Technical lossless / near lossless data compression

Nigel Atkinson (Met Office, UK)

Contents

- Survey of file compression tools
- Studies for AVIRIS imager
- Study by Tony Lee on IASI
- Compression applied to real IASI data
- Implications for IASI-NG and MTG-IRS
- Conclusions and discussion

Definitions

- Lossless
  - Can reconstruct the input exactly
  - Typical input: IASI level 1C scaled radiances
    - 16 bit scaled radiances, gaussian apodised
- Near-lossless
  - Can reconstruct the input with a defined maximum error
  - Error typically a defined (small) fraction of instrument noise
  - Example: digitisation error (or quantisation error)
    - Max error = $\delta y / 2$
    - RMS error = $\delta y / \sqrt{12}$

Standard lossless file compression techniques

- bzip2 (v1.0 released in 2000)
  - Compresses data in blocks of size between 100 and 900 kB
  - Uses a combination of *run-length encoding*, *Burrows-Wheeler transform* (a statistical modelling technique) and *Huffman encoding*
    - ***Huffman encoding*** *assigns short codes to frequently occurring numbers, and long codes to rare numbers*
    - *A form of entropy coding; Arithmetic coding is another.*
  - Good compression ratio
  - Decompression is faster than compression
  - Widely used on EUMETCast!
- gzip (1993) zlib uses the same algorithm — ***built into netCDF4***
  - Uses "DEFLATE" which is a combination of *LZ77 (Lempel-Zif - 1977)* and *Huffman encoding*
  - Faster than bzip2 but not as effective
  - *LZ77* works by looking for repeated references within a sliding window, keeping track of the lengths and distances

Standard file compression techniques (2)

- lzip (2008)
  - uses Lempel–Ziv–Markov chain algorithm (*LZMA*)
  - Dictionary compression scheme similar to LZ77
  - Said to have a high compression ratio (but not installed on Met Office desktops so can't verify)
- xz
  - uses *LZMA2*, so probably similar performance to lzip
- compress (1985)
  - uses Lempel–Ziv–Welch (*LZW*)
  - fell out of favour because of patent on LZW (now expired)

BUFR

- BUFR compression
  - Look at all occurrences of an element within a message
  - Express as a minimum, an increment width and a set of increments (in a reduced number of bits) to be added to that minimum.
  - Longer messages (i.e. many subsets) usually compress better than short ones.

