

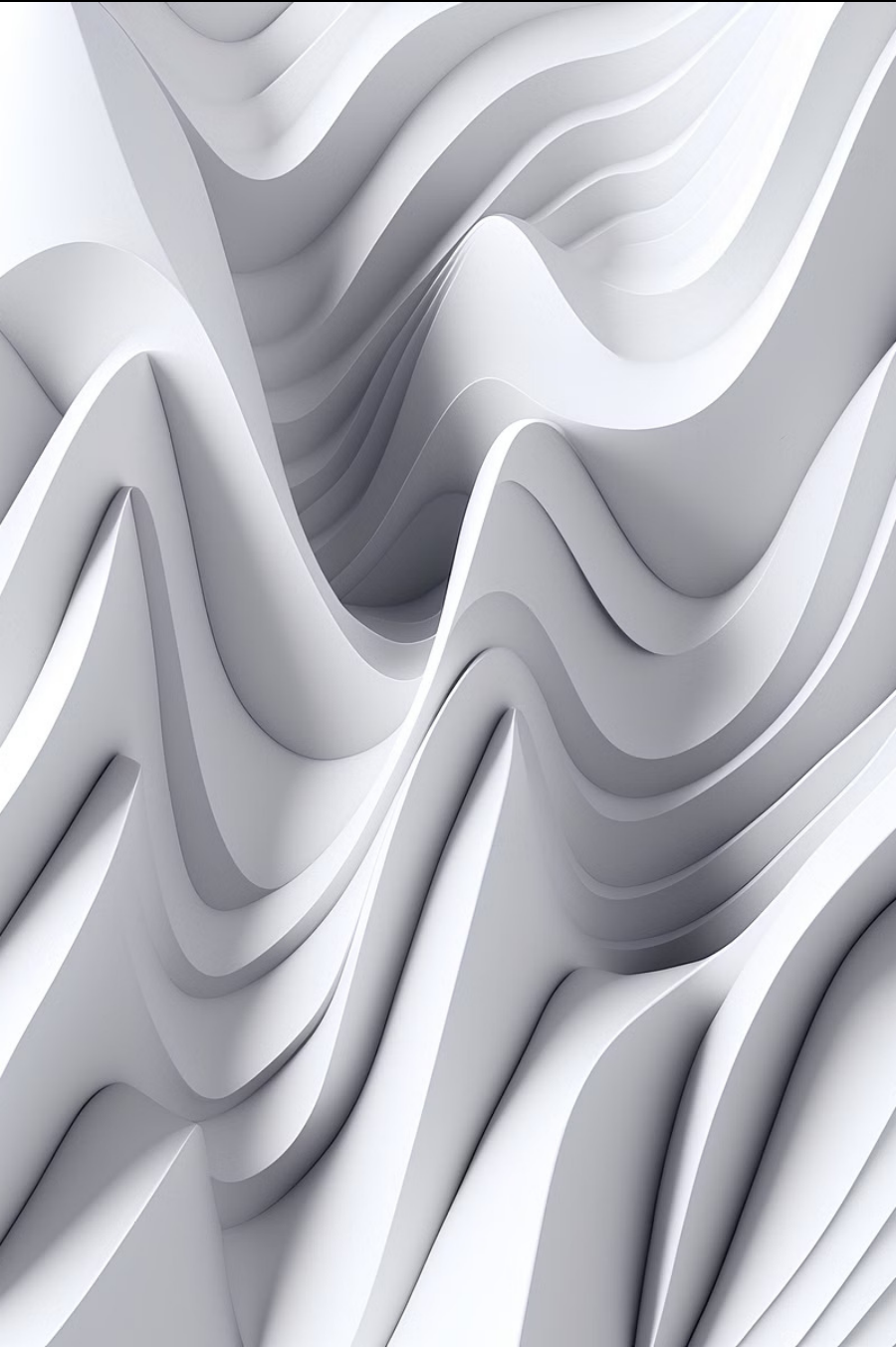


Исследование и сравнительный анализ методов сжатия данных в задачах прогноза погоды и моделирования климата

