

Исследование и сравнительный анализ методов сжатия данных в задачах прогноза погоды и моделирования климата

Научный руководитель - Гойман Гордей Сергеевич, к.ф.-м.н., ИВМ РАН

Студент - Кузнецов Иван, МФТИ ФПМИ Б05-205

Обзорный доклад - 17.03.2025

План доклада

- Постановка задачи
- Краткое напоминание модели
- Теоретическое объяснение
- Метрики и их возникновение
- Текущие и ожидаемые результаты

Постановка задачи

- Ежегодно центры метеорологических и климатических исследований создают огромные массивы данных, измеряемые сотнями петабайт. Такой объем невозможно хранить
- Необходимо эффективное сжатие для упрощения хранения и передачи данных
- Наша цель:
 - Адаптация существующих алгоритмов сжатия к отечественным форматам данных
 - Сравнение алгоритмов на практике
 - Представление общих рекомендаций по выбору метода для конкретных видов данных

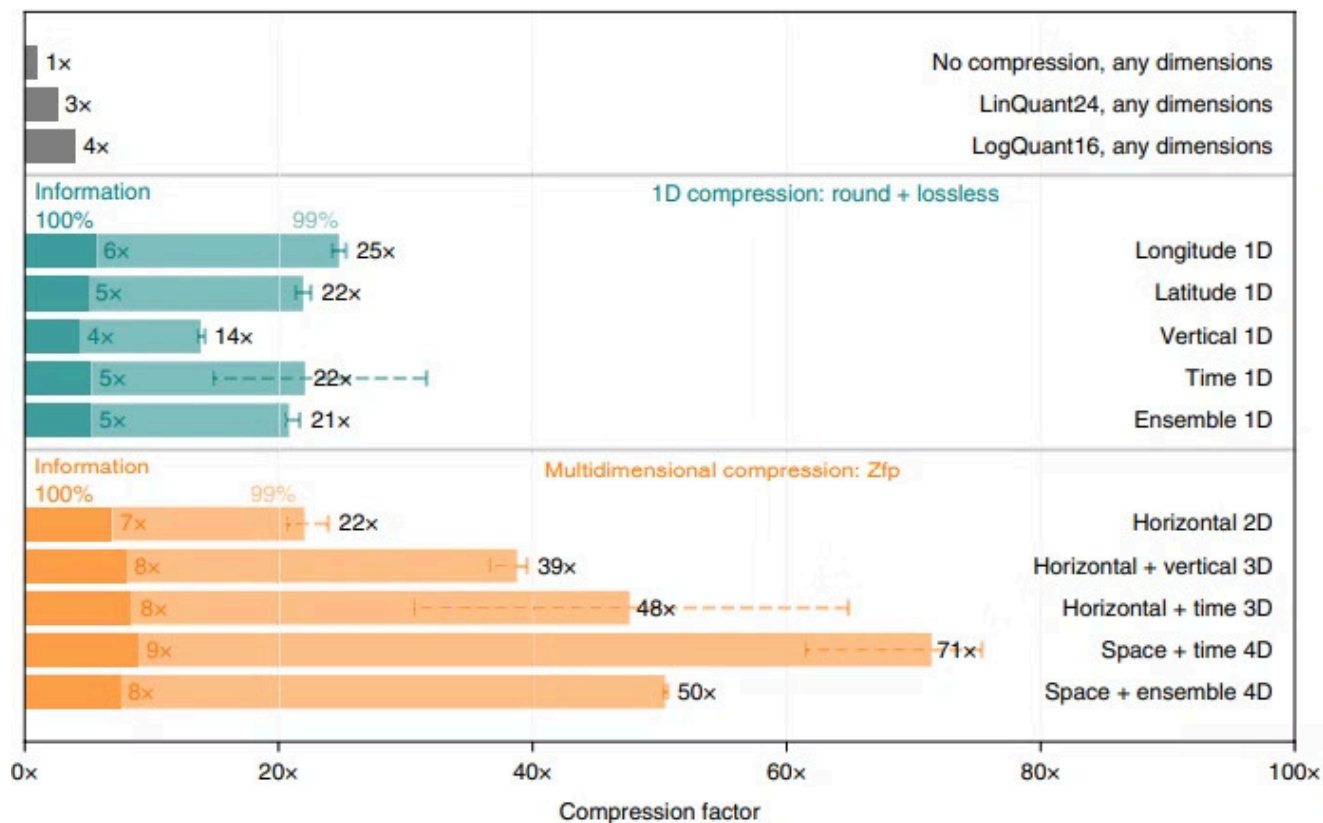
Данные



- система ПЛАВ — модель прогноза погоды
- Шаг сетки 10 и 20 км
- Ансамблевый прогноз
- Вес до сжатия ~ 23GB данные + 5GB прогноз

