Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

21. Juni 2022

Compressing atmospheric data into its real information content.

Komprimierung atmosphärischer Daten auf ihren echten Informationsgehalt.

Universität Hamburg
Scientific Computing // wissenschaftliches Rechnen
Seminar Supercomputer: Forschung und Innovation

Aleksandar Petrusov
Betreuerin: Anna Fuchs

Gliederung

- Einführung
  - Atmosphärische Daten
  - Besonderheiten atmosphärischer Daten
- Bitweiser Informationsgehalt
- Bitweiser echter Informationsgehalt
  - Funktionsweise
  - Vorteile
- Schwächen derzeitiger Kompression
- Multidimensionale Kompression
- Geschwindigkeiten und Bewertungsansätze der Kompressoren
- Zusammenfassung

1 Einführung

- Was sind atmosphärische Daten?
- Was sind die Besonderheiten solcher Daten?
- Welche Vorteile bringt es diese zu komprimieren?
- Was sind die schwächen derzeitiger Kompressoren?
- Welche Ansätze existieren?

Atmosphärische Daten

- gasförmige Hülle um die Erde
- Wind und Wetter

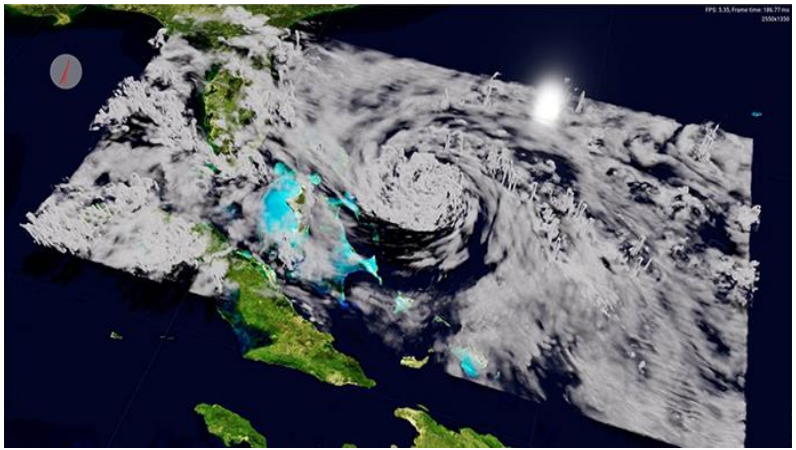


Abb.1: 1km Simulation des Kategorie 5 Hurrikans
Quelle: Literaturverzeichnis 1)

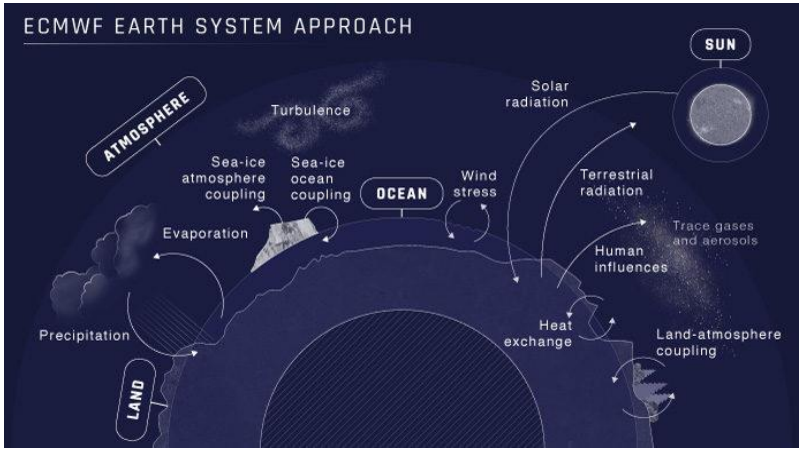


Abb.2: ECMWF-Ansatz für das globale klima System
Quelle: Literaturverzeichnis 2)