Image compression standards

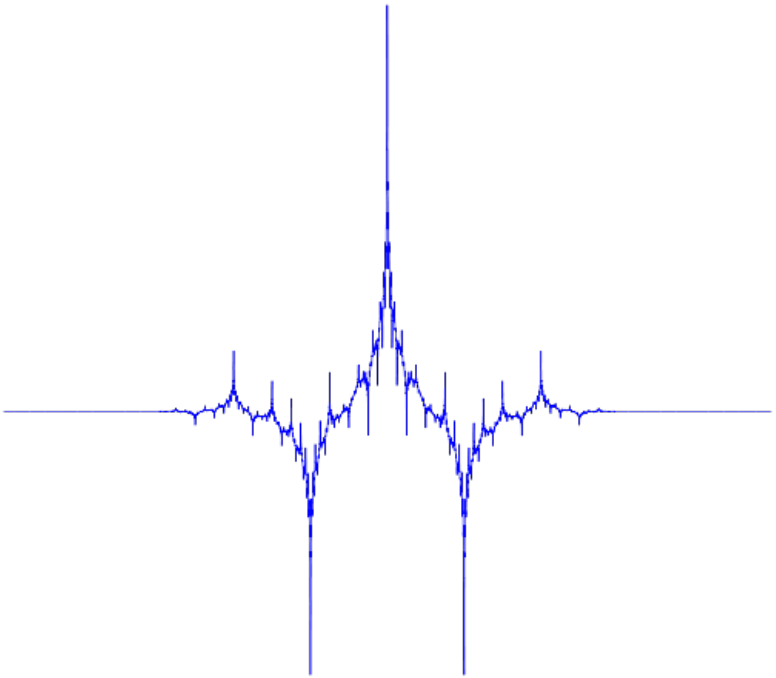
- designed for 2D images

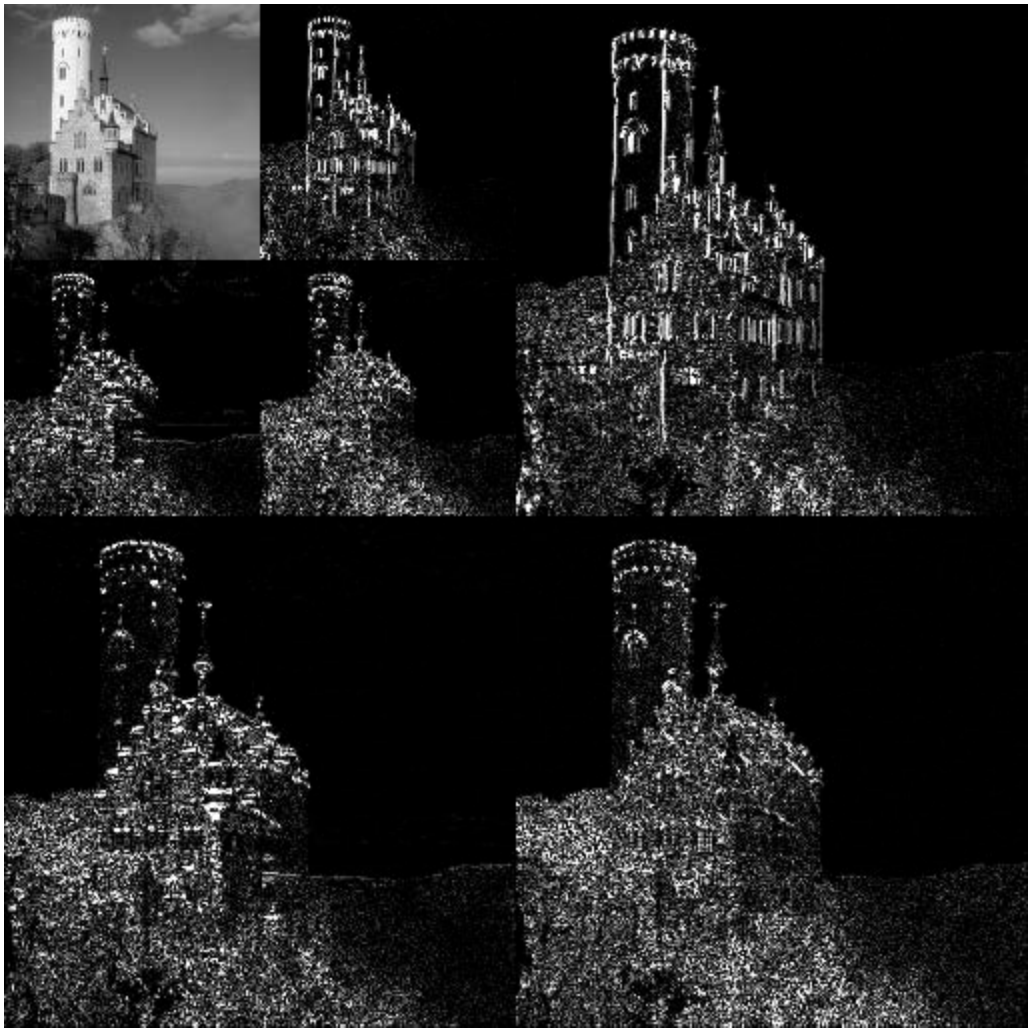
- JPEG-LS (LS=lossless)
  - The *LOCO-I* algorithm
    - Decorrelation/prediction – edge detection using neighbouring samples
    - Context modeling – local gradients
    - Coding corrected prediction residuals
    - Run length coding in uniform areas
- JPEG2000
  - Uses an integer wavelet transform
    - Decomposes the image into multiple resolutions
    - Lossless and lossy options

Image compression (2)

- szip
  - Can be used internally by HDF5 (netCDF4 can read but not write)
  - Designed for compressing data on the surface of a sphere (astronomical or environmental maps)
  - Uses wavelet transform on a sphere
  - Otherwise, similar to JPEG2000

Wavelet transform in JPEG2000





Some studies on compression of hyperspectral images

Jarno Mielikainen and Bormin Huang (2012) – Lossless compression of hyperspectral images using clustered linear prediction with adaptive prediction length