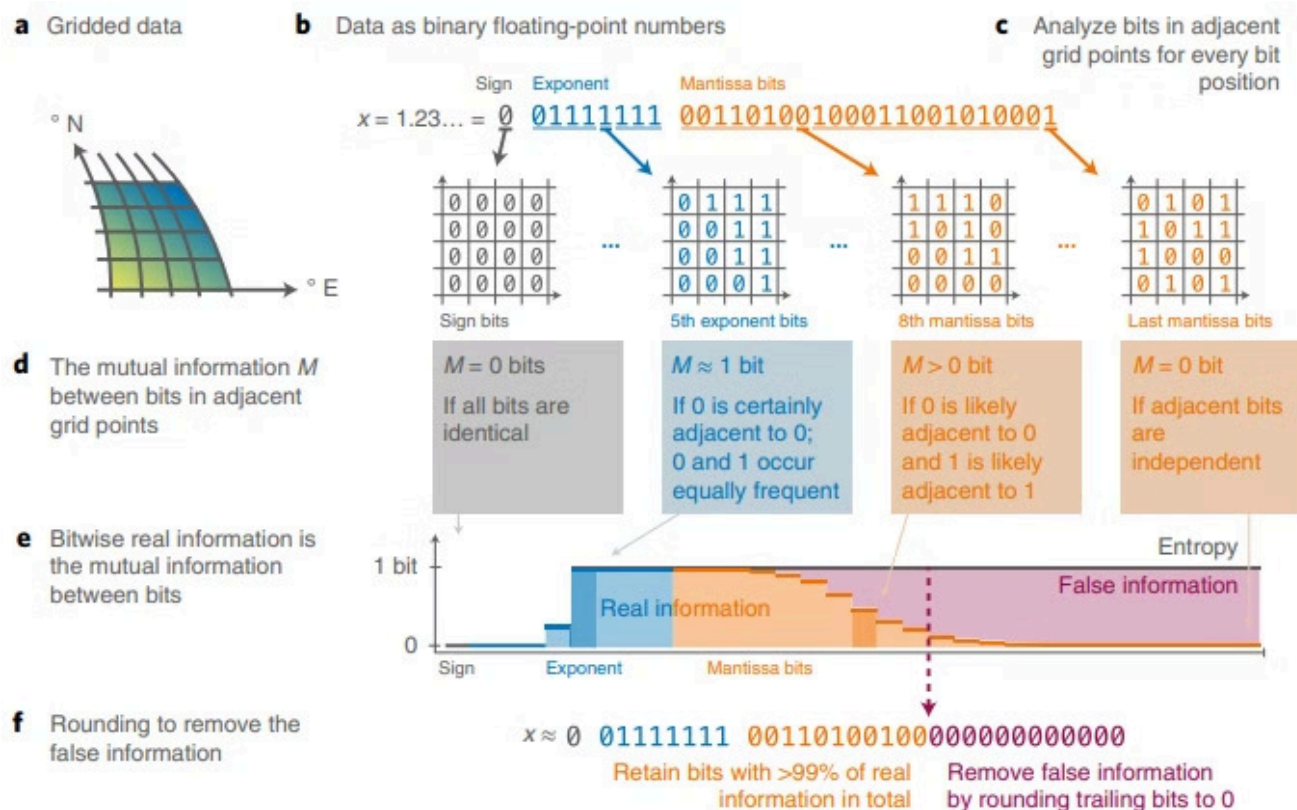
Методы и подходы

- Большинство гидрометцентров используют линейное и логарифмическое квантование — **"простое фиксированное округление"**
- Существует много различных подходов. [\[3\]](#)
- Один их методов: **"умное округление"**



Сжатие с сохранением реальной информации о температуре в различных измерениях [\[1\]](#)

