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

Итоговый доклад - 20.05.2025

План доклада

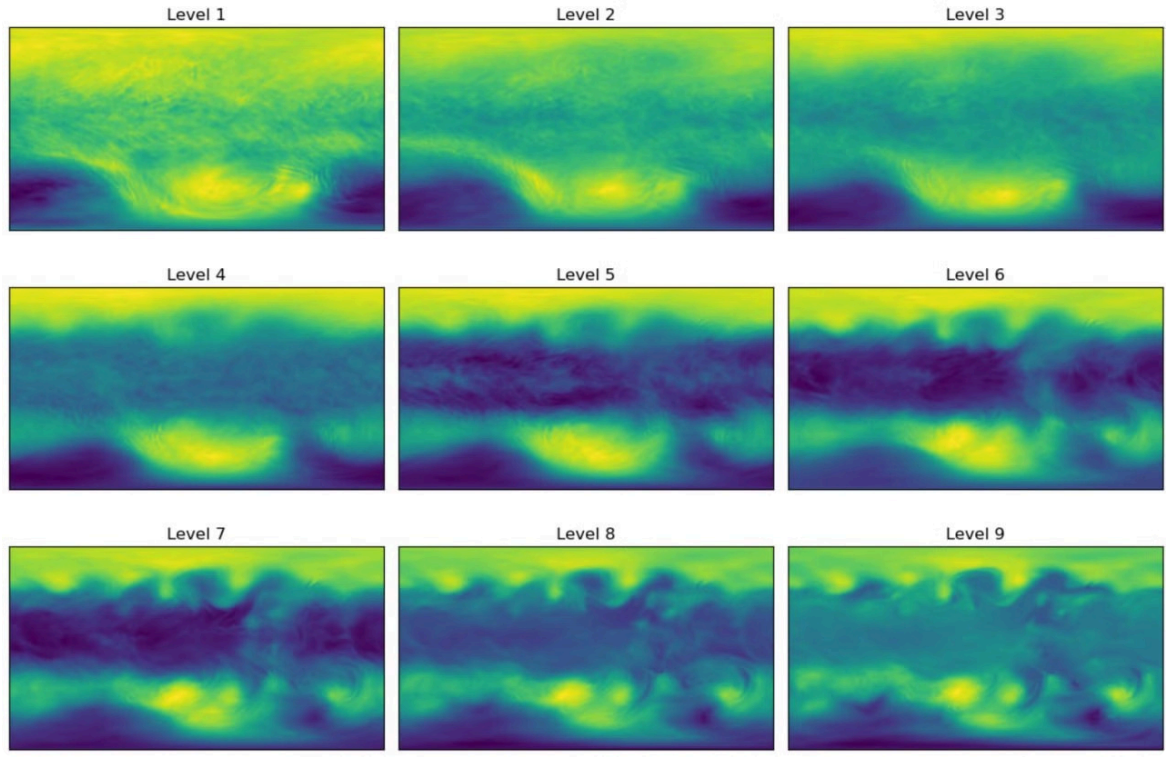
- Мотивация
- Краткое напоминание модели
- Формат данных и утилита NCO
- Алгоритмы сжатия в NCO
- Сравнительный анализ