- Wolken
- Regen
- Temperatur

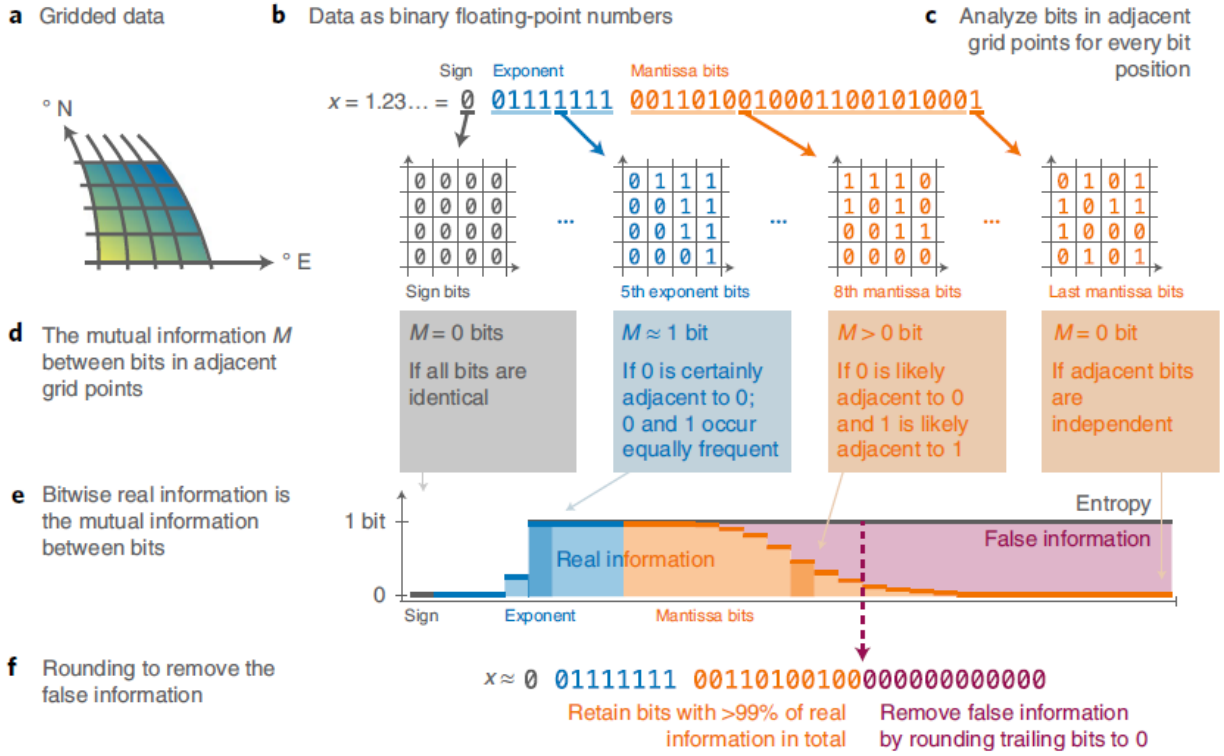
Besonderheiten atmosphärischer Daten

- Hunderte Petabytes an Daten
- 64-Bit-Gleitkommazahlen (IEEE-Standard 754)
- in 6 Dimensionen
  - 3 im Raum
  - Zeit
  - Vorhersagezeit
  - Ensemble-Dimension
- chaotisches System

2 Bitweiser Informationsgehalt

- Beruht auf Shannonsche Informationstheorie
- Zur Vorhersagbarkeit von dynamischen Systemen formuliert
- Quantifiziert wie viele Bits zum Informationsgehalt beitragen
- Erweiterung => bitweiser **echter** Informationsgehalt
  - Unterscheidet zwischen falschen und echten Informationen in den Bits