Jarno Mielikainen and Pekka Toivanen (2003) – Clustered DPCM for the lossless compression of hyperspectral images

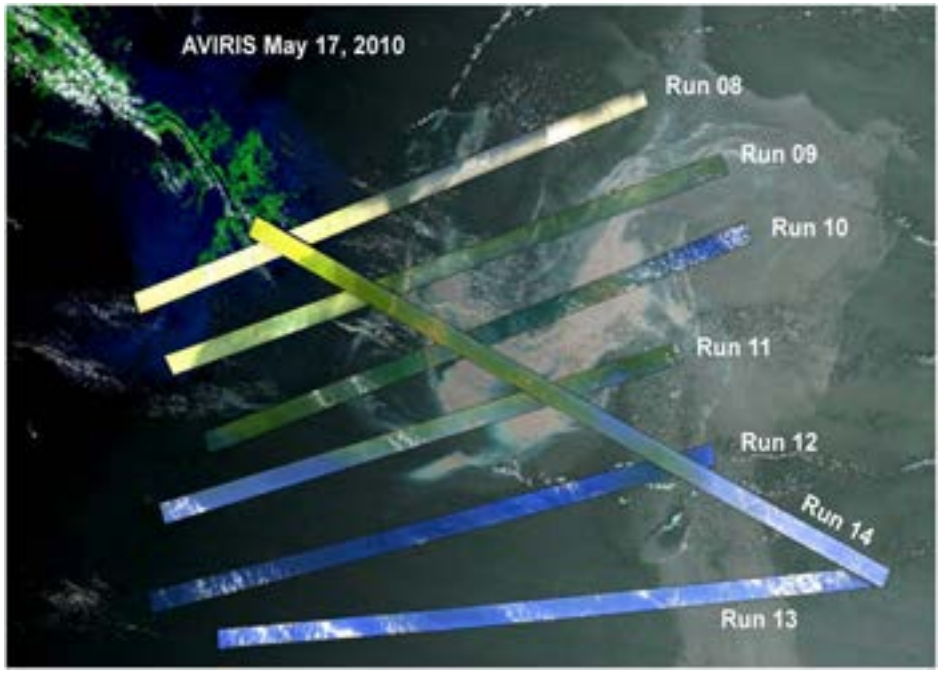
Arto Kaarna (2001) – Integer PCA and wavelet transforms for multispectral image compression

Mark Pickering and Michael Ryan (2001) – Efficient spatial-spectral compression of hyperspectral data

Michael Ryan (1997) – The lossless compression of AVIRIS images by vector quantization

AVIRIS studies

- AVIRIS is an aircraft-borne visible/near IR imager
- 224 spectral channels, 400-2500 nm
- 677 pixels wide



Mielikainen and Huang (2012)

Worked with 677 x 512 AVIRIS images

- Clustering
  - Generate 16 “clusters” – each cluster contains spectra with similar properties
- Prediction
  - For each cluster, look at differences between channels and derive prediction coefficients
- Coding
  - Entropy coding for each cluster

Example of AVIRIS results

TABLE III
BIT RATES IN BITS PER PIXEL FOR LOSSLESS COMPRESSION OF THE CALIBRATED YELLOWSTONE IMAGES OF THE CCSDS 2006 AVIRIS DATA SET

Algorithm	Scene 0	Scene 3	Scene 10	Scene 11	Scene 18	Average
JPEG-LS	6.95	6.68	5.19	6.24	7.02	6.42
LUT	4.81	4.62	3.95	4.34	4.84	4.51
LAIS-LUT	4.48	4.31	3.71	4.02	4.48	4.20
FL#	3.91	3.79	3.37	3.59	3.90	3.71
BG	4.29	4.16	3.49	3.90	4.23	4.01
A1	4.81	4.69	4.01	4.41	4.77	4.54
C-DPCM	3.61	3.43	2.97	3.28	3.49	3.36
S-RLP	3.58	3.43	2.95	3.27	3.46	3.34
S-FMP	3.54	3.39	2.94	**3.25**	3.44	3.31
C-DPCM-APL	**3.52**	**3.36**	**2.93**	**3.25**	**3.42**	**3.29**

- These are for 16-bit calibrated images
- Compression factor ~0.2 (compression ratio ~5)
- Useful improvement over JPEG-LS (differential, i.e. channel differences)
- BUT … "The whole compression process takes 20 hours"

