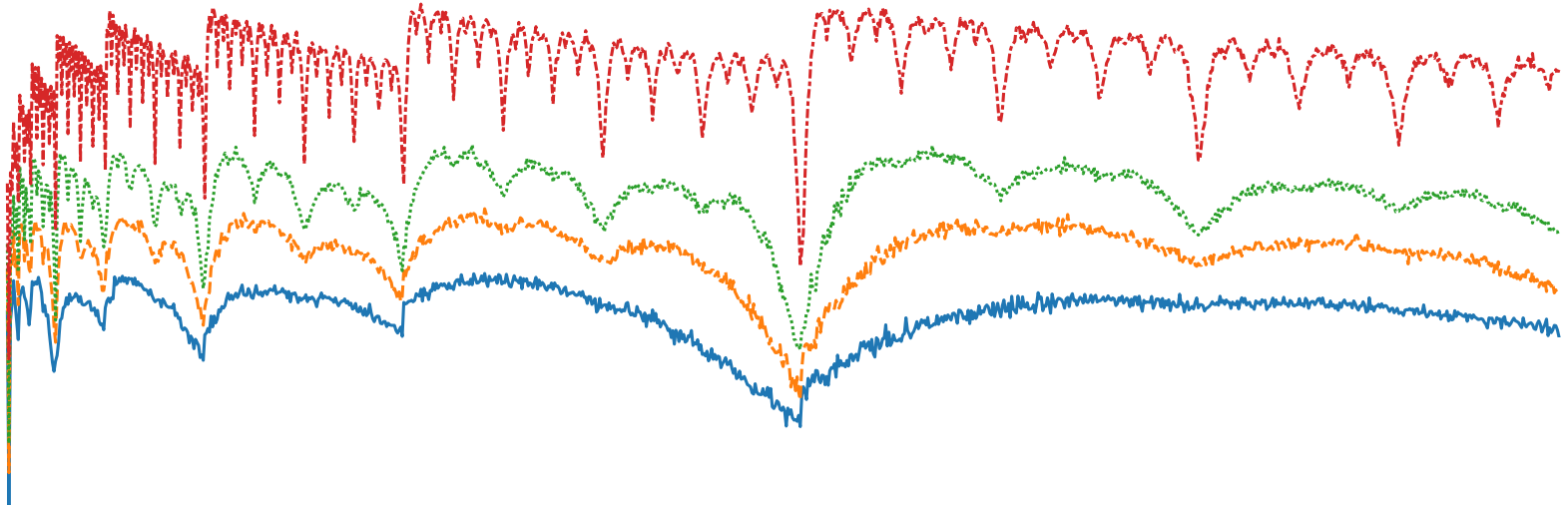
Karlsruher Institut für Technologie

Kompressionsmethoden für strukturierte Gleitkommazahlen und ihre Anwendung in den Klimawissenschaften

von Uğur Çayoğlu

STEINBUCH CENTRE FOR COMPUTING (SCC) und
INSTITUT FÜR METEOROLOGIE UND KLIMAFORSCHUNG (IMK-ASF)



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem Hohes Datenaufkommen durch Klimasimulationen

ERA5 Datensatz für die Initialisierung und Validierung von Simulationsläufen umfasst 10.89 PiB

IMK-ASF Einer der größten Speicherplatzbenutzer am SCC mit >770 TiB (steigend)

Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem

Hohes Datenaufkommen durch Klimasimulationen (ERA5, 10.89 PiB)

Aktuelle Lösung

Reduzierung der zeitlichen Auflösung und gespeicherten Variablen

Folgen

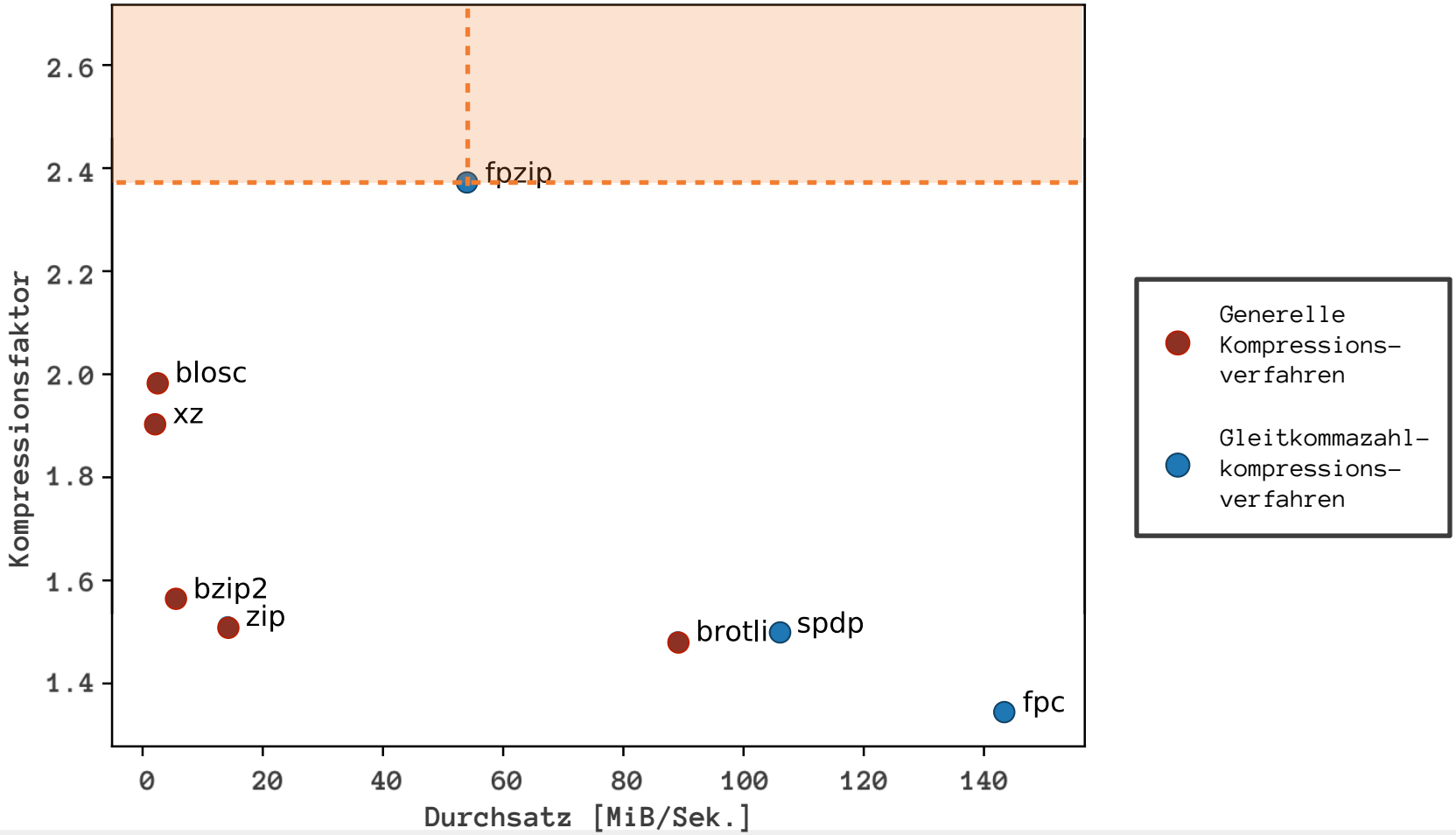
- Benutzung von Interpolationen
- Klimaereignisse (z.B. Entstehung von Stürmen) möglicherweise nicht abgebildet
- Neuberechnung von Simulationen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor

Kompressionsfaktor und Durchsatz

$$\text{Kompressionsfaktor} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Komprimierte Dateigröße [Bytes]}} \qquad \text{Durchsatz} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Kompressionszeit [Sek.]}}$$

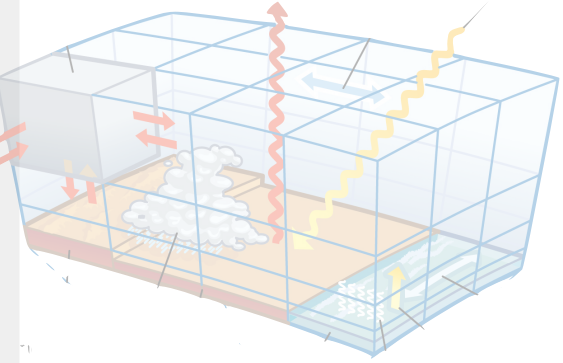


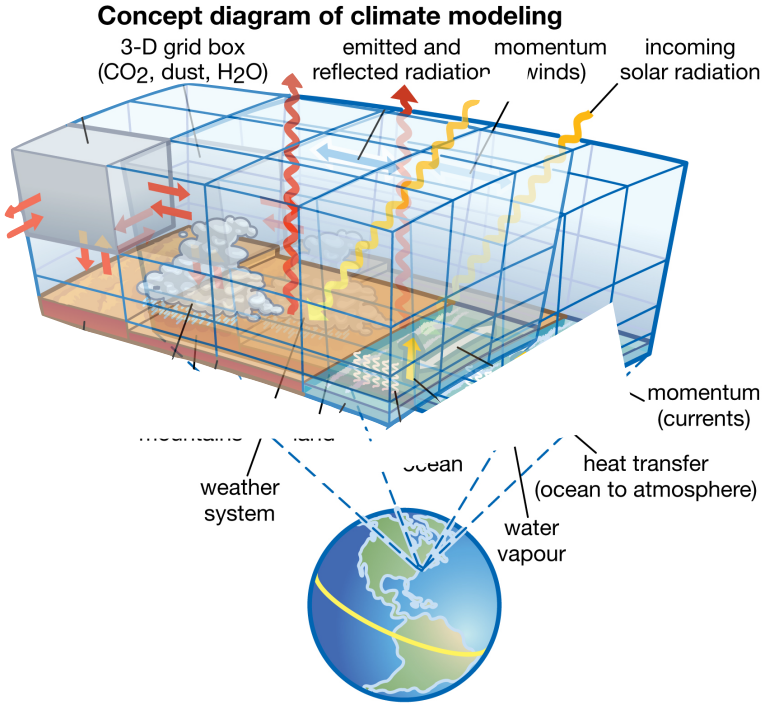
Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Klimadaten