Abb.3: Gitterpunktdarstellung des Bitweise echten Informationsgehaltes

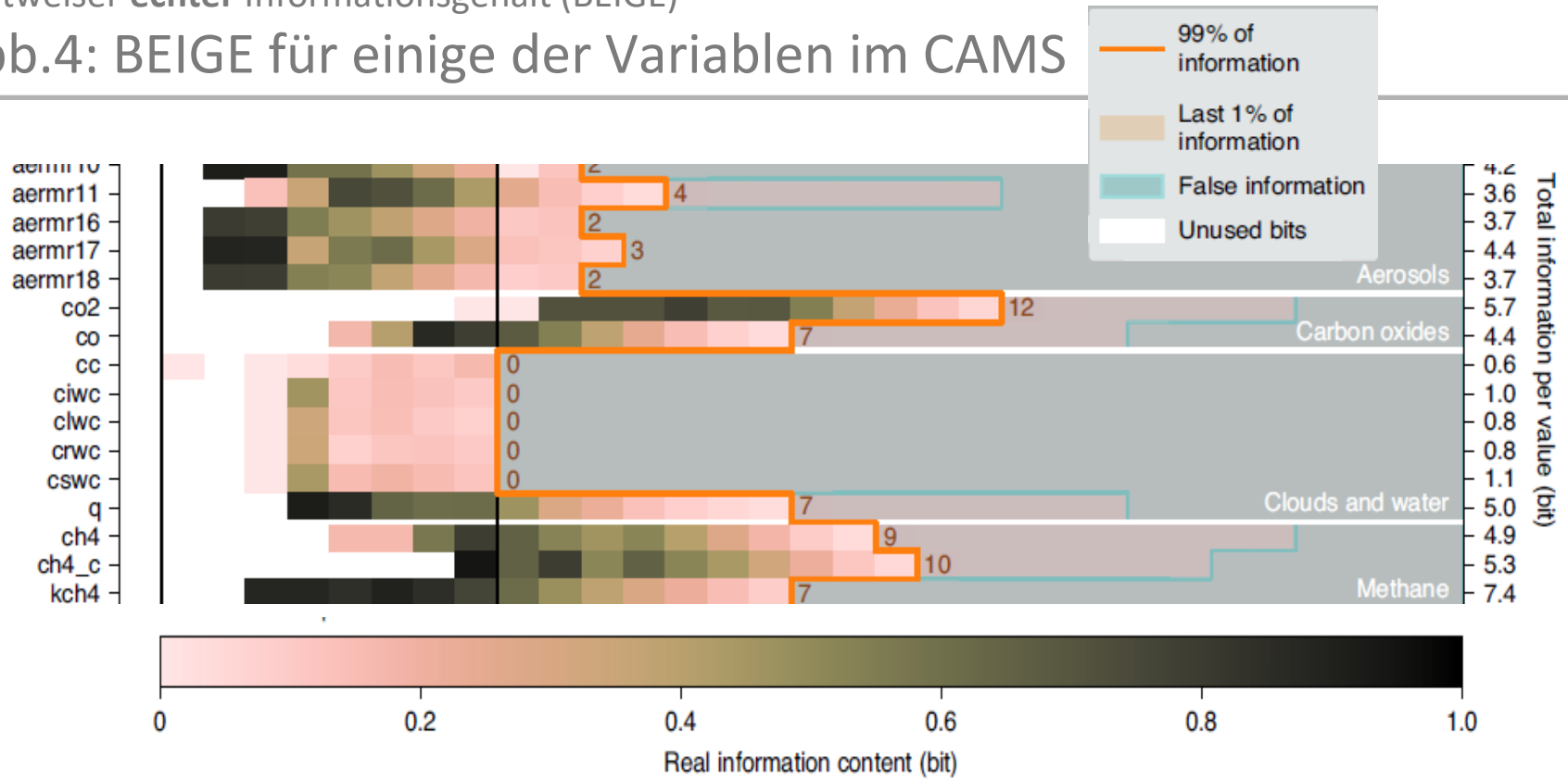


Quelle: Literaturverzeichnis 3)

3 Bitweiser **echter** Informationsgehalt (BEIGE)

- Information befinden sich in den Exponenten Bits
- Hintere Mantissen Bits unabhängig und mit ähnlicher Wahrscheinlichkeit
  - Hohe Informationsentropie
  - Enthalten nur geringe Mengen an Informationen
  - Abhängigkeit unterscheidet sich unwesentlich von Null
- mehr Informationen je stärker statistische Abhängigkeit
- Benachbarte Bits in jeder Dimension zu finden
  - => räumliche und zeitliche Korrelation

Abb.4: BEIGE für einige der Variablen im CAMS



Quelle: Literaturverzeichnis 3)