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

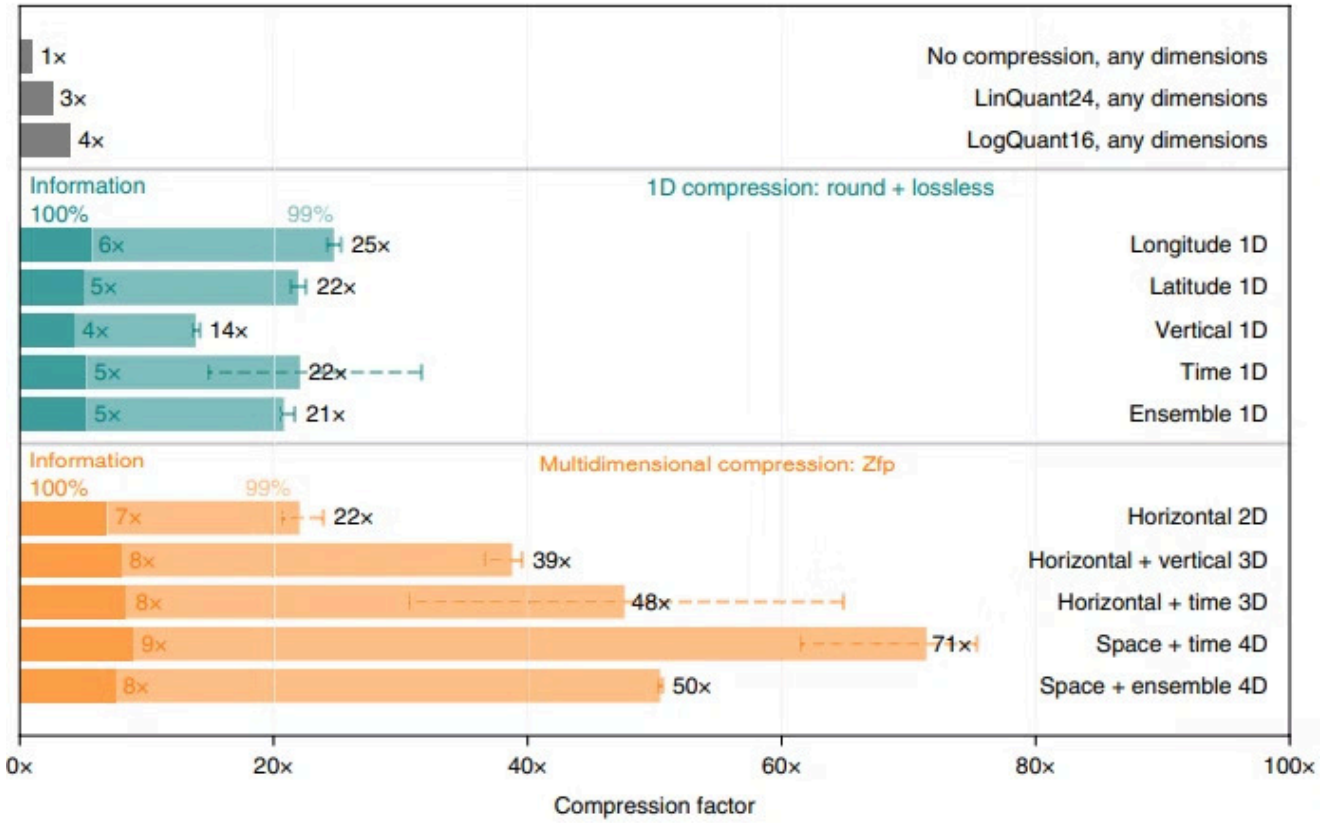
Модель

- 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, они делятся на 2 типа:
  - **lossless** алгоритмы
  - **lossy** алгоритмы



Сжатие с сохранением реальной информации о температуре в различных измерениях

HDF5 - это ZIP в науке

- **HDF5 (Hierarchical Data Format version 5)** — это открытый формат файлов и библиотека, разработанные для хранения и управления **большими объемами сложноструктурированных научных данных** (*astro, quantum, DL, climate, MPT*)
- *По сути, это "контейнер", оптимизированный для работы с многомерными массивами.*

Формат	Плюсы	Минусы	Когда использовать?
HDF5	Поддержка иерархии, сжатие, чанкинг	Сложность для простых данных	Научные данные, многомерные массивы
NetCDF	Удобство при обработке	Менее гибкий, чем HDF5	Метеорология, океанография
ZIP	Универсальность	Нет структуры, медленный доступ	Архивирование файлов

Формат данных NetCDF

- **NetCDF** (Network Common Data Form) — это **формат хранения научных данных** (например, метеорологических).
- Он определяет:
  - **Структуру данных** (переменные, размерности, атрибуты).
  - **Физическое хранение** (как байты записываются на диск).

Версия	Основа	**Поддержка сжатия**	Особенности
NetCDF-4	На основе HDF5	**Да** (сжатие + чанкинг)	Гибкость, но накладные расходы.

Утилита NCO

- **NCO** (NetCDF Operators) — это набор **утилит командной строки** (`ncks`, `ncatted`, `ncpdq` и др.) для обработки NetCDF-файлов.
- Она позволяет:
  - Изменять данные (например, обрезать, объединять, переименовывать переменные).
  - **Сжимать файлы** (используя возможности NetCDF-4).