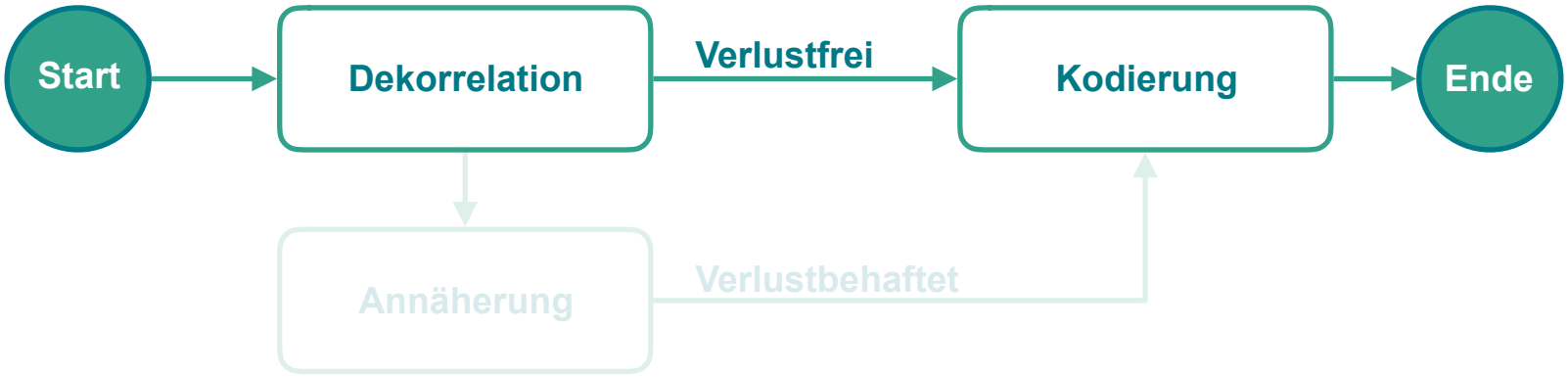
4D Daten (Längen- u. Breitengrad, Höhe, Zeit)





Quelle[1]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



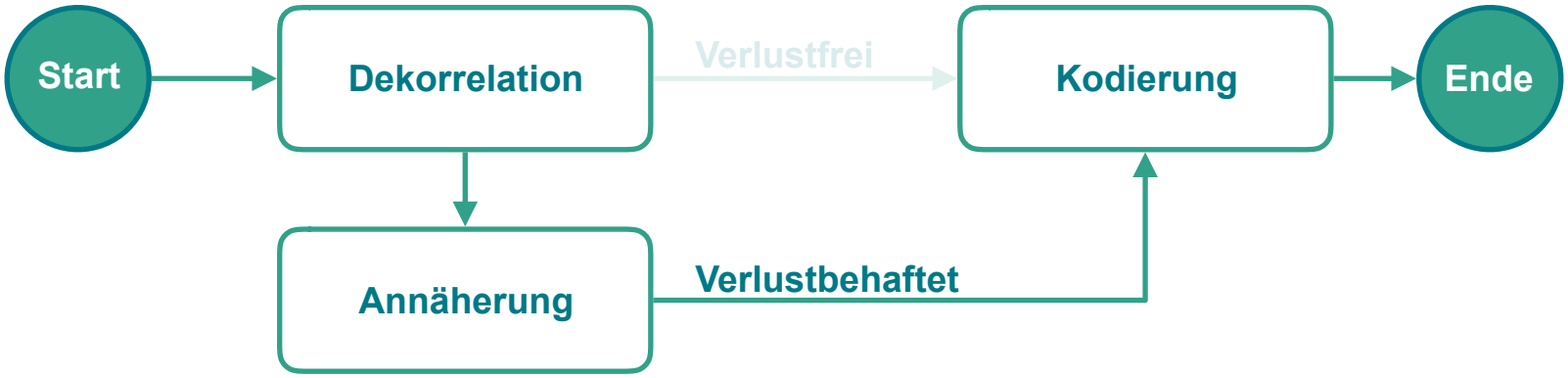
Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

JPEG

Kompressionsarten

- **Verlustbehaftete**
- Verlustfreie

Datentypen

- **Integer**
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- **Bild**
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- **Lauflängenkodierung**
- Wörterbuch-Verfahren
- **Transformation**
- Vorhersagebasierte Verfahren
- **Entropie-basierte Verfahren**

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Beiträge zur Informatik

Karlsruher Institut für Technologie

pzip

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

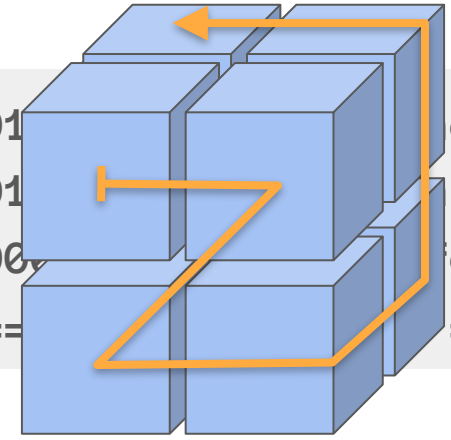
Direkter Beitrag | Einflussbereich

Pascal Zip (pzip)

Vorhersagebasiertes Kompressionsverfahren

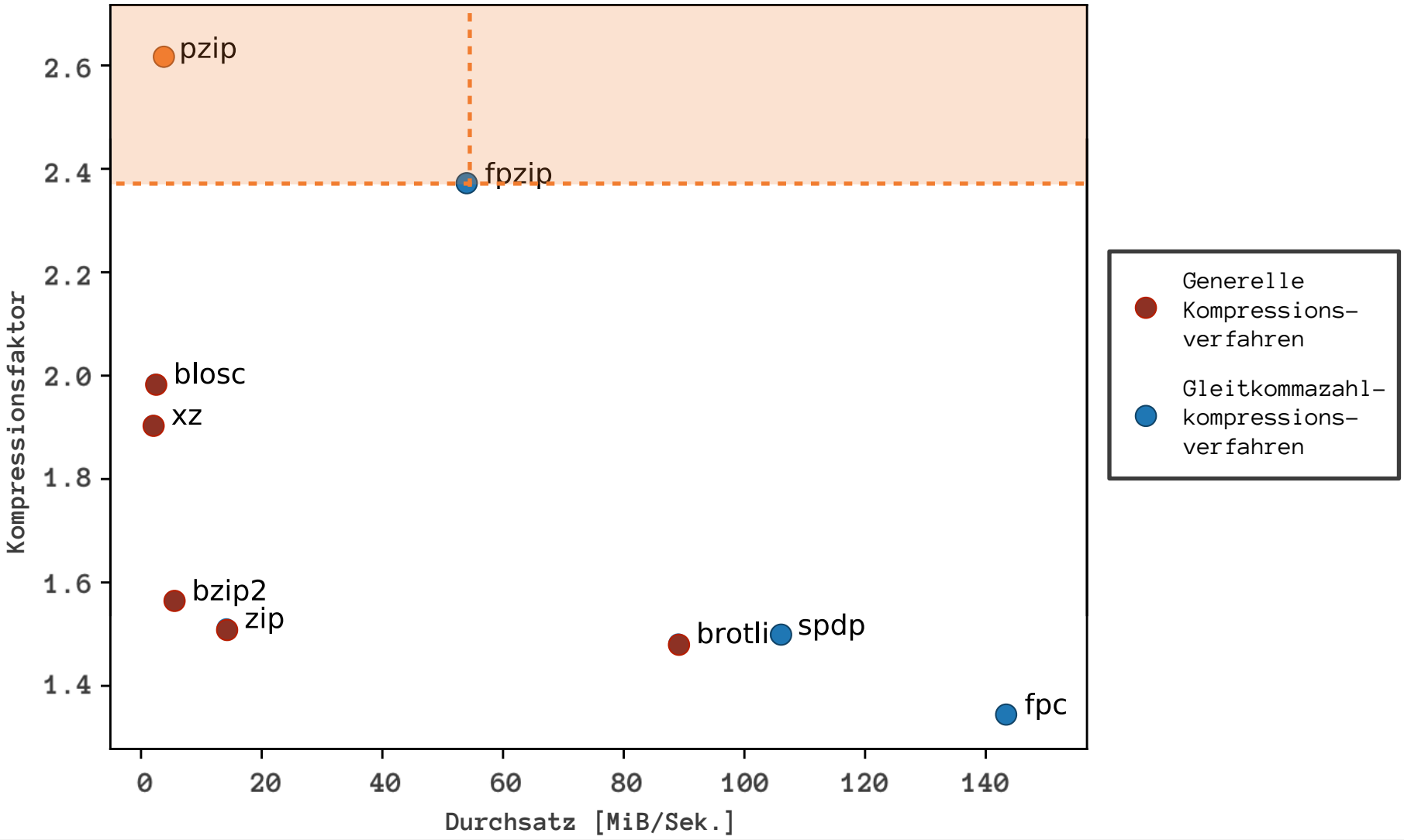
Methode

Für jeden einzelnen Datenpunkt wird (basierend auf vorhergehenden Werten) eine Vorhersage gegeben und die Differenz zum wahren Wert (Residuum) gespeichert