Funktionsweise BEIGE

- nutzt Information in benachbarten Gitterpunkten
- nutzt räumliche und zeitliche Kohärenz
- Analyse der Bits mit echtem Informationsgehalt (Keep-Bits)
- rundet Bits ohne echten Informationsgehalt auf Null

- => Komprimierung nur der echten Informationen

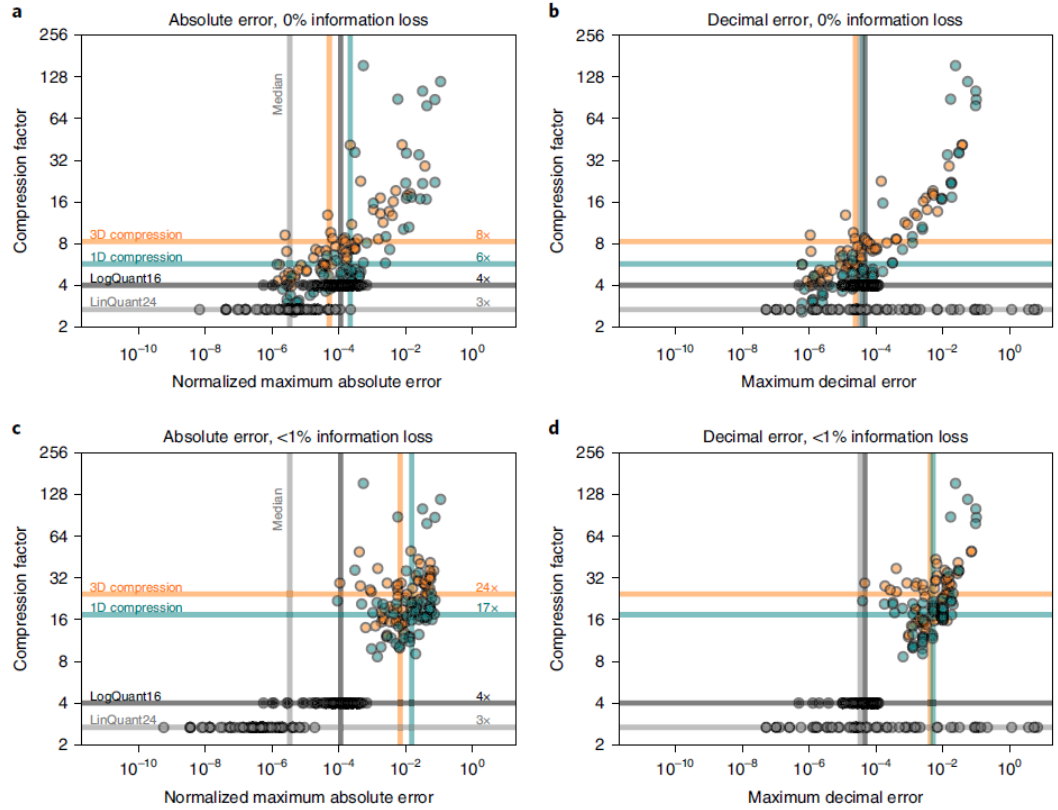
Vorteile der Komprimierung mit BEIGE

- Komprimierung immer ein Kompromiss
  - Größe, Genauigkeit und Geschwindigkeit
- Speicherbedarf wird stark verringert
- gemeinsame Daten Nutzung erleichtert
- Analyse sowie Rundung sind deterministisch
- 1% Informationsverlust erhöht Komprimierungsfaktor immens

Schwächen der derzeitigen Kompressoren (CAMS)

- nutzt lineare Quantisierung
  - 24-Bit-Version des GRIB237
- Korrelation wird sich selten zu nutze gemacht
  - Da nur 1 Dimension komprimiert wird
- schlechte Übereinstimmung Daten und quanti. Werte
  - meisten Bits ungenutzt