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

Студент - Преин Дмитрий, группа Б05-205

Первый доклад - 11.03.2025



План доклада

- Постановка задачи
- Описание модели
- Описание данных
- Ожидаемые результаты
- Актуальные методы
- Текущие результаты
- Дальнейшее исследование

Постановка задачи

- Ежегодно центры прогнозирования погоды и климата генерируют сотни петабайт данных
- Сжатие данных требуется для сокращения объема хранения и упрощения обмена информацией
- Цель исследования - Изучение и сравнение существующих методов сжатия метеорологических данных
- Практическая значимость - Результаты исследования помогут определить оптимальные методы сжатия для работы с большими объемами метеорологических данных



Описание модели



- работаем с данными модели ПЛАВ среднесрочного и долгосрочного прогноза погоды из Гидрометцентра России и ИВМ РАН
- ПЛАВ с шагом сетки 10 км -- среднесрочный и авиапрогнозы
- ПЛАВ с шагом 20 км -- ансамблевый прогноз

Tolstykh M. et al. Vorticity-divergence semi-Lagrangian global atmospheric model SL-AV20: dynamical core // Geoscientific Model Development. – 2017. – Т. 10. – №. 5. – С. 1961-1983. <https://doi.org/10.5194/gmd-10-1961-2017>

Описание данных

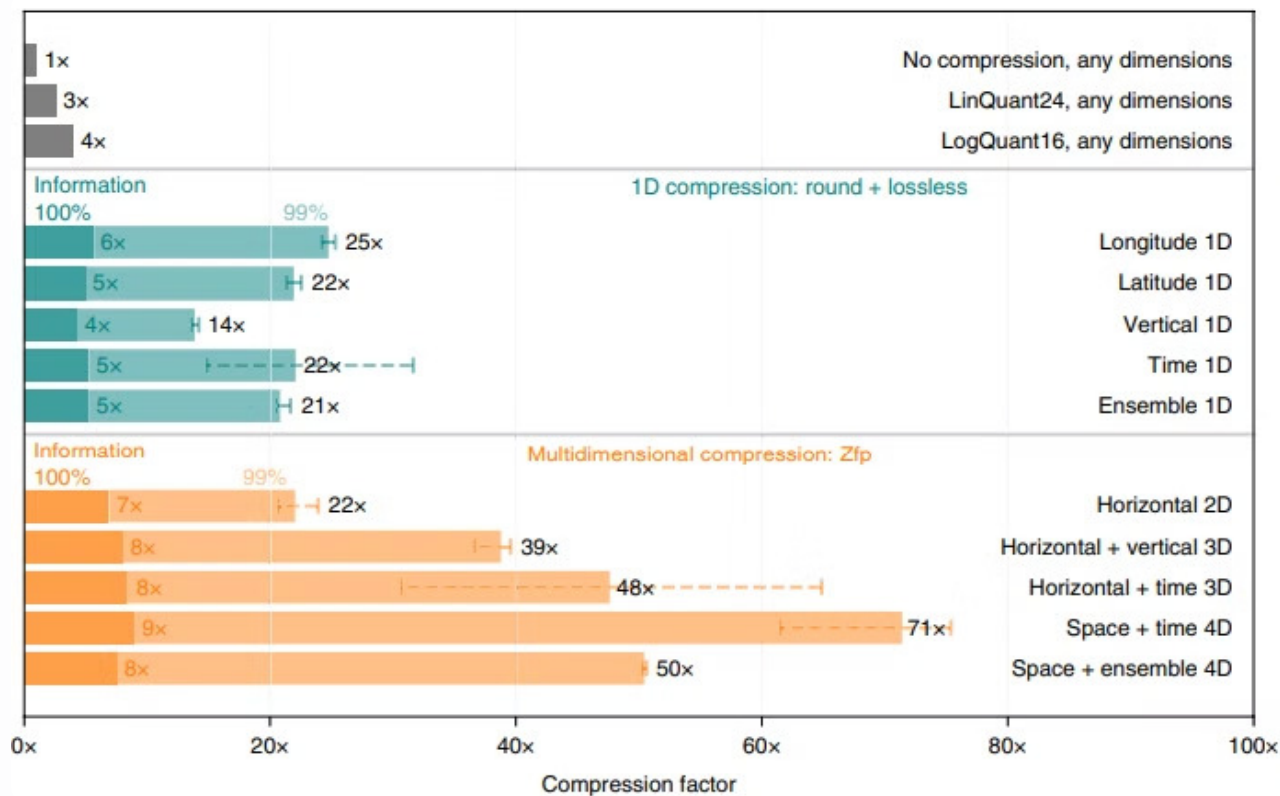
Configuration id	SLAV072L96	SLAV20	SLAV10
Number of grid-nodes in longitude	400 (0.9°)	1600 (0.225°)	3600 (0.1°)
Number of grid nodes in latitude	251 (0.72°)	865	1945
Number of vertical levels	96	51	104
Time step, s	1440	450	270
Number of time-steps per model day	60	192	320
Size of the compressed control-point file, Gb	0.3	1.9	9.9
Size of the raw control-point file, Gb	0.3	2.8	22.7
Size of the output file, Gb	—	0.7	3.3