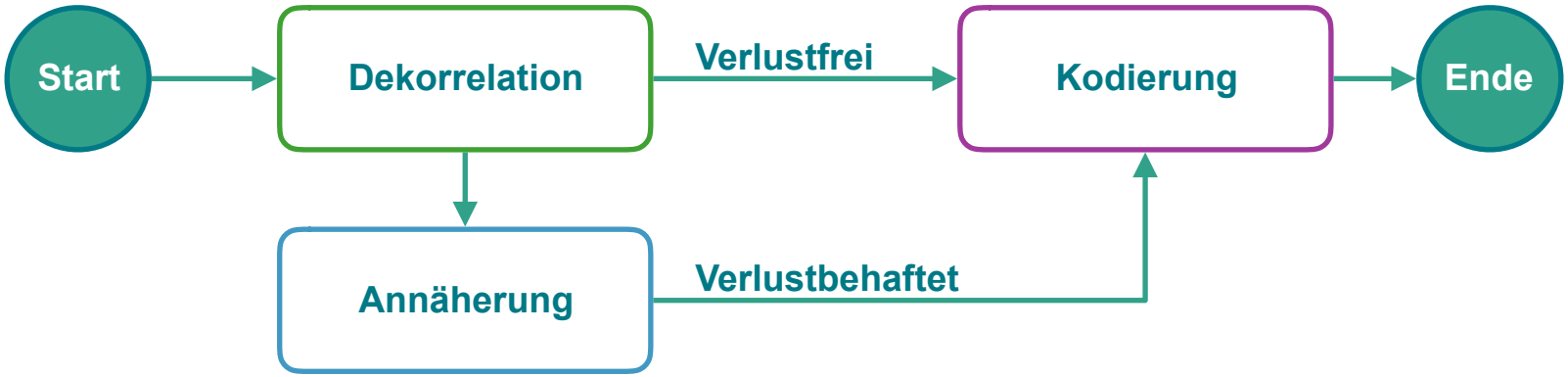


```
01000010100001111011010001                ersage = 67.853
01000010100100101000000001                hrheit = 73.251
00000000000101010011010000                erenz
          *============                    = 11 -> 32 - 11 - 1 = 20
```

Kompressionsverfahren im Vergleich



Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

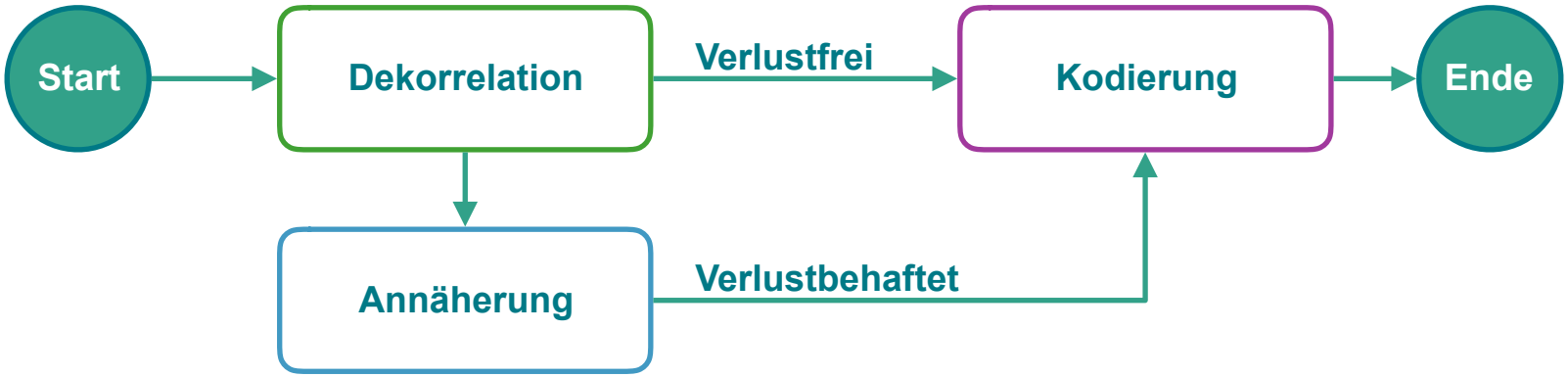
Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

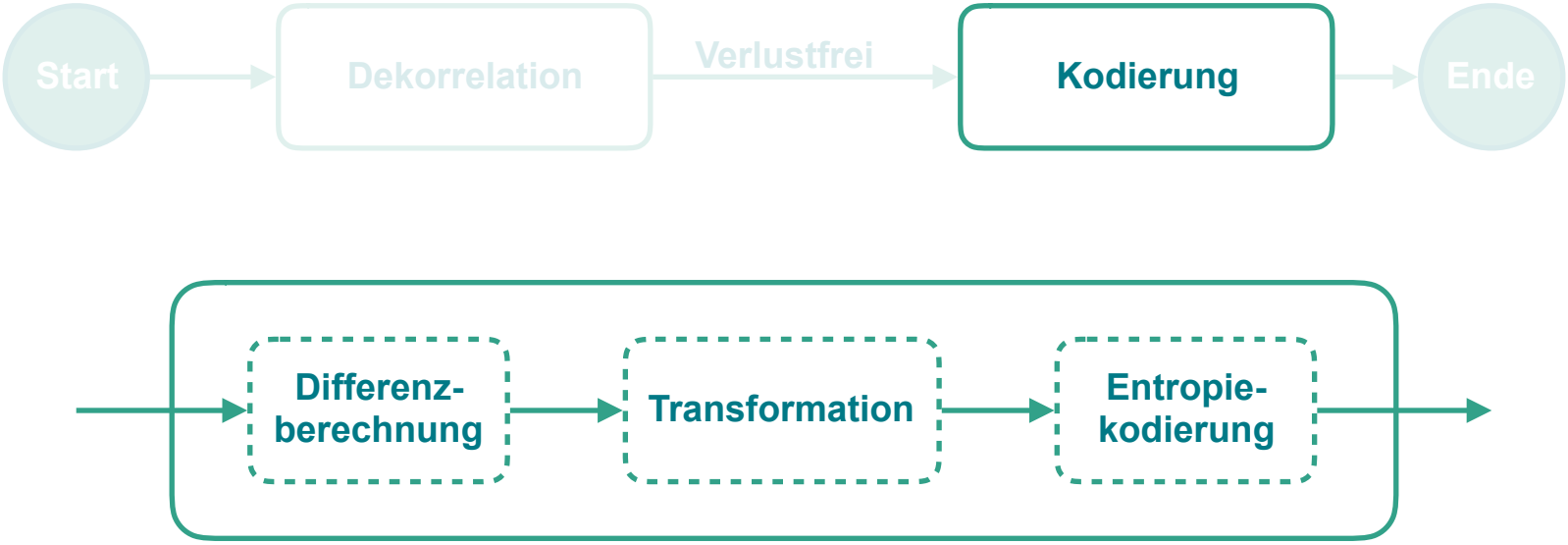
Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