Abb. 5: Kompressionsfaktor im Vergleich zum Kompressionsfehler

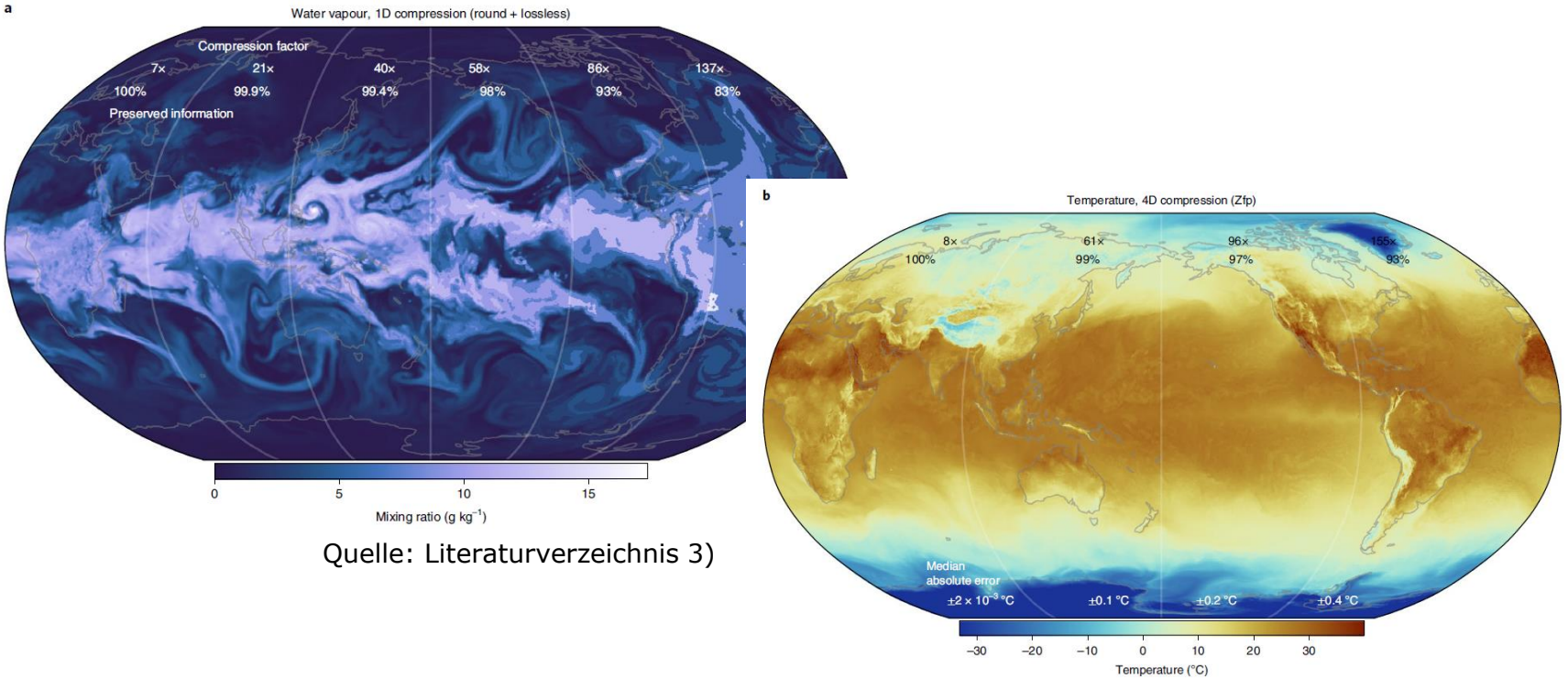


Quelle: Literaturverzeichnis 3)