сейчас используется сжатие на основе bit-grooming, реализованное в рамках стандартного пакета nco

Ожидаемые результаты

- Ранжирование методов по эффективности
- Оценка влияния сжатия на точность прогноза и других приложений
- Рекомендации по выбору алгоритмов для конкретных типов данных и задач
- Программное обеспечение для сжатия и восстановления выходной продукции и начальных данных метеорологических моделей.

Актуальные методы

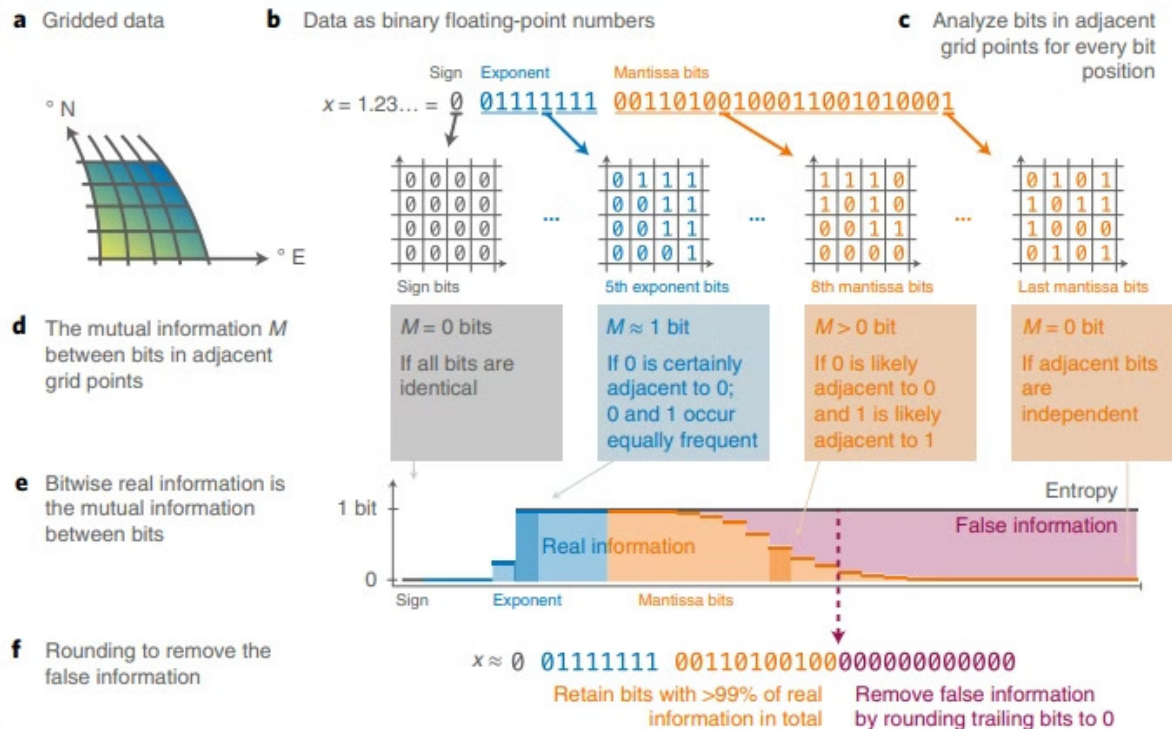