Kaarna (2001) – PCA and wavelet transform

- Also working with AVIRIS images
- Selected a small number of points in the image and used to generate PCs – just used 8 PCs
- Compute PC scores and residuals
- Applied 3-D wavelet transform to the residuals (2 spatial dimensions and 1 spectral)

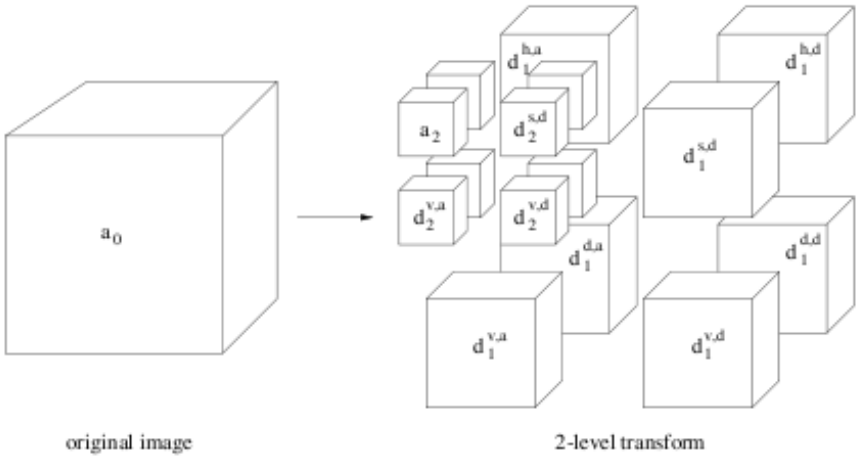


Fig. 1. The three-dimensional transform with two levels.

Kaarna - results

Before

TABLE I

ORIGINAL ENTROPIES AND COMPRESSION RATIOS OF THE FOUR TEST IMAGES.

Image	entropy bits/sample	CR	bitrate bits/sample
Jasper	11.19	2.01	7.94
Moffet	11.55	1.97	8.11
Lunar Lake	12.17	2.27	7.07
Cuprite	12.07	2.20	7.29

using bzip2

After

TABLE III

ENTROPIES AND ACTUAL COMPRESSION RATIOS FOR THE TEST IMAGES.

Image	ent_b	ent_a	ent_f	CR	bitrate	ratio
Jasper	6.14	5.24	5.62	2.83	5.65	1.99
Moffet	6.38	5.36	5.74	2.79	5.73	2.01
Lunar Lake	5.84	5.14	5.50	2.79	5.73	2.21
Cuprite	6.12	5.15	5.51	2.90	5.52	2.19

c.f. Clustered DPCM

(Mielikainen + Toivanen, 2003)

TABLE III

COMPRESSION RATIOS FOR THE FOUR COMBINED AVIRIS'97 IMAGES

image	[12]	New Method	JPEG-LS [12]	JPEG2000 [12]
Jasper Ridge	2.82	3.46	1.91	1.78
Moffet Field	2.94	3.46	1.99	1.82
Lunar Lake	3.23	3.37	2.14	1.96
Cuprite	3.13	3.42	2.09	1.91
Average	3.05	3.43	2.03	1.87

Comments

- For the Kaarna study, there must be significant signal in the PC residuals – otherwise wavelet transform would have no effect
- The clustered DPCM performs better, but may be too slow to be practical
- Instrument noise is not mentioned – how does it relate to the 16-bit words?
- Not obvious whether these results would apply to IR sounders such as IASI or MTG-IRS