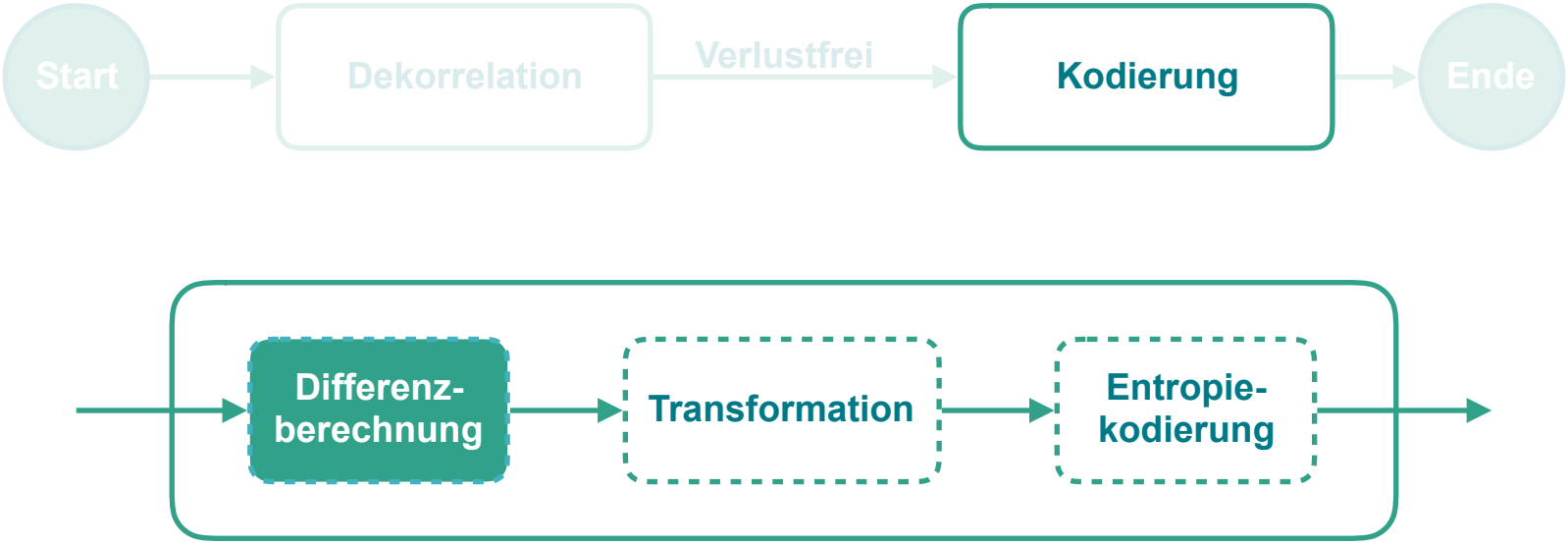
Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Datenkodierung bei der verlustfreien vorhersagebasierten Kompression

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Zwei Arten der Berechnung von Residuen

Karlsruher Institut für Technologie

Abs. Differenz

$$d = |v - w|$$

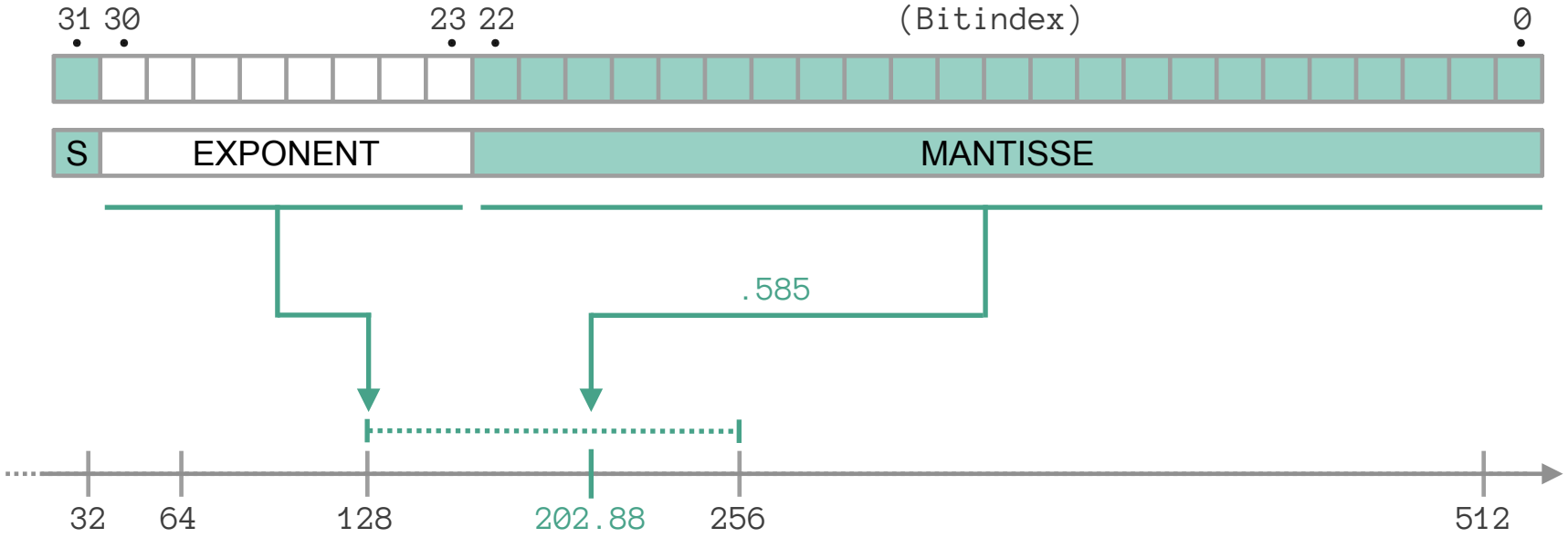
- ⊕ Kleine Residuen
- ⊖ Underflow
- ⊖ Zwei Operationen
- ⊖ Bit für Vorzeichen

XOR

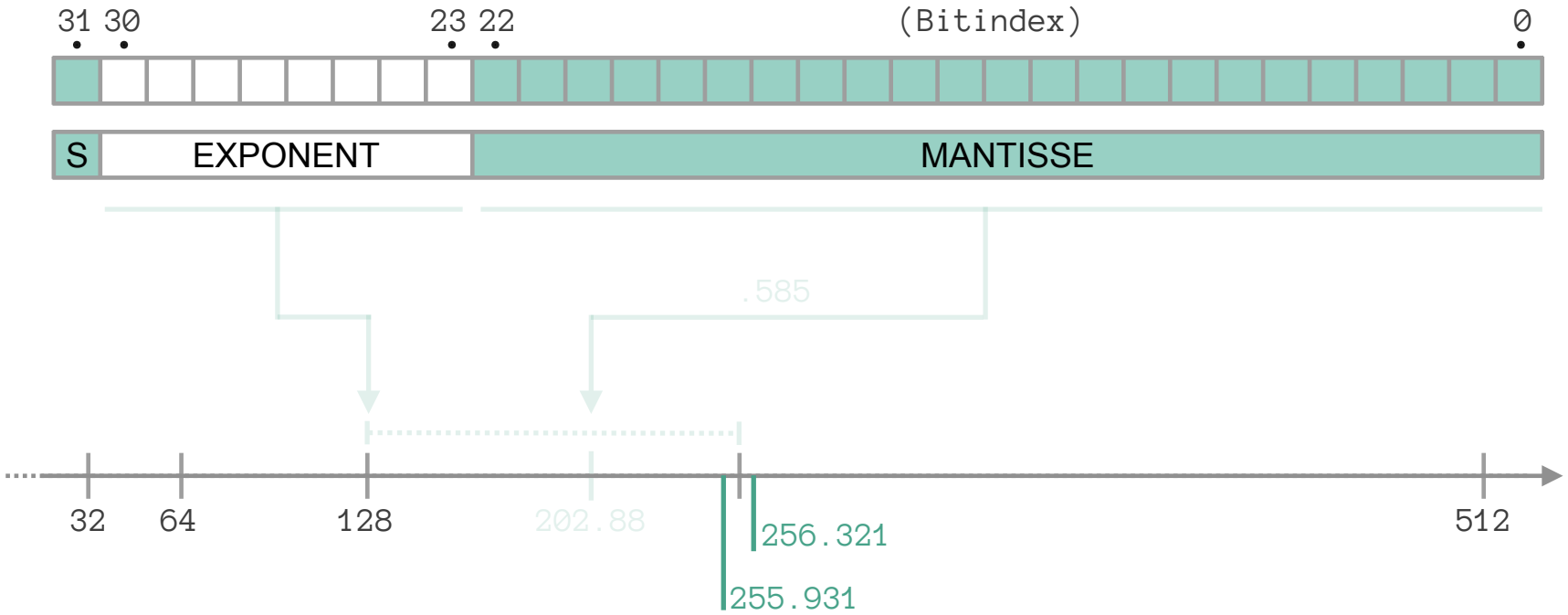
$$d = v \oplus w$$

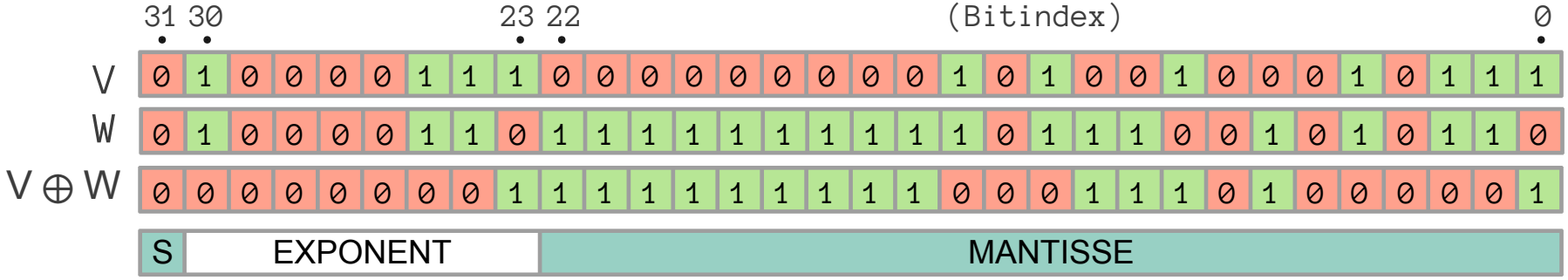
- ⊕ Eine Operation
- ⊕ Kein Underflow
- ⊖ Bitflip-Problem (große Residuen)