Сейчас большинство гидрометцентров используют линейное и логарифмическое квантование, которые хоть и ограничивают абсолютную ошибку, но подходят далеко не для всех данных

Мы ожидаем большего сжатия с использованием альтернативных подходов, которые получили развитие в последние годы

Сжатие реальной информации о температуре в различных измерениях [1]

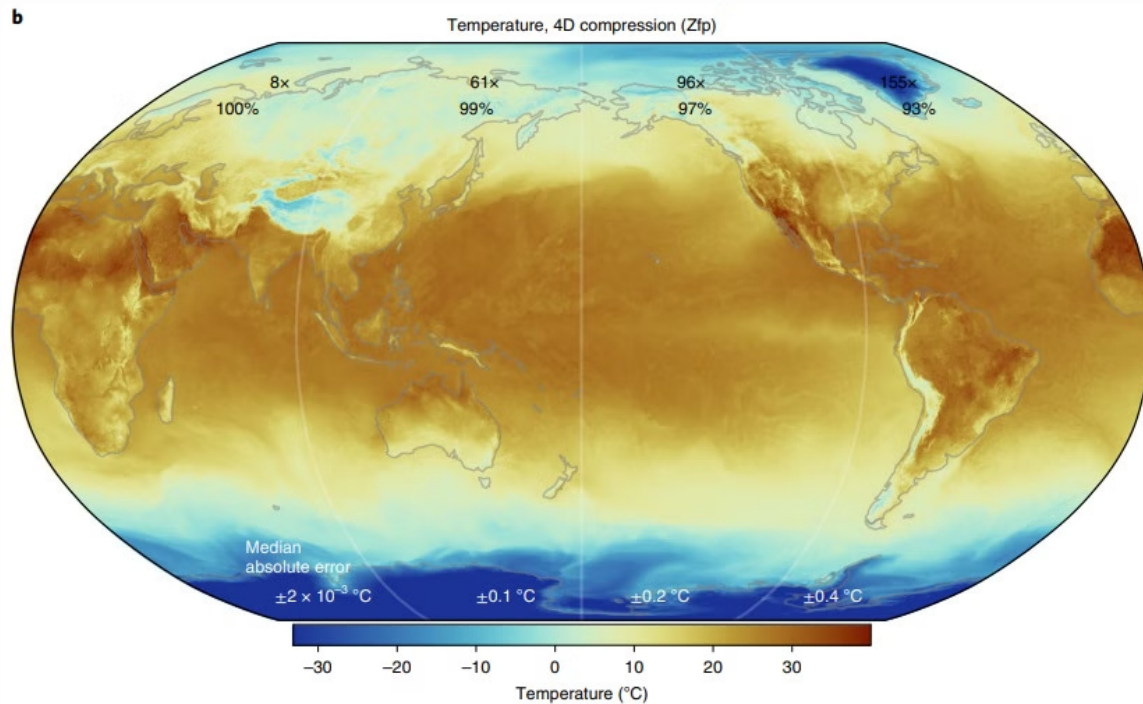
Сжатие только реальной информации



- 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

Побитовое содержание реальной информации [1]

