fraction along the length of the smaller axis (i.e. this yielded four statistics).

   The above was done for separations in only the rows or columns, in which case the points at which the curve fell below the threshold were represented as the fraction along the length of the corresponding axis (i.e. this yielded eight statistics in total).

7. the mean (or mean of absolute values) divided by the standard deviation,
8. same as above, except for non-zero values only,
9. the range of the exponent field,
10. the standard deviation of the exponent field, and
11. the logarithm of the largest non-zero value divided by the smallest non-zero value.

As well as these, two global statistics were calculated:

1. the fraction of values equal to the mode (i.e. the most common value) in the entire variable and
2. the fraction of hyperslices where all values were identical.

References

Brasseur, G., Hauglustaine, D., Walters, S., Rasch, P., Müller, J.-F., Granier, C., and Tie, X. (1998). MOZART, a global chemical transport model for ozone and related chemical tracers: 1. Model description. *Journal of Geophysical Research: Atmospheres*, 103(D21):28265–28289.

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. The. *The American Statistician*, 35:54.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., and Bechtold, P. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Dennis, J. M., Edwards, J., Evans, K. J., Guba, O., Lauritzen, P. H., Mirin, A. A., St-Cyr, A., Taylor, M. A., and Worley, P. H. (2012). CAM-SE: A scalable spectral element dynamical core for the Community Atmosphere Model. *Int. J. High Perform. Comput. Appl.*, 26:74–89.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., et al. (2011). MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate*, 24(14):3624–3648.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers, J. G. (2005). A description of the advanced research WRF version 2. Technical Report NCAR/TN-468 STR, National Center For Atmospheric Research, Boulder, Colarado, USA.

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). *Environmental Modelling & Software*, 23(10):1338–1342.

Zender, C. S., Bian, H., and Newman, D. (2003). Mineral Dust Entrainment And Deposition (DEAD) model: Description and 1990s dust climatology. *J. Geophys. Res.*, 108(D14):4416.