Сжатие только реальной информации

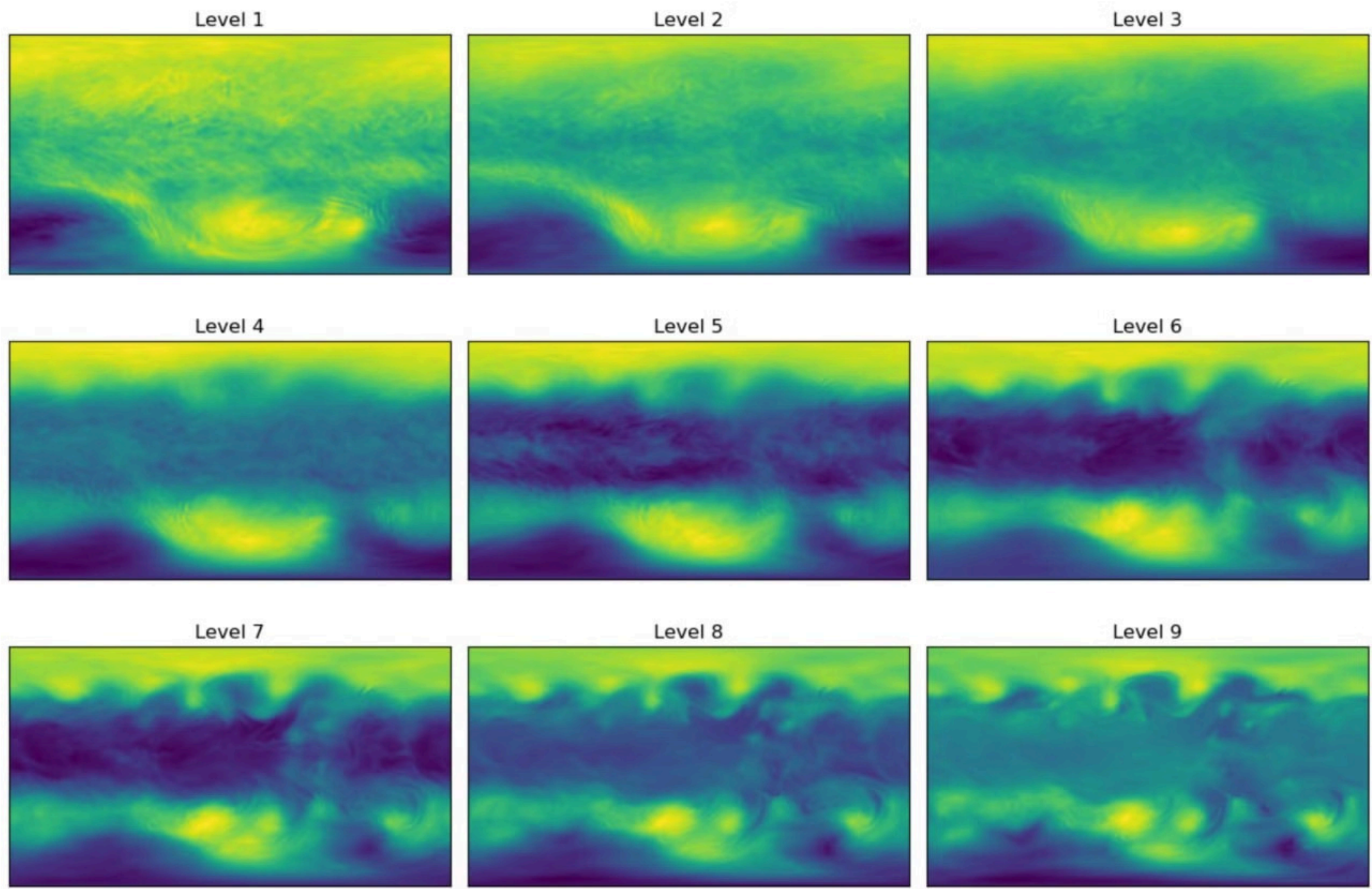


Побитовое содержание реальной и случайной информации [1]

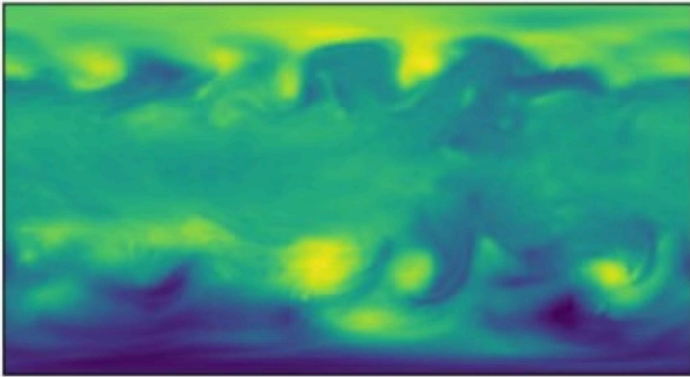
- Часть битов в записи числа может быть случайными и не нести реальной информации
- Можно оценить рассматривая корреляцию по пространству
- Зануление случайных битов позволяет эффективно сжимать применяя lossless алгоритмы
- Итог: малая потеря данных при занулении случайных битов и последующее lossless сжатие