Multidimensionale Datenkompression

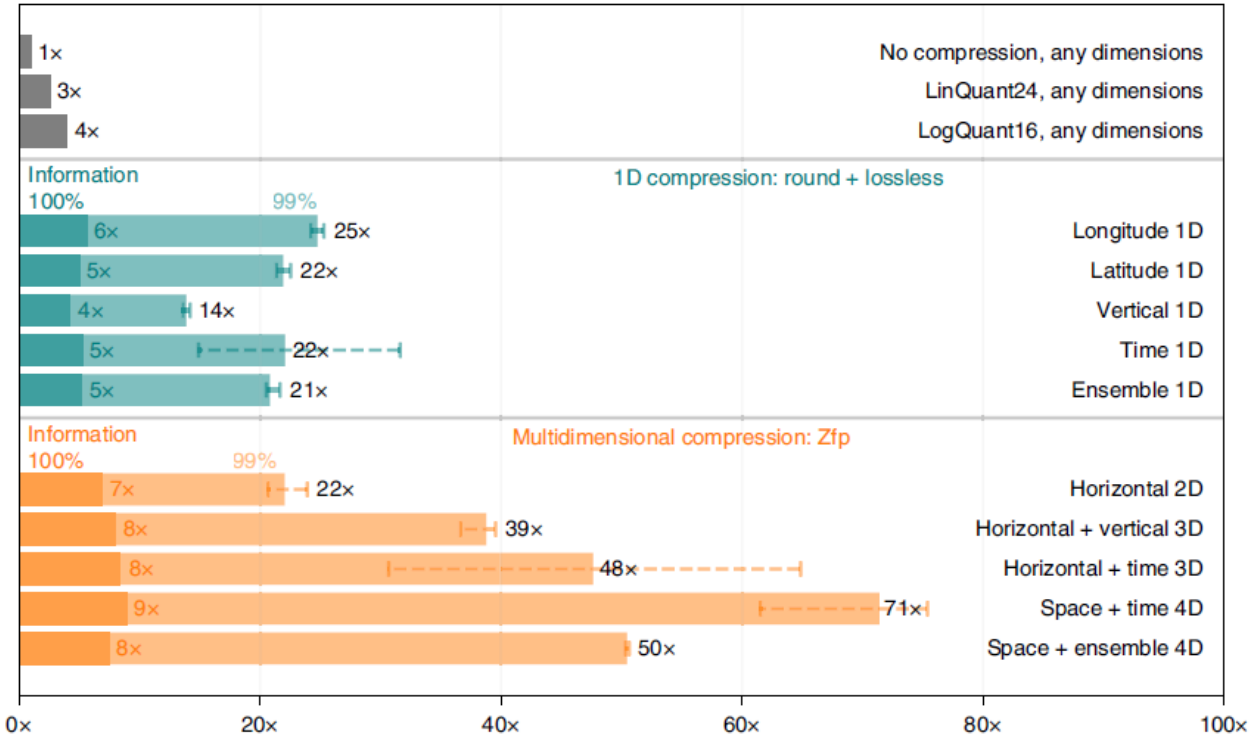
- Komprimierung mehrerer Dimension
  - Zfp für 2-4 Dimension geeignet
- Unterteilt ein d-dimensionales Feld in 4D-Wert-Blöcken
  - führt automatisch zu Korrelation
- 4D-Kompression am effektivsten
  - Mittlerer Absolute Fehler bei 0,1°C bei 99% Info

Abb. 6: Komprimierung auf verschiedenen Ebenen der erhaltenen Informationen



Quelle: Literaturverzeichnis 3)

Abb. 7: Komprimierung der realen Information der Temperatur in verschiedenen Dimensionen



Quelle: Literaturverzeichnis 3)