Gleitkommazahlen und der Bitflip

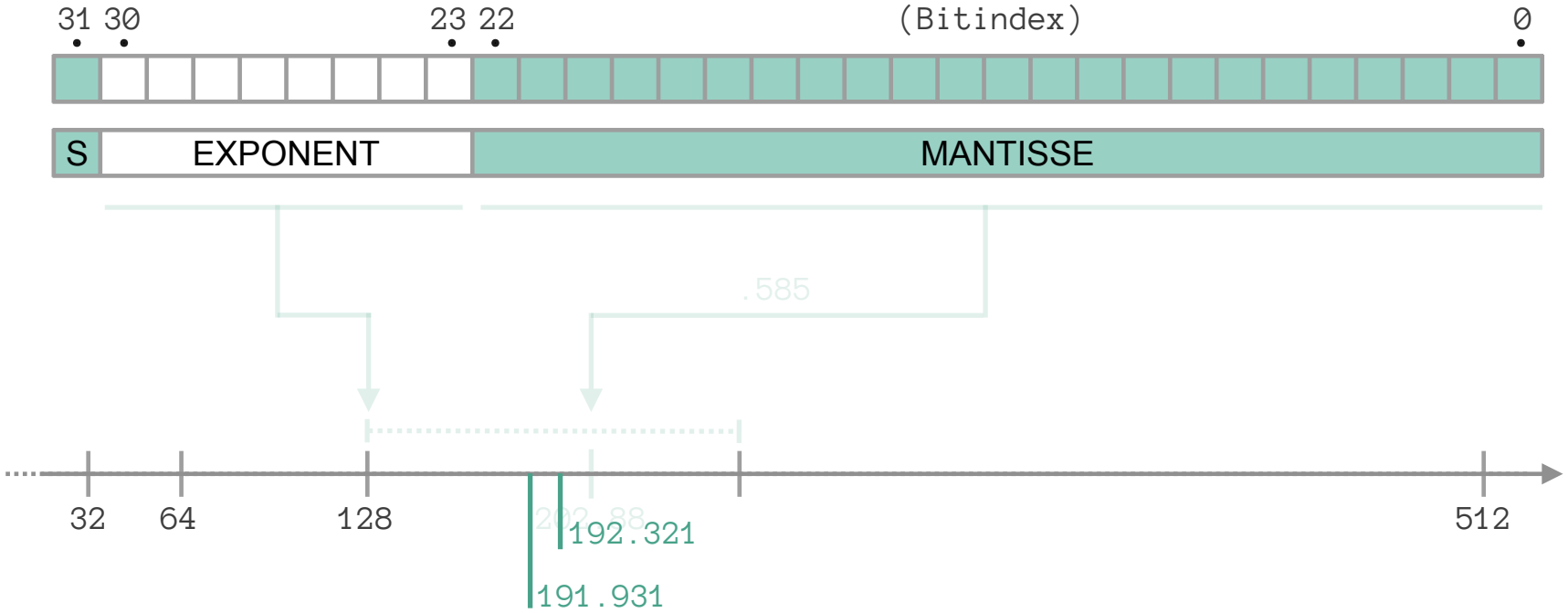


Gleitkommazahlen und der Bitflip

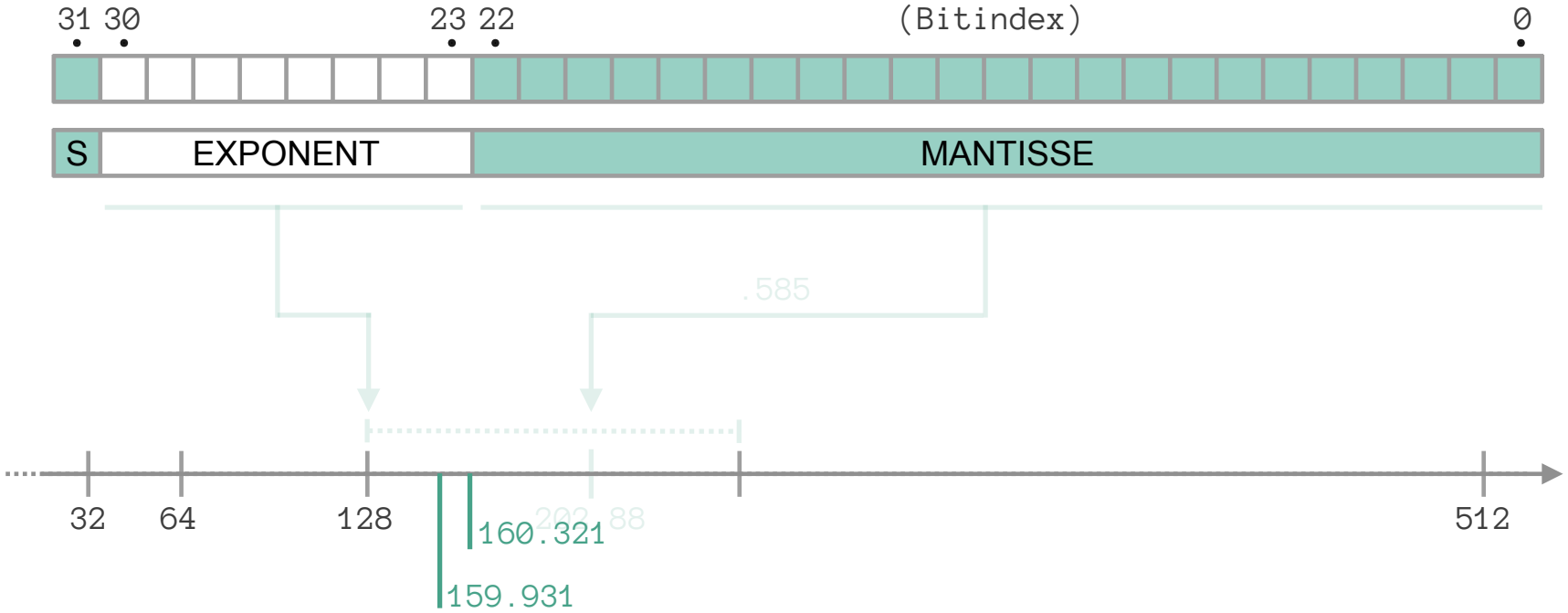




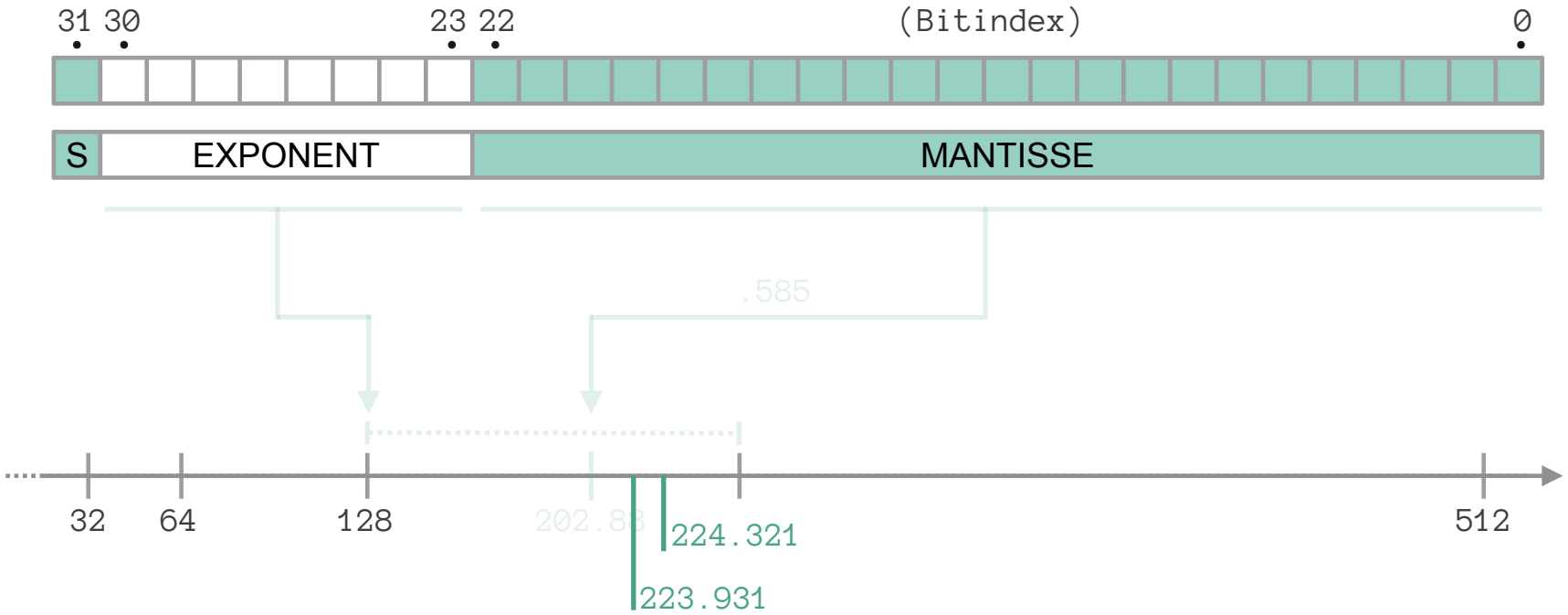
Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