Воспроизводимость для наших данных

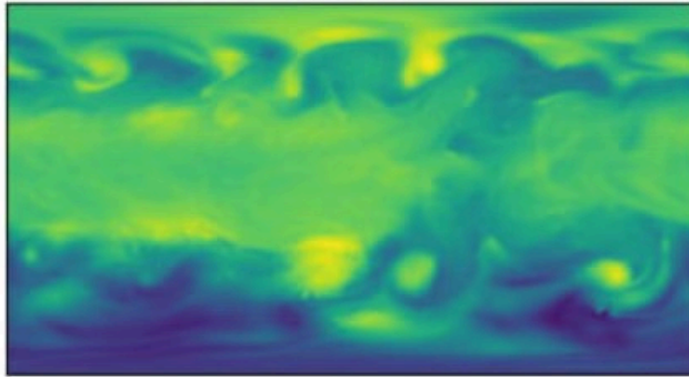
Профиль поля температур изобарических уровней до компресии



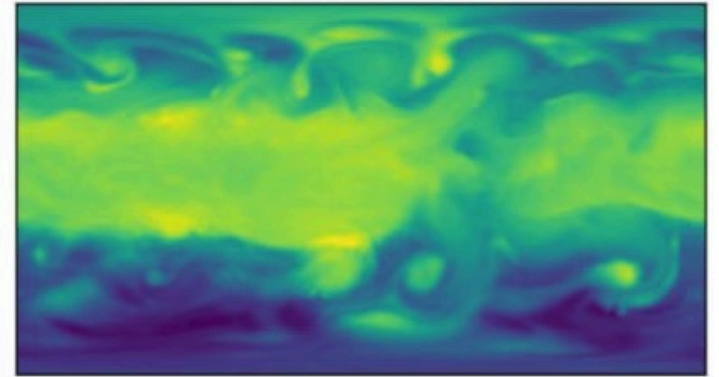
Level 10



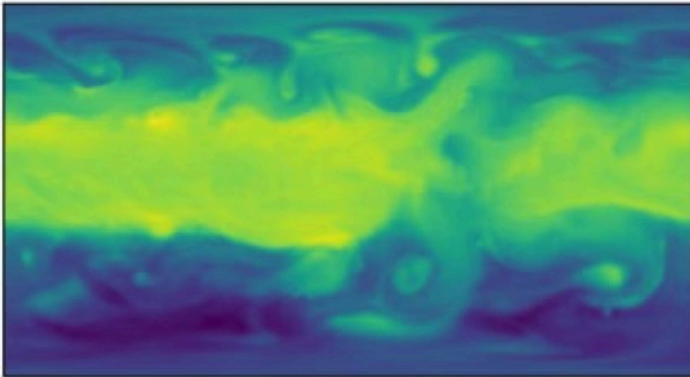
Level 11



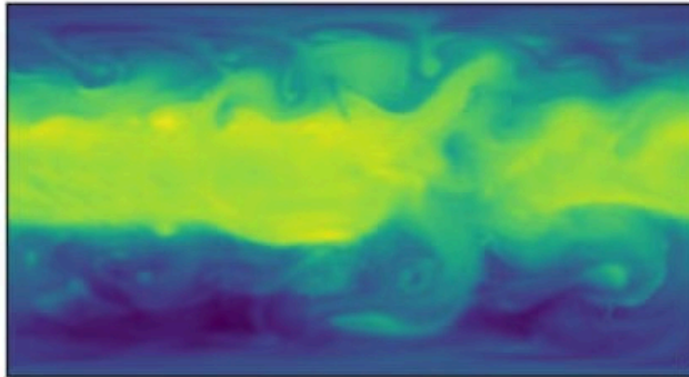
Level 12



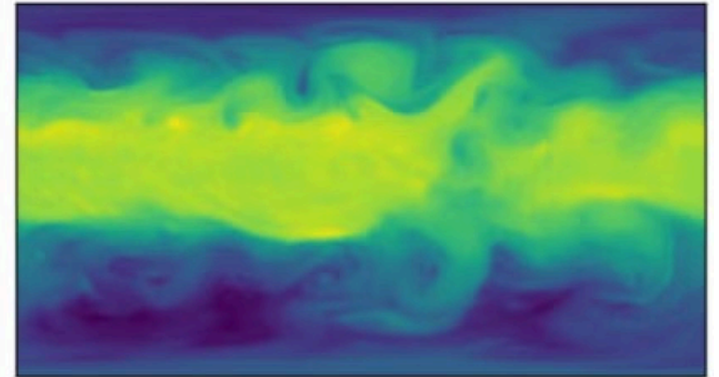
Level 13



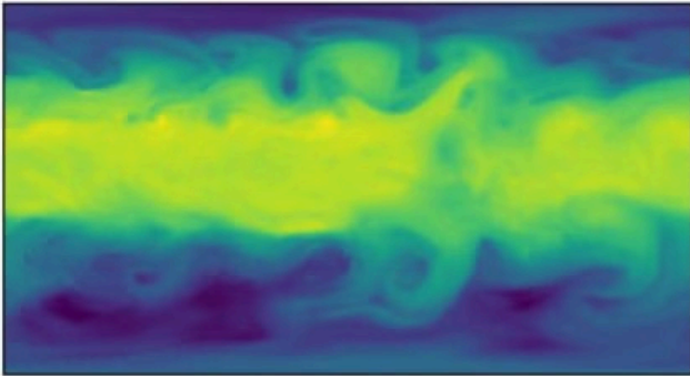
Level 14



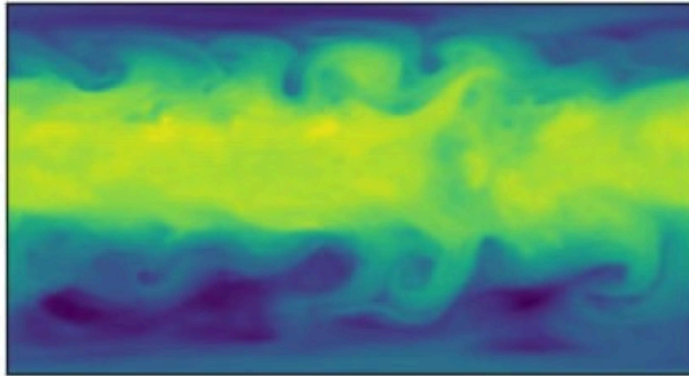
Level 15



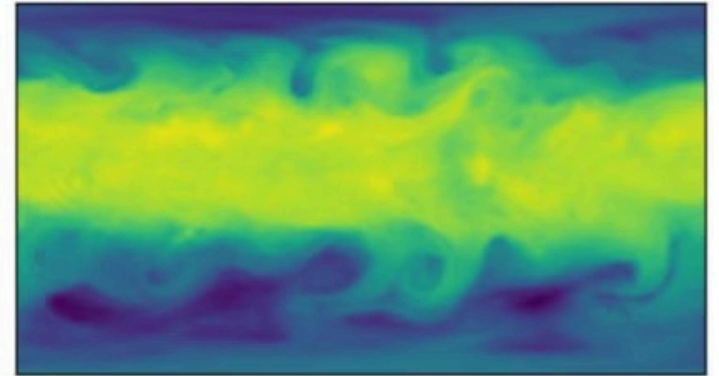
Level 16



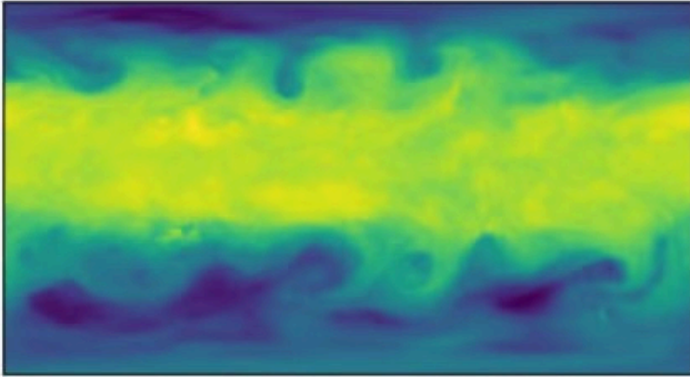
Level 17



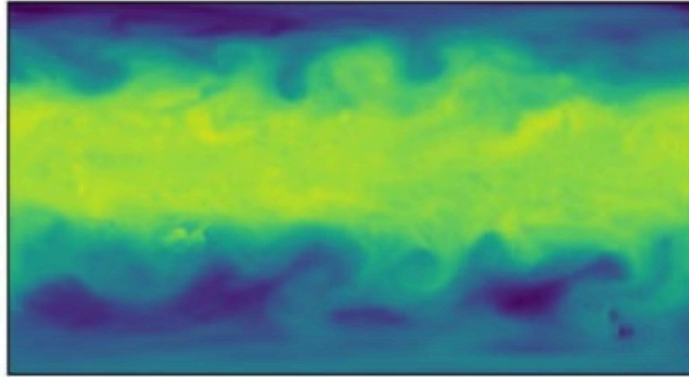
Level 18



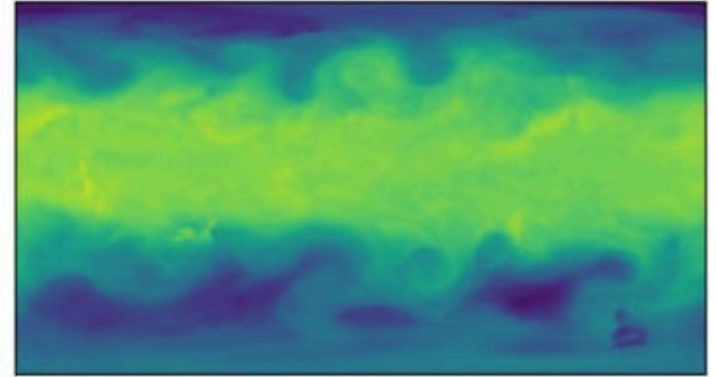
Level 19



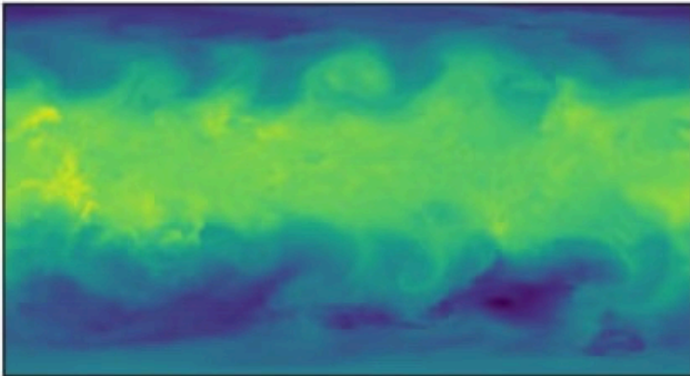
Level 20



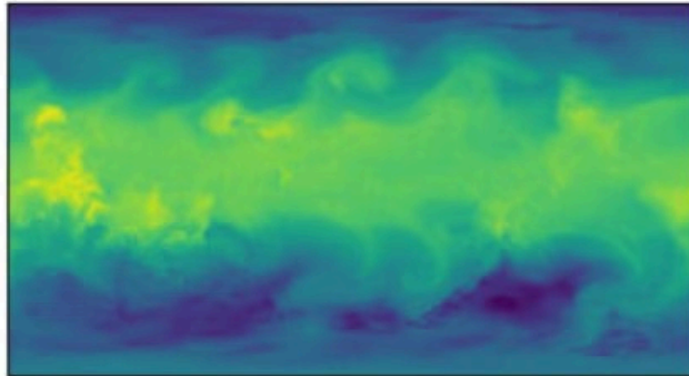
Level 21