Tony Lee's 2004 study on IASI

EUMETSAT Contract EUM/CO/03/1155/PS

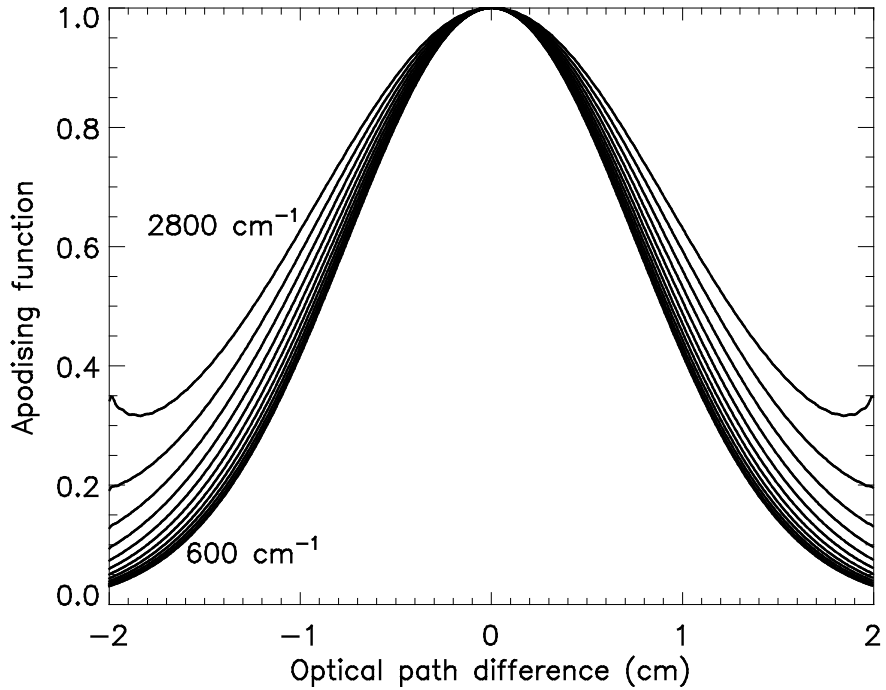
Approach was as follows:

1. Noise normalise the IASI level 1C spectra
2. Compute PC scores (predictors) and residuals (correctors)
3. Quantise the correctors at half NEΔT level
   - should increase noise by only 1%
4. Huffman encode the correctors (most probable values - close to zero - encoded with small bit length)
5. Disseminate predictors and correctors separately. Some users would only need the predictors.

Tony Lee's 2004 study (2)

- Results – for simulated IASI spectra
  - IASI predictor plus corrector volume was ~3.6 bits per channel
  - This is 22.5% of the baseline volume for 16-bit spectra – a 4.4-fold reduction
- Note:
  - The noise normalisation assumes full noise covariance matrix – equivalent to working with de-apodised spectra and a diagonal noise
  - Not fully lossless, due to the additional quantisation step, but noise increase is small
  - Outlier spectra will have a higher corrector volume (due to the Huffman encoding) but will be accurately represented