Geschwindigkeit der Kompression und Dekompression

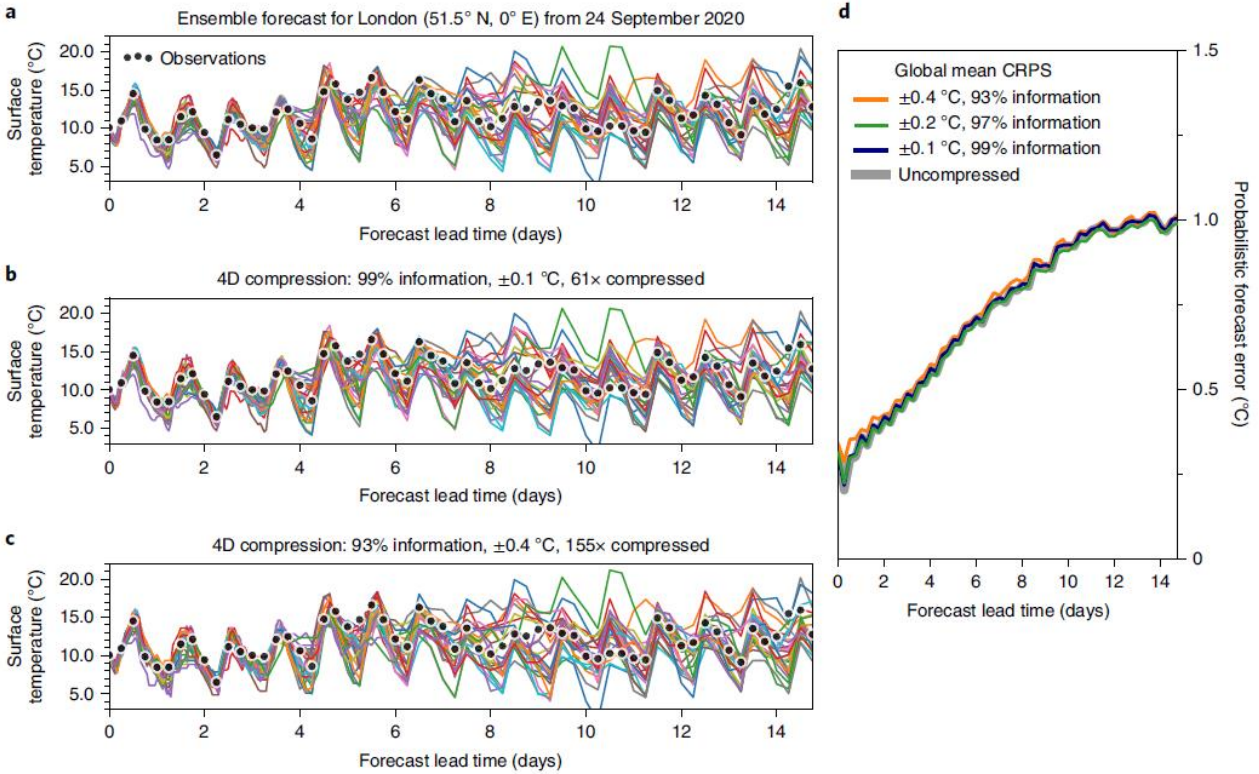
- angemessene Geschwindigkeiten
- niedrige Kompressionsfaktoren
  - => höhere Geschwindigkeiten
- essentiell für gemeinsame Datennutzung

Bewertungsansätze

- Komprimierung an Scores bewerten
  - Score Cards
  - Müssen bestanden werden wie bei Turing-Test
- Tests von zusätzlichem Wert
  - nicht bestehen = Probleme
  - Bestehen = Potenzial

- CRPS (Continuous ranked probability score)
  - quadratischer Fehler zwischen Beobachtung und Vorhersage

Abb. 8: Überprüfung einer Ensemble-Vorhersage



Quelle: Literaturverzeichnis 3)

Zusammenfassung

- Was sind atmosphärische Daten?
  - Wind und Wetter
- Was sind die Besonderheiten solcher Daten?
  - chaotische und sehr große Datenmengen
- Welche Vorteile bringt es diese zu komprimieren?
  - Speicherbedarf verringern und gemeinsame Datennutzung erleichtern
- Was sind die schwächen derzeitiger Kompressoren?
  - keine Unterscheidung zwischen echten und falschen Daten
- Welche Ansätze existieren?
  - BEIGE und moderne Algorithmen

Literatur

- 1) ECMWF, 2022. *1km Simulation des Kategorie 5 Hurrikans Dorian*. [image] Available at: <https://www.ecmwf.int/sites/default/files/medialibrary/2022-04/1km-simulation-dorian.jpg> [Accessed 19 June 2022].
- 2) ECMWF, 2022. *Erdsystem Ansatz*. [image] Available at: <https://www.ecmwf.int/sites/default/files/styles/news_item_main_image/public/our-earth-system-690px.jpg?itok=7hQExTD3> [Accessed 19 June 2022].
- 3) Klöwer, M., Razinger, M., Dominguez, J.J. *et al.* Compressing atmospheric data into its real information content. *Nat Comput Sci* **1,** 713–724 (2021).https://doi.org/10.1038/s43588-021-00156-2