Алгоритмы сжатия в NCO

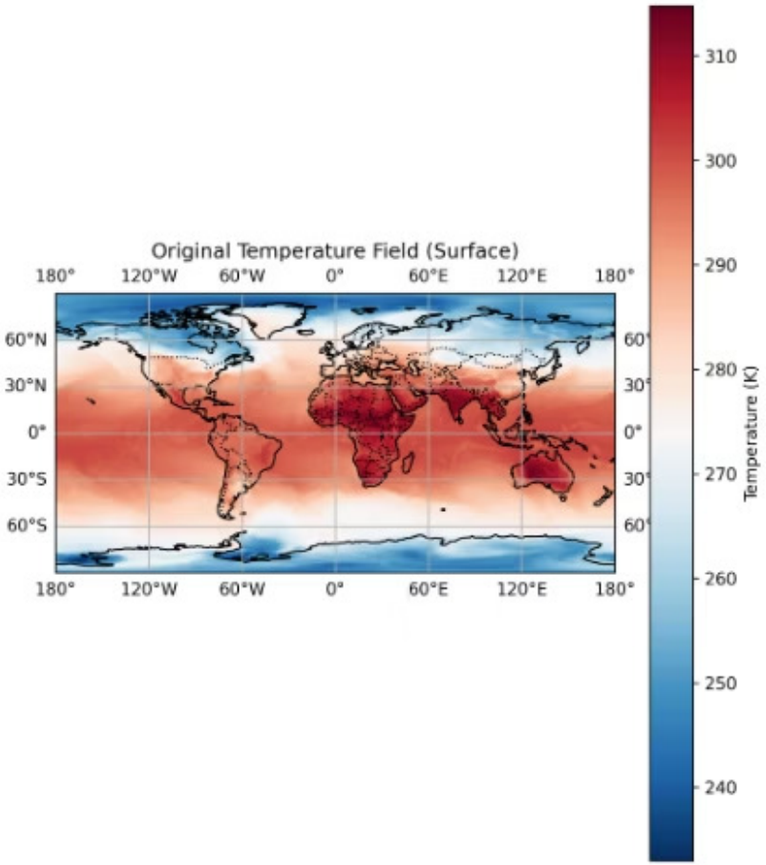
- **NCO** поддерживает два типа сжатия: **[2]**

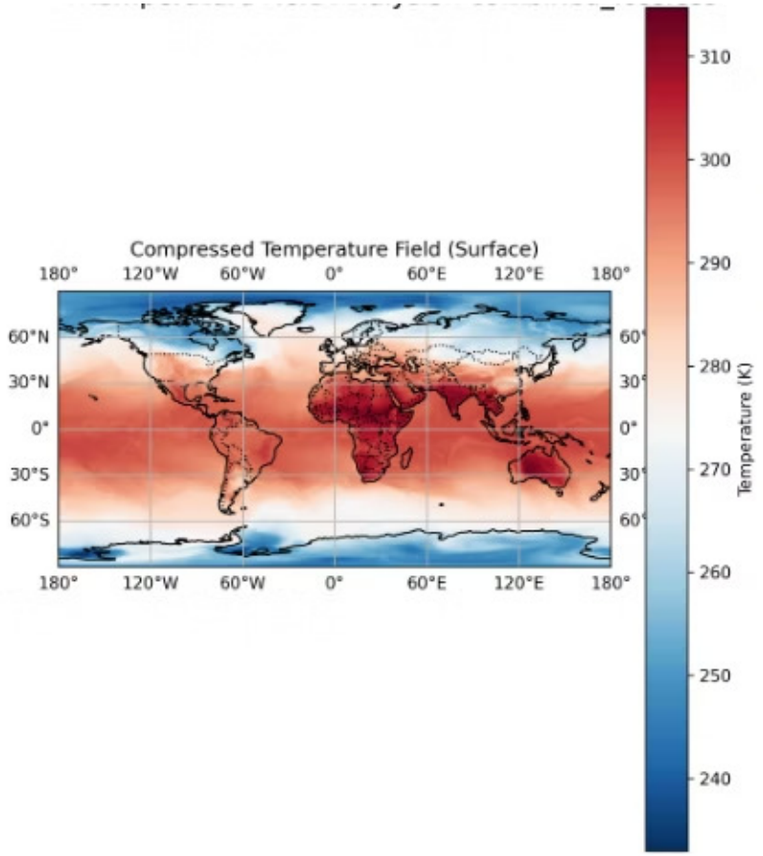
1. **Сжатие с потерями (lossy)** — квантование, **Precision-Preserving Compression, Bit Grooming**
2. **Сжатие без потерь (lossless)** — **DEFLATE + Shuffle | Chunking**... ... ... ⋮

	Тип	**Как работает**	Особенности
DEFLATE (как в ZIP)	Без потерь (lossless)	Ищет повторяющиеся байты и заменяет их кодами	Управление уровнями сжатия от 1 до 9
Shuffle Filter	Без потерь	Переставляет байты, чтобы улучшить сжатие **DEFLATE**.	Это увеличивает повторяемость байт
Precision-Preserving Compression	С потерями	Округляет числа до заданного количества значащих цифр.	Удобен
Квантование	С потерями	Данные преобразуются в целые числа	Удобен