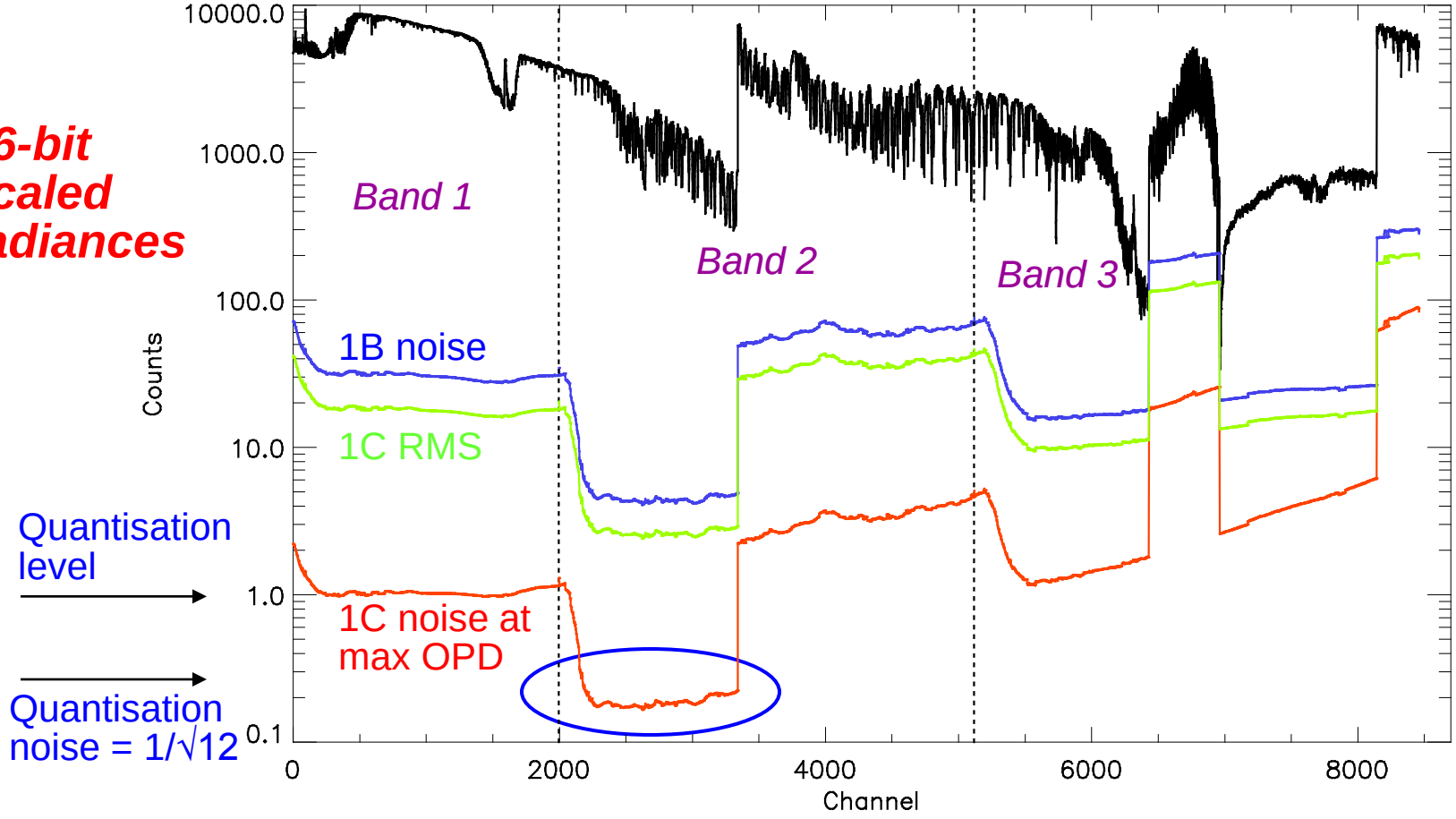
A closer look at apodisation



- Current practice (e.g. EARS-IASI) is to use 1C spectra with diagonal normalisation. Actual noise at high OPD is much lower than at low OPD – hence PC scores are probably not optimum
- This may not matter too much because it just transfers information to the correctors (which we can encode losslessly)
- However, the 1C spectrum is already degraded – next slide!

Quantization of apodised spectrum

16-bit scaled radiances



- We are all working with lossy compression!
- Does it matter in practice? (mainly band 2)

Quantisation (2)

- Quantisation level for IASI 1c seems rather arbitrary
- Important to get this right for IASI-NG and MTG-IRS
  - choosing the right quantisation is key to effective compression – regardless of which compression technique you use

Some experiments with real IASI data

In the following slides I will look at:

1. Basic formats, with standard compression
2. Try to reproduce Tony Lee's results
   - but with real data instead of simulated
3. Experiment with different quantisations
4. Compare level 1c with level 1b

Experiments with IASI data (1)

As an example, I used a direct-readout pass received at Exeter,

IASI_xxx_1C_M02_20130422084559Z_20130422085406Z_V_T_20130422085721Z

size: 166.695 MB, number of scans: 61 (= 488 sec), number of spectra: 7320

First, look at the standard formats. Reference is AAPP l1c format which uses 16 bit integers for the spectra. Size = 135.4 MB (spectra comprise 91.5% of this)

Format	**Volume w.r.t. AAPP l1c**
Native PFS	1.231 *Larger than AAPP due to IIS etc.*
PFS + gzip	0.912 (11 sec)
PFS + xz	0.735 (85 sec)
PFS + bz2	0.722 (25 sec)
BUFR	0.680
AAPP l1c + bz2	0.627

- bz2 has little effect if applied to the BUFR file – as expected
- But AAPP format – which contains the same information – can be compressed below the BUFR data volume

Experiments with IASI data (2)

Now try some simple channel difference techniques, using the AAPP l1c data

Format	**Volume w.r.t. AAPP l1c**
AAPP l1c + bz2	0.627 (from previous table)
Difference from prev channel + bz2	0.586
Difference from mean + bz2	0.547
Diff from mean then diff from prev channel + bz2	0.516 *18% reduction*

- This is surprising – I would have expected bzip2 to do operations like this automatically
- Only effective with 1c data – no significant effect on 1b (see later slide)
  - apodisation produces noise correlations