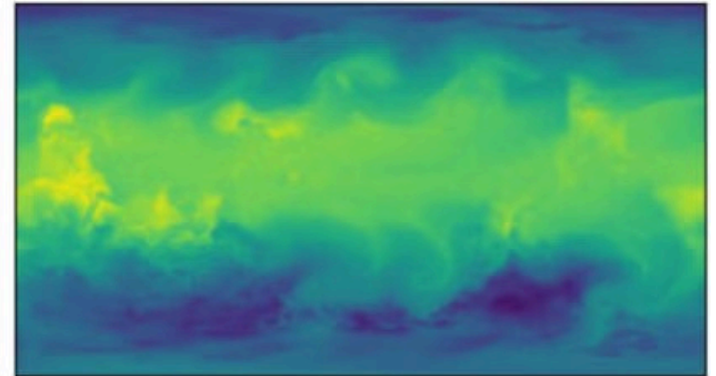
Level 22



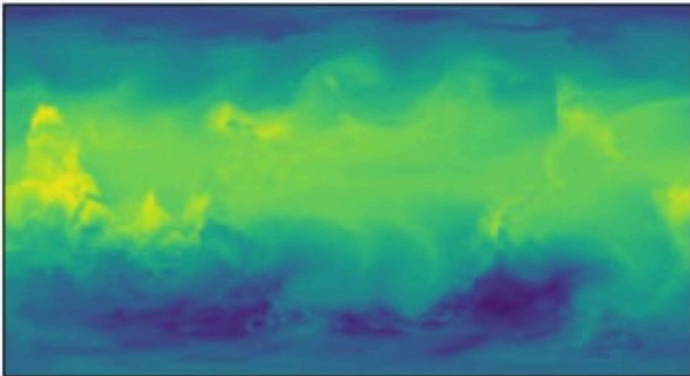
Level 23



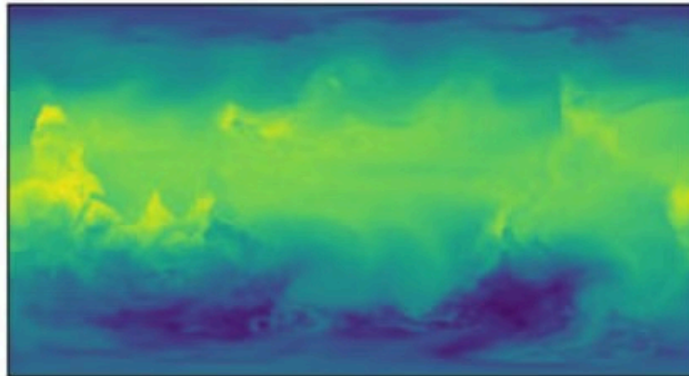
Level 24



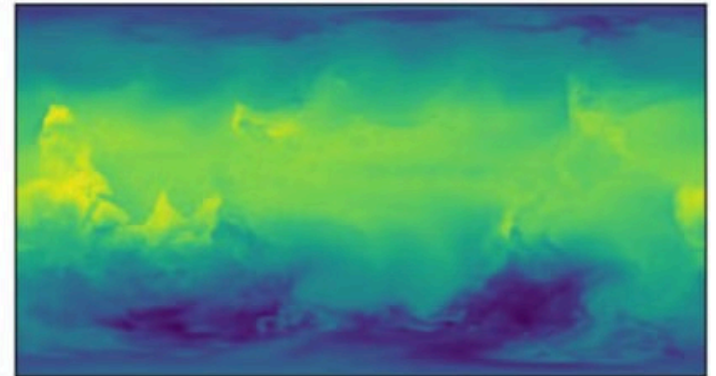
Level 25



Level 26

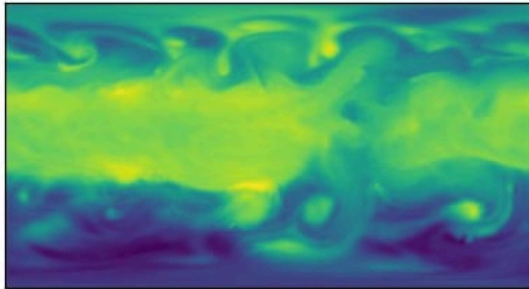


Level 27

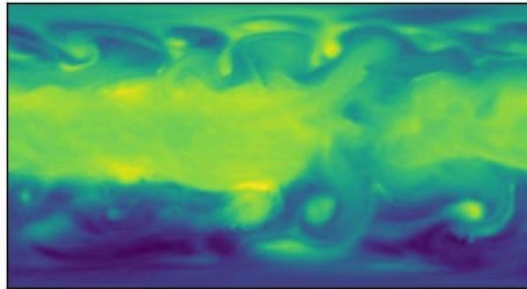


Для 12 уровня наглядная иллюстрация сжатия

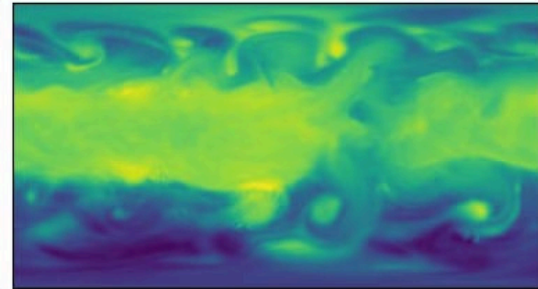
100% information



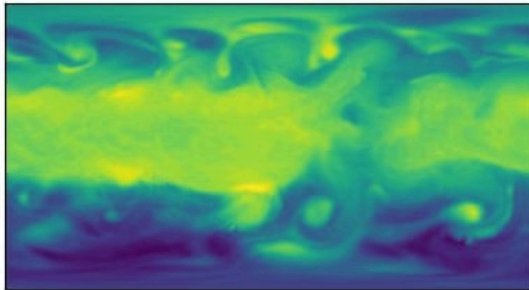
90% information



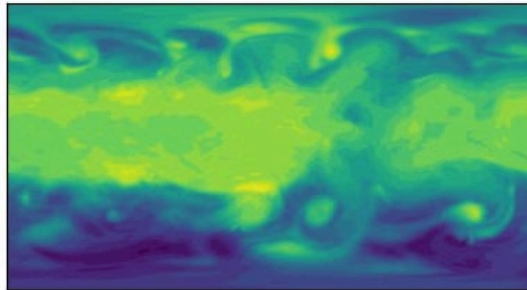
80% information



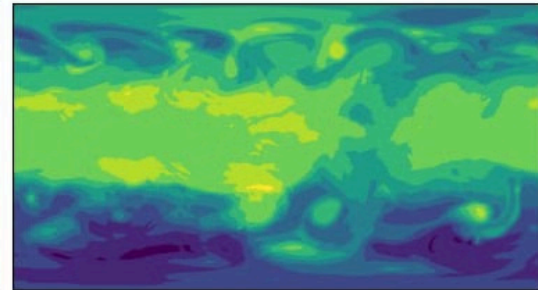
70% information



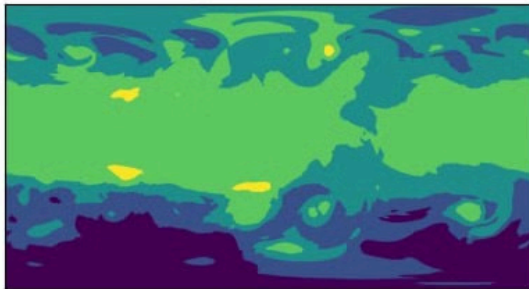
60% information



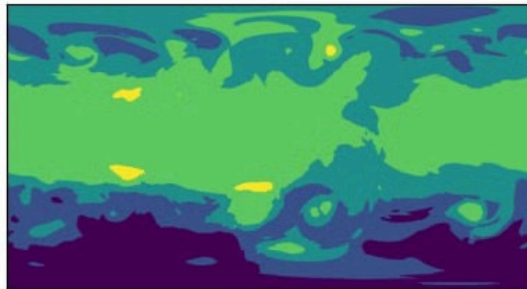
50% information



40% information



30% information