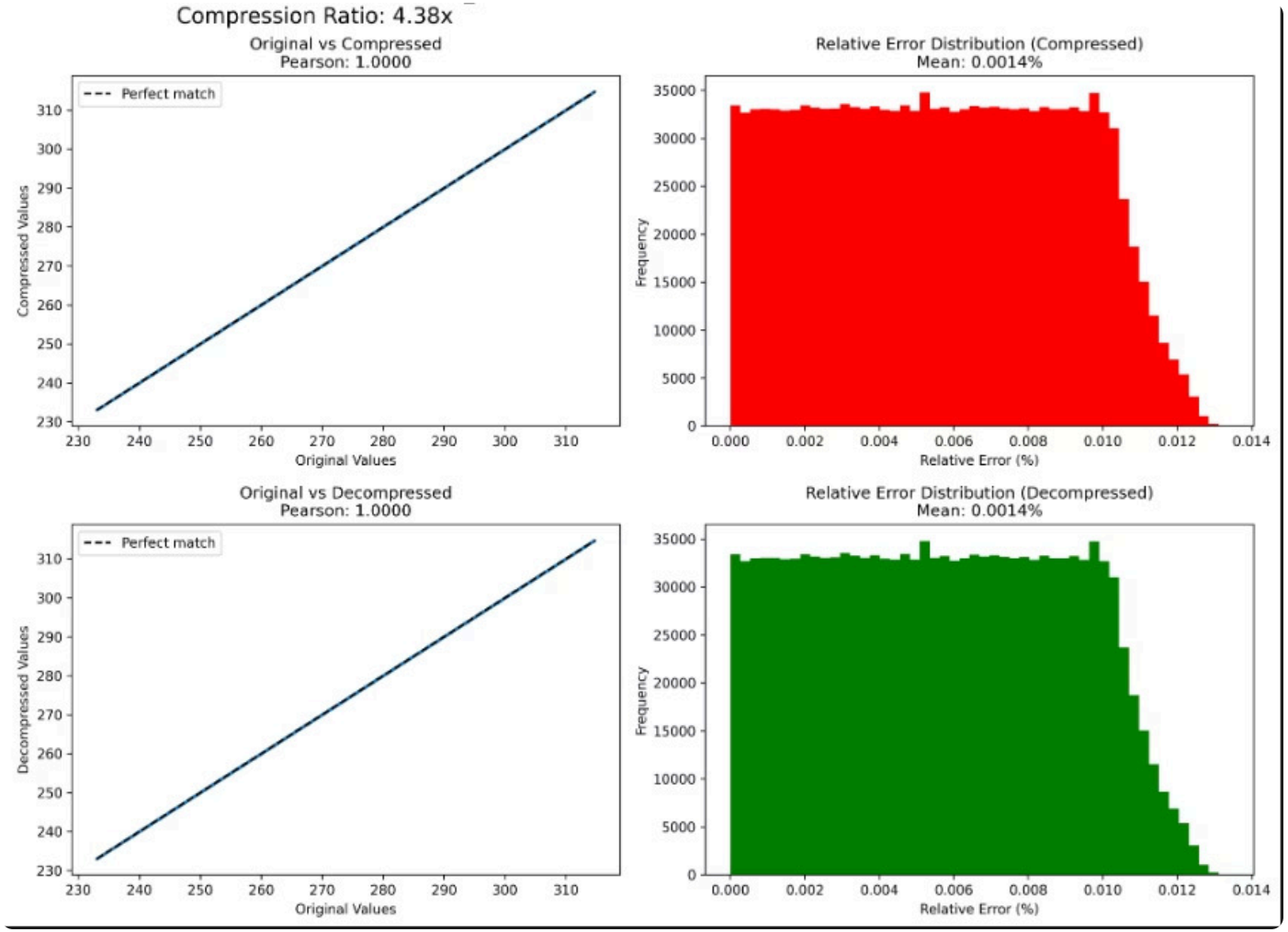
Сравнительный анализ

lossless combined





Сравнительный анализ

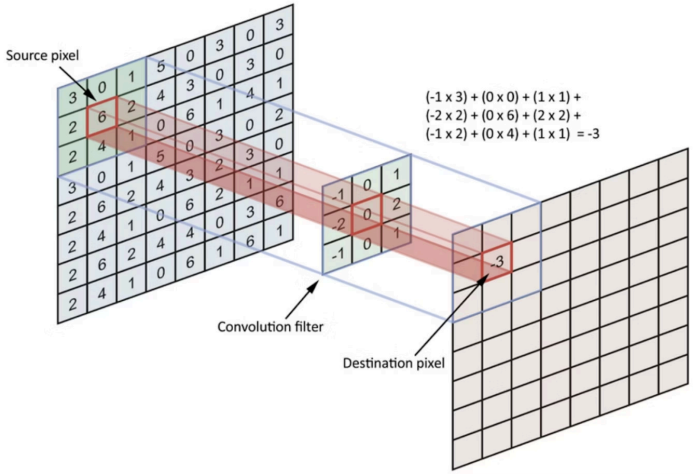


Precision-Preserving Compression (min)