Experiments with IASI data (3)

- PC compression

Format	Volume w.r.t. AAPP l1c	Volume w.r.t. BUFR l1c
AAPP PC format + bz2 (300 PCs) + bz2	0.062 *PCs only*	
16-bit residuals + bz2	0.406 *Residuals only*	
PCs + 16-bit residuals + bz2	0.469 *Lossless*	0.690
PCs + residuals quantised at 0.5 x NEΔT	0.230 *Near Lossless?*	0.338

- This is consistent with Tony Lee's result for simulated spectra
- As noted previously it is *not* near-lossless if we are using apodised spectra
- The compression w.r.t. l1c BUFR is a factor 3

Experiments with IASI data (4)
- change the quantisation

How close can we get if we don't use PCs, but just quantise the basic l1c spectra at 0.5 x NEΔT?

Format	**Volume w.r.t. AAPP l1c**
AAPP l1c quantised at 0.5 x NEΔT + bz2	0.318

- As expected, this is not as good as PCs + residuals (0.230), but much better than the basic l1c (0.63)
- Has the advantage of being a simple scheme
- *Most of the benefits of near-lossless compression comes from the quantisation.*

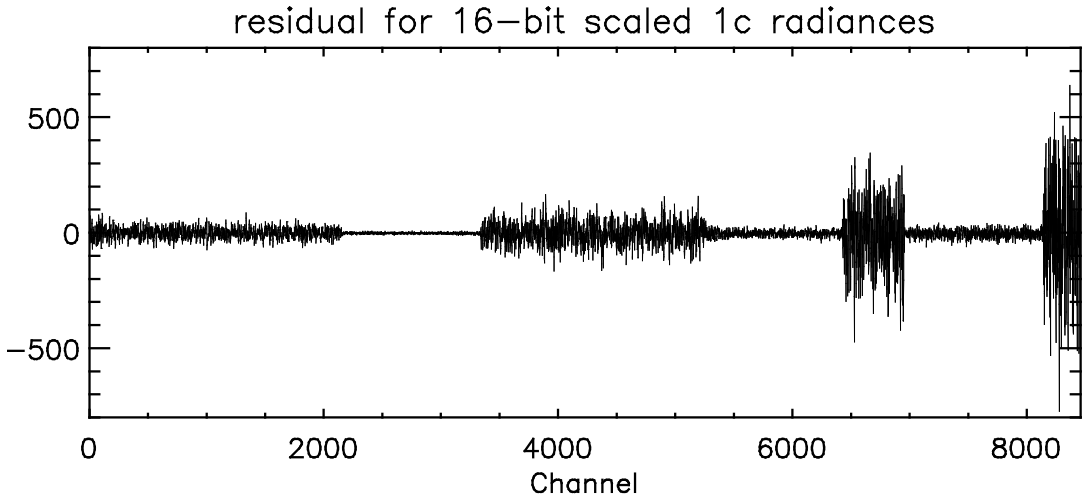
Experiments with IASI data (5)

Comparison of l1b and l1c (a different pass – 20/09/2013)