20% information



Теоретическое обоснование

$$H = \sum_{i=1}^{2^n} p_i \log_2 p_i - \text{энтропия Шенона - формализация понятия "Информация"}$$

$$H \leq n \rightarrow H = n \iff \text{uniform distribution}$$

$$H_f = n - H - \text{свободная энтропия}$$

$$\text{for bitstream : } b = b_1 b_2 \dots b_k \dots b_l \rightarrow H = -p_0 \log_2 p_0 - p_1 \log_2 p_1$$

with p_0, p_1 being the probability of a bit b_k in b being 0 or 1

Для двух разных узлов у нас есть два потока $r = r_1 r_2 \dots r_k \dots r_l$ И $s = s_1 s_2 \dots s_k \dots s_l$

Взаимная информация:

$$p_{rs} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

Взаимная информация двух потоков:

$$M(r, s) = \sum_{r=0}^1 \sum_{s=0}^1 p_{rs} \log_2 \left(\frac{p_{rs}}{p_{r=r} p_{s=s}} \right)$$

Разделим один поток на два

$$r = b_1 b_2 \dots b_{l-1}; s = b_2 b_3 \dots b_l$$

Безусловная энтропия

$$H = H(r) = H(s)$$

Условная энтропия

$$H_0 = -p_{00} \log_2 p_{00} - p_{01} \log_2 p_{01}$$
$$H_1 = -p_{10} \log_2 p_{10} - p_{11} \log_2 p_{11}$$

Ложная информация $\leftarrow$ Максимум энтропии $\leftarrow$ вероятность значения бита не зависит от состояния предыдущего бита

реальная информация : энтропия - ложная

$$I = H - p_0 H_0 - p_1 H_1$$

Сохранение информации

избыточная информация в s при аппроксимации r

$$R(r, s) = \frac{2M(r, s)}{H(r) + H(s)}$$

сохраненная информация

$$P(r, s) = R(r, s)I(r)$$

обобщение на массивы X и Y (compressed)

$$P(X, Y) = \frac{\sum_{q=1}^n R(x_q, y_q)I(x_q)}{\sum_{q=1}^n I(x_q)}$$

Метрики

- Базовые: rmse, mae, max abs error, structure error
- Время работы
- Тест Тьюринга
- Коэффициент сжатия данных
- Коэффициент сохранения информации P (зависимость от других метрик)
- Дивергенция Кульбаха - Лейблера

Дальнейший план исследования

- продолжение обзора методов
- сравнение методов
- подбор оптимального коэффициента {сжатие}/{потери} для разных видов набора данных
- поиск эвристик

Ссылки

- [1] методы, основанные на анализе реального информационного содержания
 - <https://doi.org/10.1038/s43588-021-00156-2>
- [2] Стандартные алгоритмы сжатия с потерями
 - <https://doi.org/10.5194/gmd-17-8909-2024>
- [3] нейросетевые алгоритмы
 - <https://arxiv.org/abs/2410.09109>
 - <https://arxiv.org/abs/2407.11666>

Спасибо за внимание!

Кузнецов Иван

kuznetsov.ia@phystech.edu

tg: @stop_pleaze