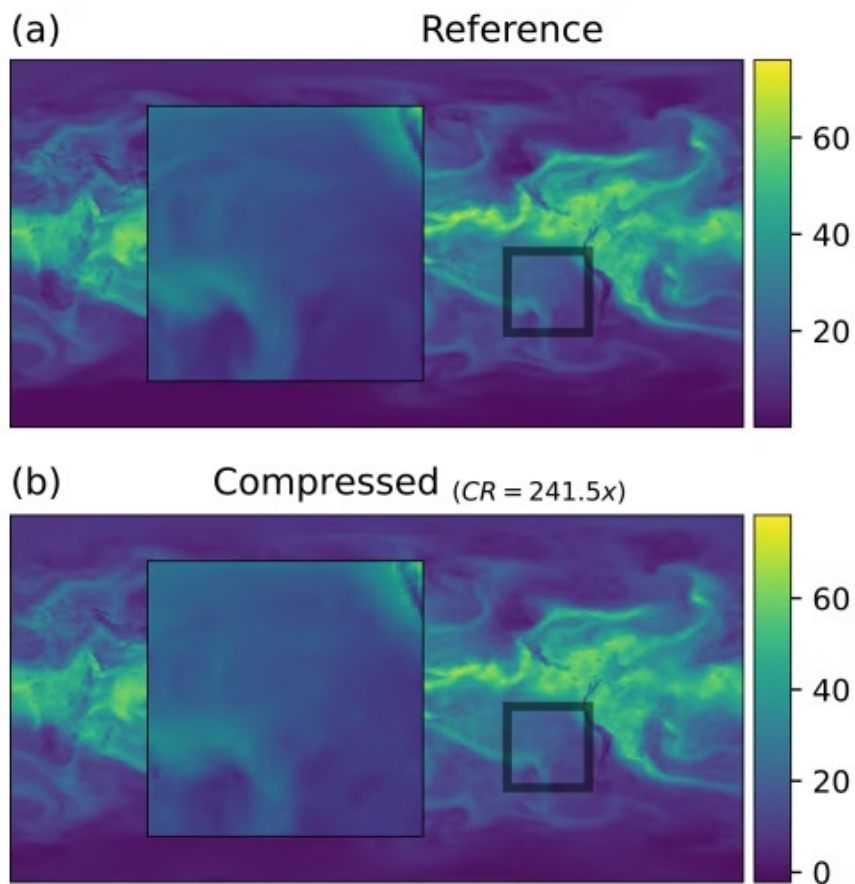
Алгоритм Zfp



- Метод разбивает d-мерный массив на блоки размерности 4d
- Разбиение осуществляется с учетом корреляции данных в каждом измерении
- Внутри каждого блока применяется преобразование
- Преобразование выполняется с заданной абсолютной или относительной погрешностью

Сжатие на различных уровнях сохраняемой информации [1]

Алгоритм SZ



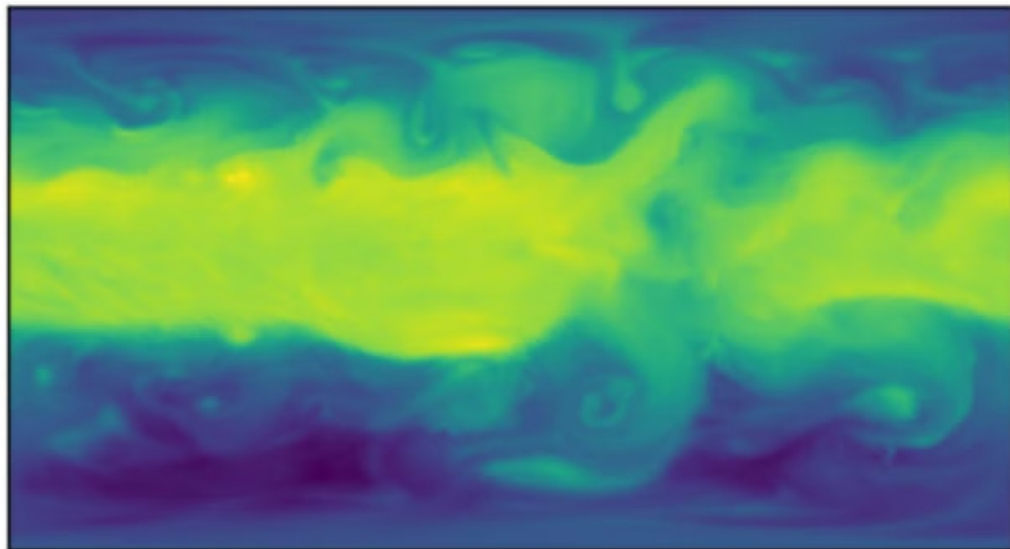
сжатие с помощью SZ [2]