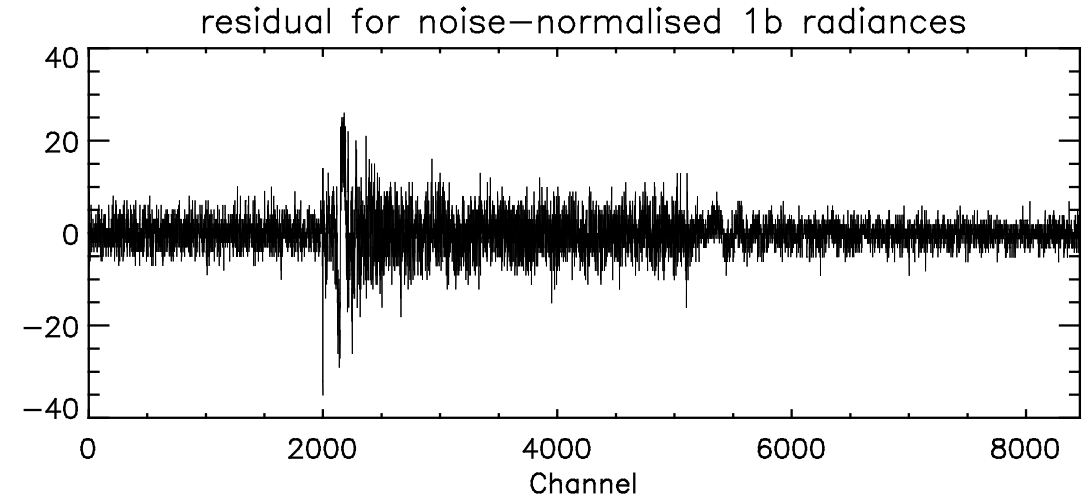
Format (all with bz2)	Volume w.r.t. AAPP (1b)	Volume w.r.t. AAPP (1c)
AAPP l1b/l1c + bz2	0.670	0.643
Quantised at 0.5xNEΔT + bz2	0.359	0.372
Quantised at 0.5xNEΔT + channel diff	0.358	0.347
Quantised at 0.5xNEΔT + diff from mean	0.347	0.280
Quantised at 0.5xNEΔT + channel diff + diff from mean	0.341	0.276
PCs + residuals quantised at 0.5xNEΔT	0.300	0.200
PCs + residuals quantised at 0.125xNEΔT	N/A	0.309

- Used EUMETSAT covariance matrix to compute new PCs for l1b
- Consistent with expectation that 1b retains more information than 1c
- 1b result could be considered near-lossless

Residuals



typical spectrum



Summary for IASI

l1c – lossless

1. l1c differences from mean, with channel differences + bz2
   - Volume 0.76 *all on this slide are compared with l1c BUFR*
2. PCs + 16-bit residuals + bz2
   - Volume 0.69

l1c – compromise

3. PCs based on l1c + residuals quantised at 1/8 NEΔT + bz2
   - Volume 0.46

l1b – near lossless, with 1% noise increase

4. l1b, quantised at half NEΔT + bz2
   - Volume 0.53
5. PCs based on l1b + quantised residuals + bz2
   - Tony Lee's recommendation
   - Volume 0.44 (c.f. 0.33 estimated by Tony with simulated data)

Warning: PCs based on l1b may not suit users of the PC-only product (Tim Hultberg did some studies on this)

Implications for IASI-NG

- IASI-NG will have doubled spectral resolution compared with IASI, and lower noise
- Design of 1c format needs care, to avoid loss of information
- Design datasets with compression in mind – not as an after-thought
- Likely that compressed data volume will be more than double that of IASI

A look at MTG-IRS

Figures derived from the spec in MTG Mission Requirements Document, plus information from MIST IX presentation

	IASI	MTG-IRS
Spectral sampling	0.25 cm^{-1}	0.625 cm^{-1}
Samples per spectrum	8461	1738 (2 bands)
Spatial sampling at nadir	25 km	4 km
Spectra per hour	54000	8.0×10^6 (full disk every hour)
Samples per hour	4.6×10^8	1.4×10^{10}
Bits per sample, for L1B, 0.5× NEΔR quantisation + diff from mean + bzip2	5.6	5.6 (using NEΔR spec)
GB per hour	0.32	9.7

coincidentally the same

Factor 30 higher than IASI, but less than the 32GB/h estimated by EUM for uncompressed data

- MTG-IRS will need lossy compression for dissemination (baseline is PC scores)
- A role for near-lossless compression in the L1 format used for archiving, etc. e.g. *appropriately scaled integers, with built-in netCDF-4 compression*

Conclusions

- Results for IASI:
  - For near-lossless compression, best is a factor of 2.2 compared with current BUFR product
- Conversion of radiances to appropriately-scaled integers is key
- Don’t degrade your radiances before you have to!
  - Implications for ground segment design
- Advantage of PC + residuals is that some users would only need the PCs
- Specialised techniques for hyperspectral imagery (e.g. AVIRIS) not tried for IASI (as far as I know) but my guess is it’s unlikely they would significantly improve on the above results
  - because PCA is effective in shifting most of the signal into the low-volume PC scores. Residuals are mostly noise.
- Is it best to use real instrument noise for the normalisation, or should we use something else, e.g. a noise spec – which would give potential for higher compression? *Discuss.*

Thank you for listening!

Questions and discussion?

nigel.atkinson@metoffice.gov.uk