

Verlustfreie Kompression von Klimadaten mit Machine Learning

Masterarbeit

Einführung

Problem

Die Klimawissenschaften generieren sehr hohe Datenmengen (zur Zeit ca. 770 TiB)

Aktuelle Lösung

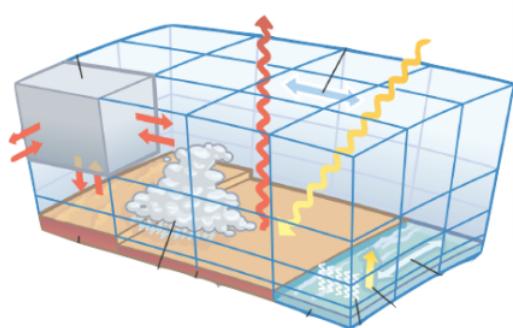
Reduzierung der zeitlichen Auflösung und gespeicherten Variablen

Folgen

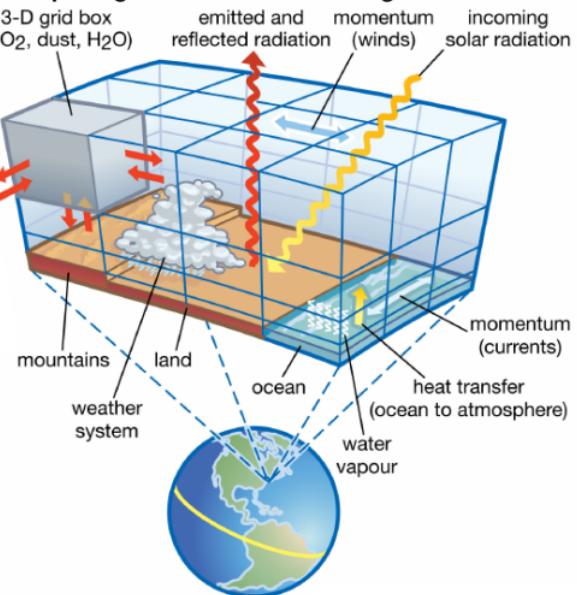
- Klimaereignisse möglicherweise nicht abgebildet
- Benutzung von Interpolationen
- Neuberechnung von Simulationen

Klimadaten

- 4D Daten (Längen- u. Breitengrad, Höhe, Zeit)

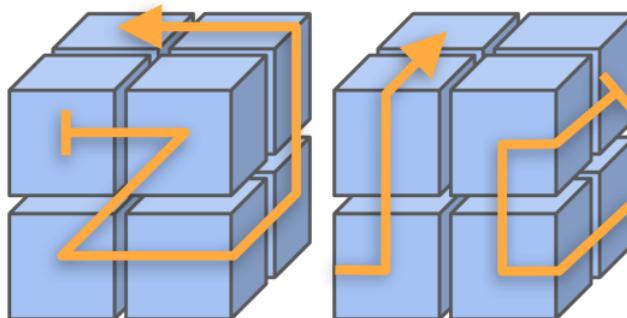


Concept diagram of climate modeling



Kompression

- Vorhersagebasierte Kompression ist am erfolgreichsten für Klimadaten (bei verlustfreier Kompression)
- Ablauf (vereinfacht)
 - Berechnung einer Vorhersage für jeden einzelnen Datenpunkt
 - Berechnung des Residuums
 - Kodierung des Residuums, sowie der Traversierungs- und Vorhersageverfahren



Wissenschaftliche Fragestellungen

- Welche Abhangigkeit besteht zwischen Traversierung und Vorhersage?
- Wie kann die Traversierung auf die Daten bzw. ihre Vorhersagbarkeit abgestimmt werden?

Wissenschaftliche Fragestellungen

- Welche Abhangigkeit besteht zwischen Traversierung und Vorhersage?
- Wie kann die Traversierung auf die Daten bzw. ihre Vorhersagbarkeit abgestimmt werden?
- Welche Feature konnen fur das Lernen der Vorhersage verwendet werden?
 - Breite- u. Langengrad, Hohe
 - Lokale Uhrzeit und Jahreszeit
 - Land/Wasser
 - Variablenabhangigkeit (Temperatur, Wasserdampf, Luftfeuchte)
- Wie konnen diese Feature kodiert werden?

Wissenschaftliche Fragestellungen

- Welche Abhangigkeit besteht zwischen Traversierung und Vorhersage?
- Wie kann die Traversierung auf die Daten bzw. ihre Vorhersagbarkeit abgestimmt werden?
- Welche Feature konnen fur das Lernen der Vorhersage verwendet werden?
 - Breite- u. Langengrad, Hohe
 - Lokale Uhrzeit und Jahreszeit
 - Land/Wasser
 - Variablenabhangigkeit (Temperatur, Wasserdampf, Luftfeuchte)
- Wie konnen diese Feature kodiert werden?
- Wie kann die Kodierung performant gestaltet werden?
 - Parallelisierung durch Datenblocke?
 - Entropy-basierte Verfahren?

Aufgaben

- Einarbeitung in die Datenformate netCDF und HDF5.
- Evaluation von ML-Verfahren für die Vorhersage von Datenpunkten (z.B. supervised, unsupervised, reinforcement learning).
- Engineering der Codierungspipeline bzgl. Performance und Kompression.

Stand der Technik

- Lossless Image Compression through Super-Resolution [2020, #image]
- High-Fidelity Generative Image Compression [2020, #image]
- Estimating Lossy Compressibility of Scientific Data Using Deep Neural Networks [2020, #float]
- Adaptive Deep Learning based Time-Varying Volume Compression [2020, #float]
- Wavefield compression for seismic imaging via convolutional neural networks [2019, #float]
- Lossless Data Compression with Neural Networks [2019, #text]
- DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Statistics of Natural Videos [2019, #video]