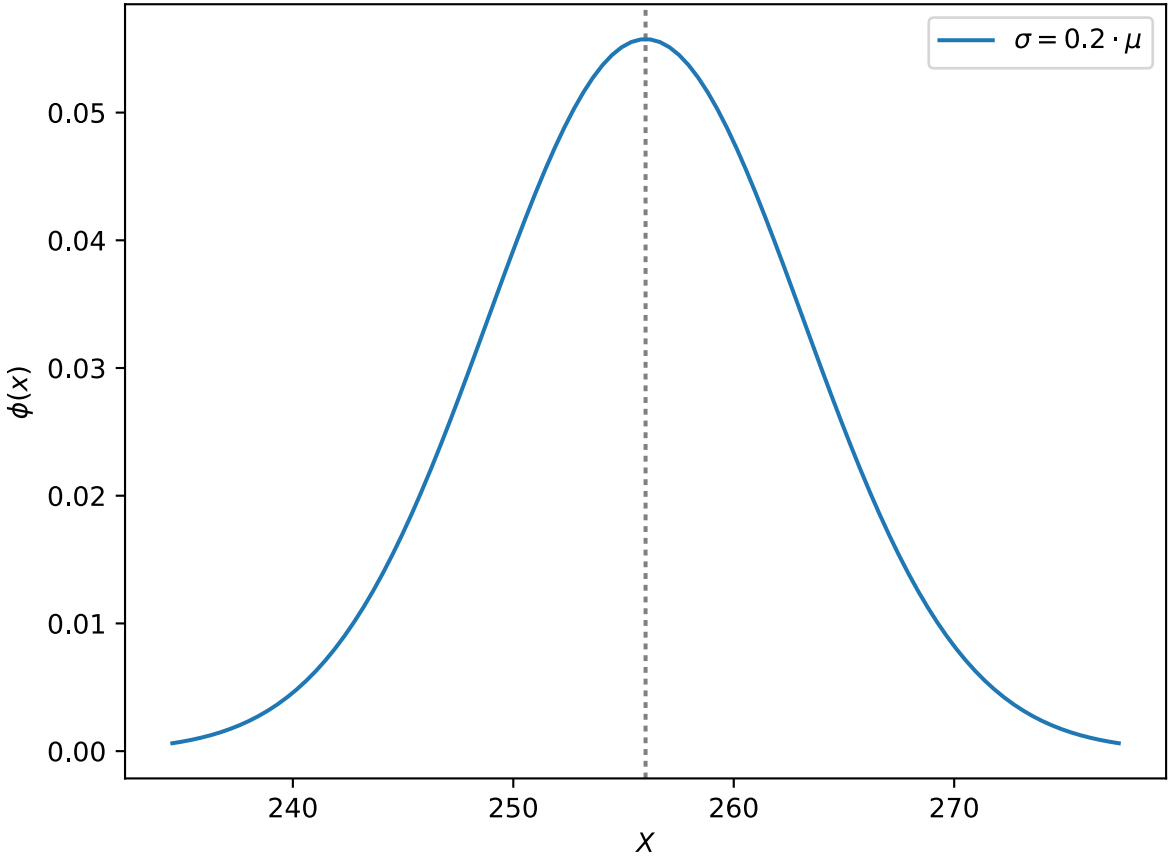
Karlsruher Institut für Technologie

Frage

Kann ich im Vorfeld bestimmen wie stark die Kompression vom Bitflip betroffen sein wird?

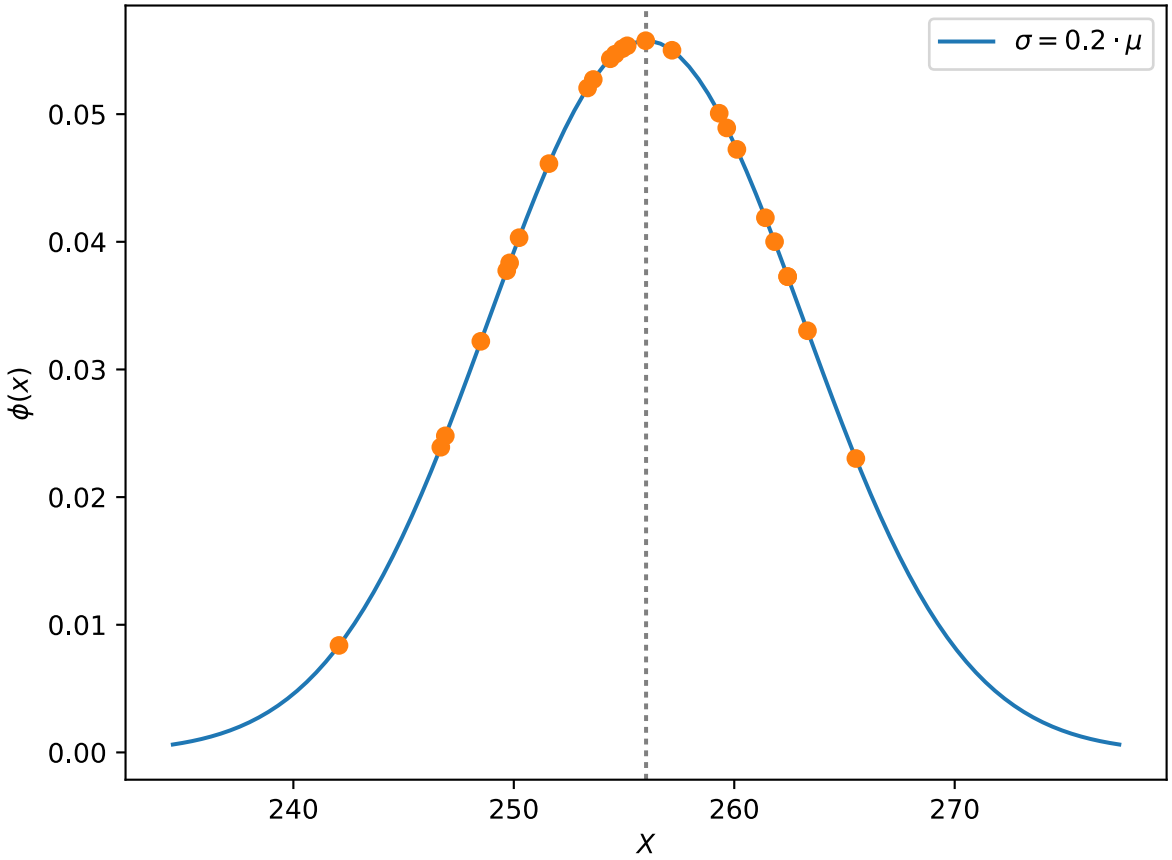
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Prämisse: Normalverteilung der Vorhersagen