- Алгоритм работает с данными блоками
 - Для **трехмерных данных** используются блоки размером $6 \times 6 \times 6$ точек сетки
 - Для **двухмерных данных** применяются блоки 12×12 точек
- Алгоритм поддерживает три режима сжатия
 - **Абсолютный режим**: задается абсолютная погрешность
 - **Относительный режим**: задается относительная погрешность
 - **Относительный по точкам**: учитывается относительная погрешность для отдельных точек данных

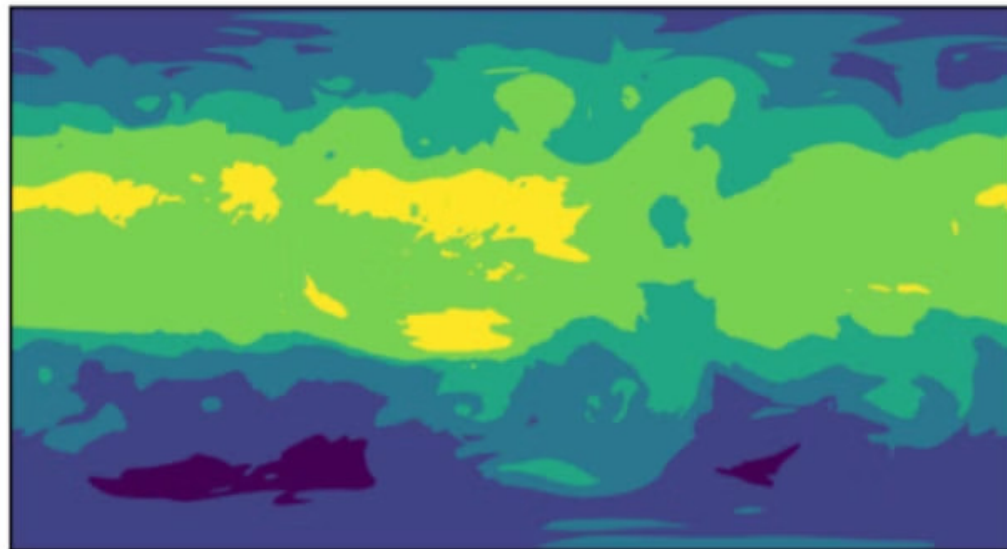
Текущие результаты

- изучили актуальные статьи на тему
- Разобрались в работе алгоритмов, основанных на анализе реального информационного содержания и стандартных алгоритмов сжатия с потерями
- Провели первичное тестирование алгоритмов на наших данных

Temperature, uncompressed



Temperature, 30% information preserved



Дальнейшее исследование

- Построить разные метрики ошибки для сжатых полей
- Анализ нейросетевых алгоритмов и методов, основанных на применении тензорных разложений и аппроксимациях наименьшего ранга
- Исследование влияния степени сжатия на различные метрики качество полей
- Исследование влияния степень сжатия начального состояния модели на качество прогноза
- Выработка методики для определения оптимального алгоритма и степени сжатия для различных полей и приложений



ССЫЛКИ

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



Спасибо за внимание!

Дмитрий Преин

prein.ds@phystech.edu

Tg: [@perin_nn](#)