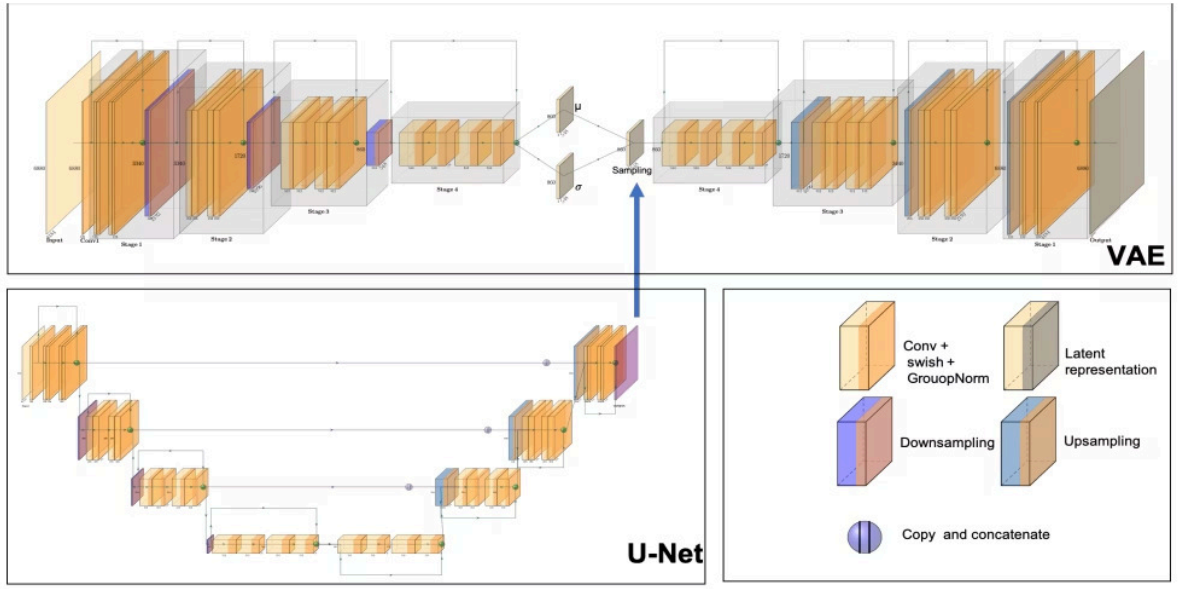
+

Deflate + Shuffle

Нейросетевой подход



Свертка с ядром



Архитектура модели [3]

Итог

- Lossless алгоритмы самые простейшие и быстрые
  - compress ratio = [1-5]
- Lossy из библиотеки NCO уступают другим lossy
  - compress ratio = [1-5]
- Нейросетевые алгоритмы перспективны и требуют изучения

Ссылки

- P.D., Palmer T.N. Compressing atmospheric data into its real information content // Nature Computational Science. 2021. Vol. 1. P. 713-724.
- Rew R., Davis G. NetCDF: An Interface for Scientific Data Access // IEEE Computer Graphics and Applications. 1990.
- Zhao S. et al. Neural Data Compression for Climate Models // 2024.
- Mirowski P., Warde-Farley D., Rosca M., Grimes M.K., Hasson Y., Kim H., Rey M., Osindero S., Ravuri S., Mohamed S. Neural Compression of Atmospheric States // 2024.
- Prims O., Redl R., Rautenhaus M., Selz T., Matsunobu T., Modali K.R., Craig G. The effect of lossy compression of numerical weather prediction data on data analysis: a case study using enstools-compression // Geosci. Model Dev. 2024. Vol. 17. P. 8909-8925.
- Liu Q., Gong B., Zhuang X., Zhong X., Kang Z., Li H. Compressing high-resolution data through latent representation encoding for downscaling large-scale AI weather forecast model

Спасибо за внимание!

Кузнецов Иван

kuznetsov.ia@phystech.edu

tg: @stop_pleaze