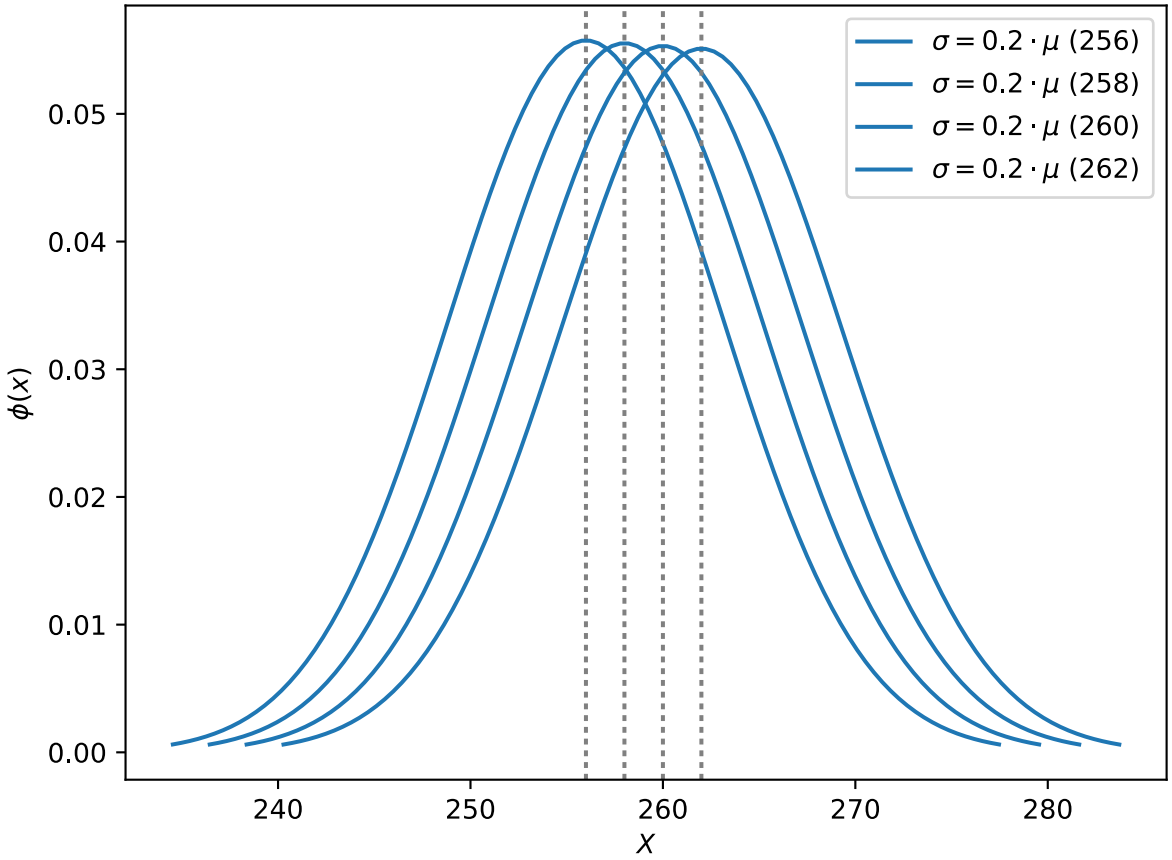
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Prämisse: Normalverteilung der Vorhersagen



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

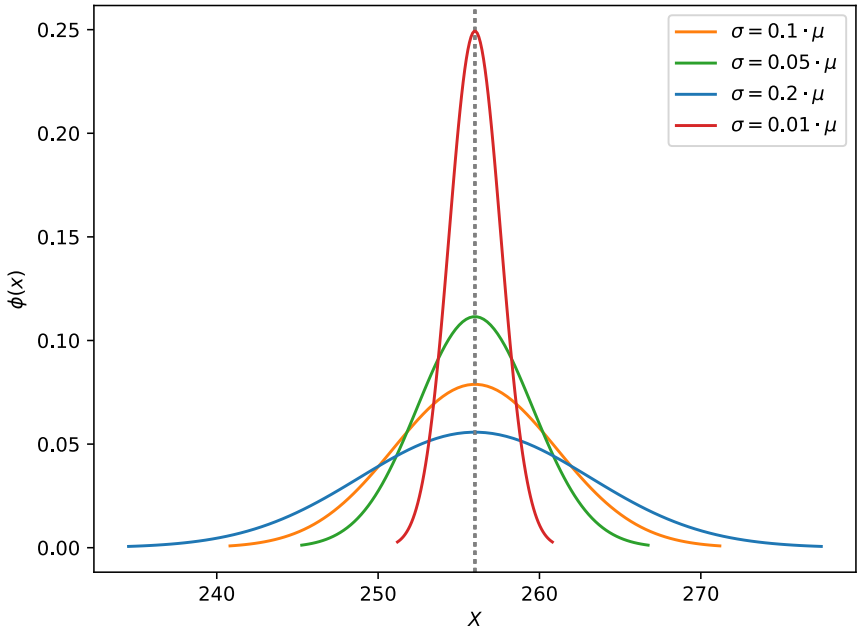
Prämisse: Normalverteilung der Vorhersagen



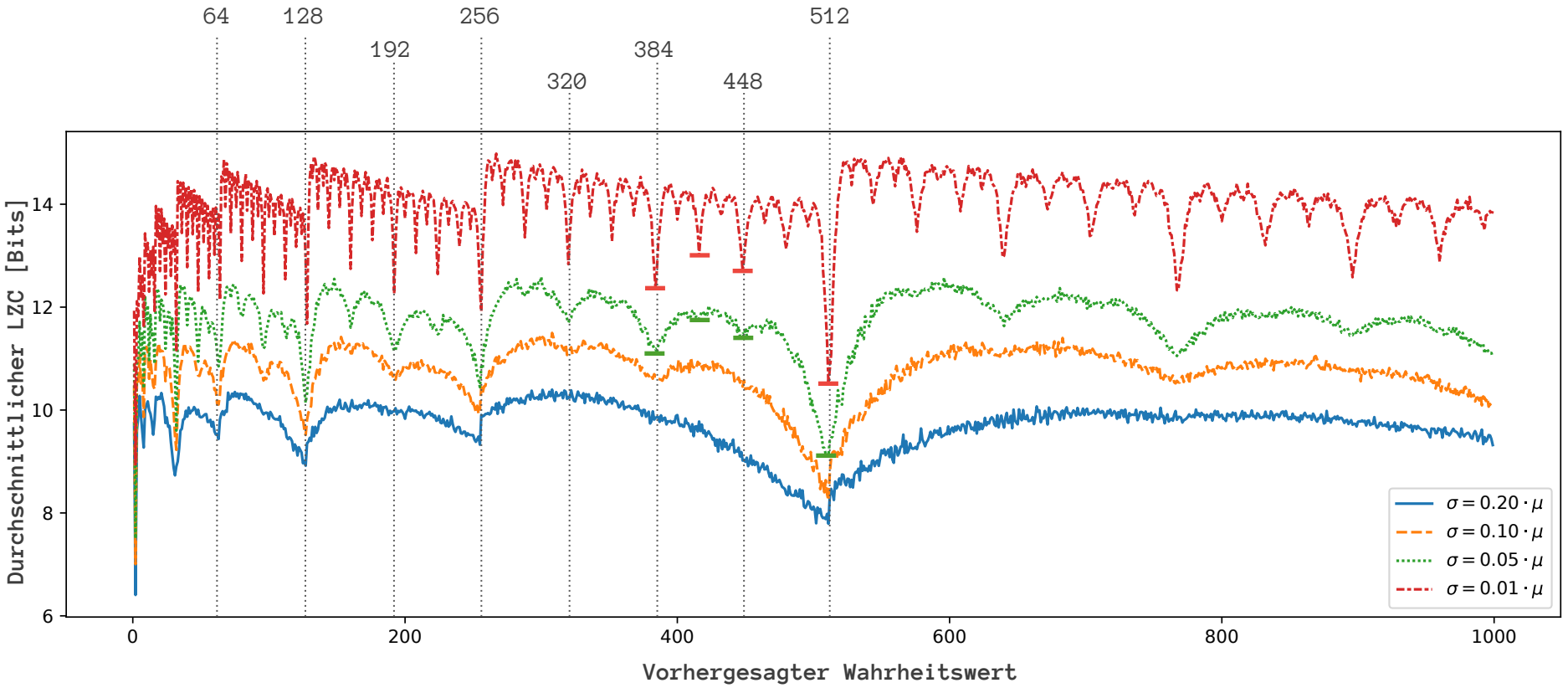
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

- Verschiedene normalverteilte Datensätze (100 Datenpunkte)
  - Erwartungswert (Wahrheit) $\mu \in [0; 1000]$ mit $\mu \in \mathbb{R}$
  - Standardabweichung (Vorhersagen) $\sigma \in \{0.2\mu, 0.1\mu, 0.05\mu, 0.01\mu\}$
- Berechnen des durchschnittlichen LZC



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor



Lösung zum Bitflip-Problem

Karlsruher Institut für Technologie

- Verschiebung der Differenzberechnung in einen anderen Wertebereich

$$\begin{aligned} \text{Vorhersage} &= V \\ \text{Wahrheit} &= W \\ V \oplus W &= R \end{aligned} \qquad \begin{aligned} V + \text{Shift} &= V' \\ W + \text{Shift} &= W' \\ V' \oplus W' &= R' \end{aligned} \qquad V + \text{Shift} = ?$$

- Eigenschaften vom Zielwert
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

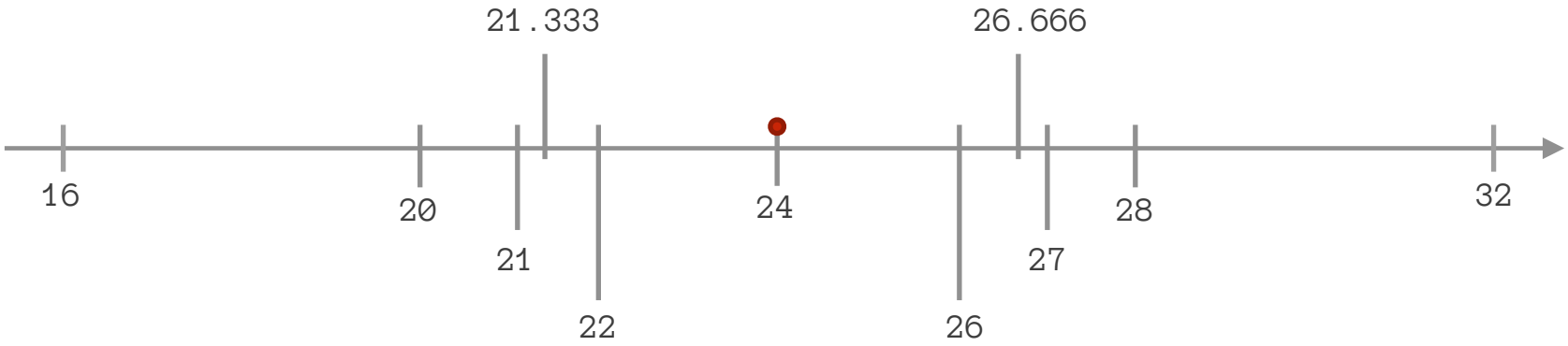
Lösung zum Bitflip-Problem

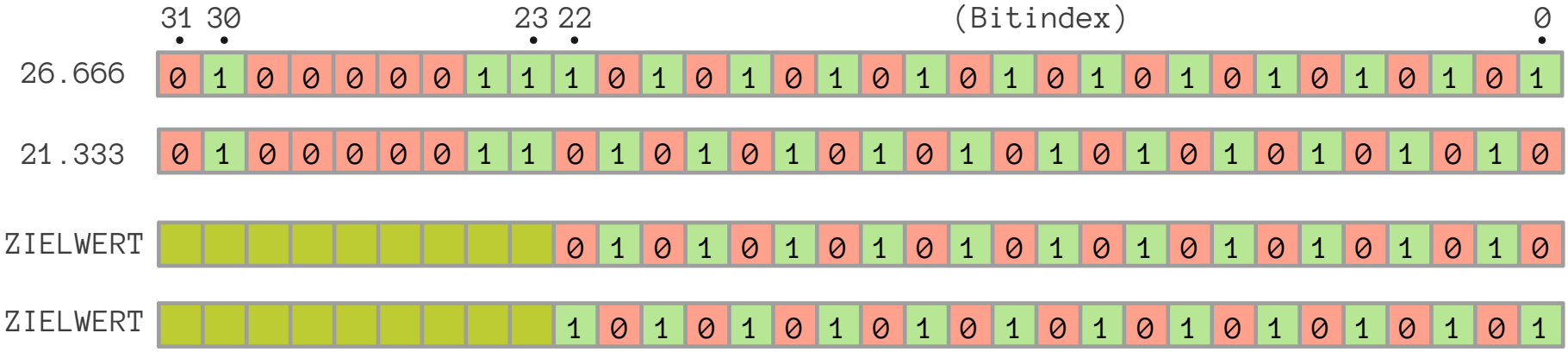
- Verschiebung der Differenzberechnung in einen anderen Wertebereich

```
Vorhersage = V        V + Shift = V'
  Wahrheit = W        W + Shift = W'        V + Shift = ?
     V ⊕ W = R          V' ⊕ W' = R'
```

- Eigenschaften vom Wertebereich
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

Vermeidung des Bitflip-Problems

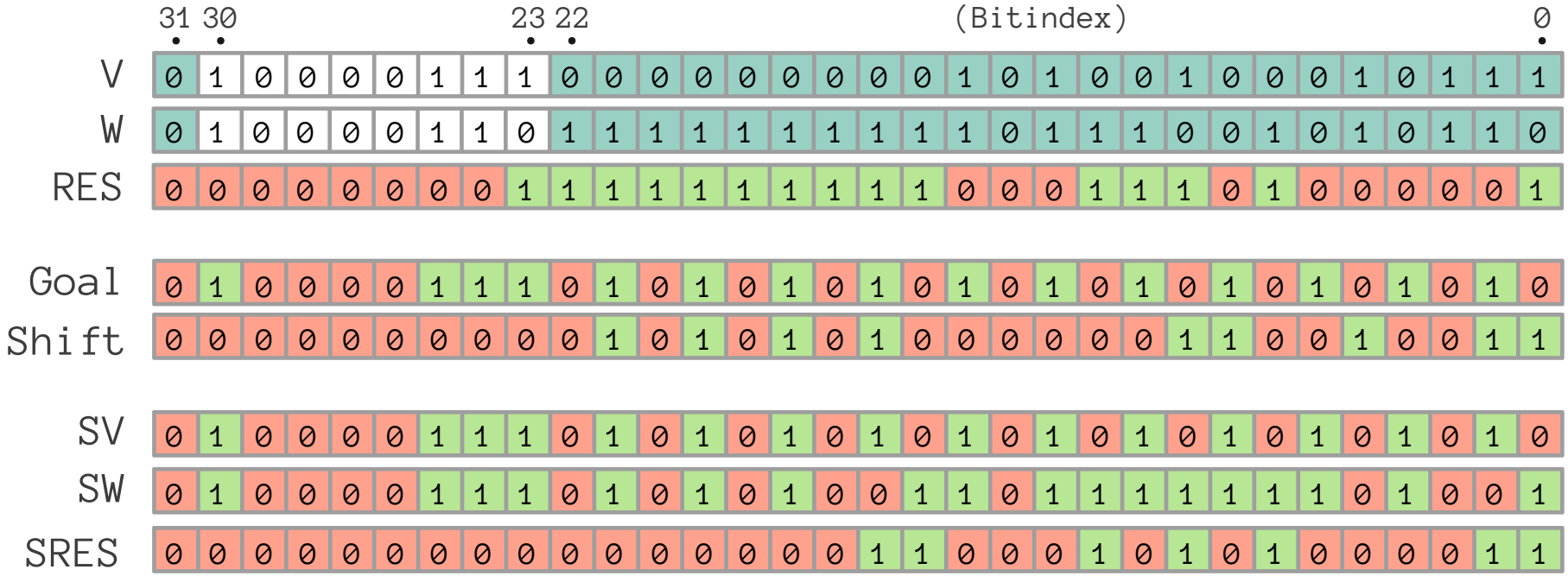




Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
Vorhersage (V):   256.321
Wahrer Wert (W):  255.931
```



LZC: 8 -> 16

Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
 Vorhersage (V):  256.321
Wahrer Wert (W):  255.931          LZC: 8 -> 16
```

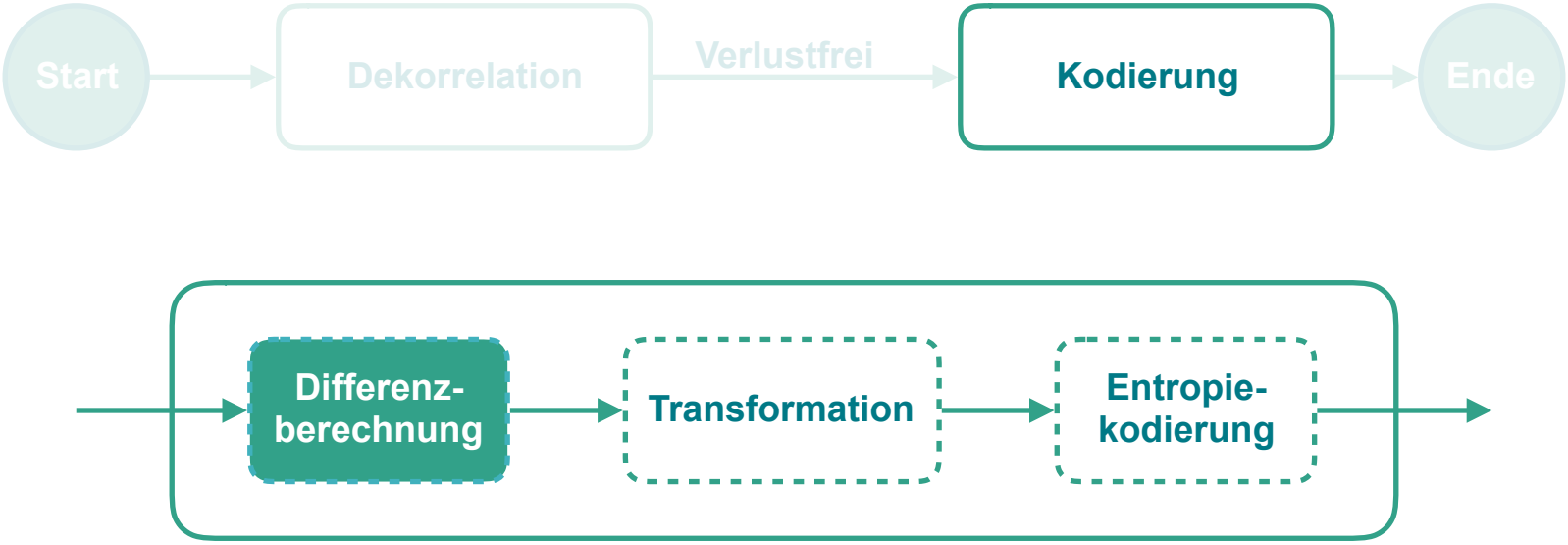
Es funktioniert besser, je näher die Zahlen an Zweierpotenzen liegen

```
 Vorhersage (V):  256.002
Wahrer Wert (W):  255.991          LZC: 8 -> 21
```

Es funktioniert besser, je größer die Zahlen sind

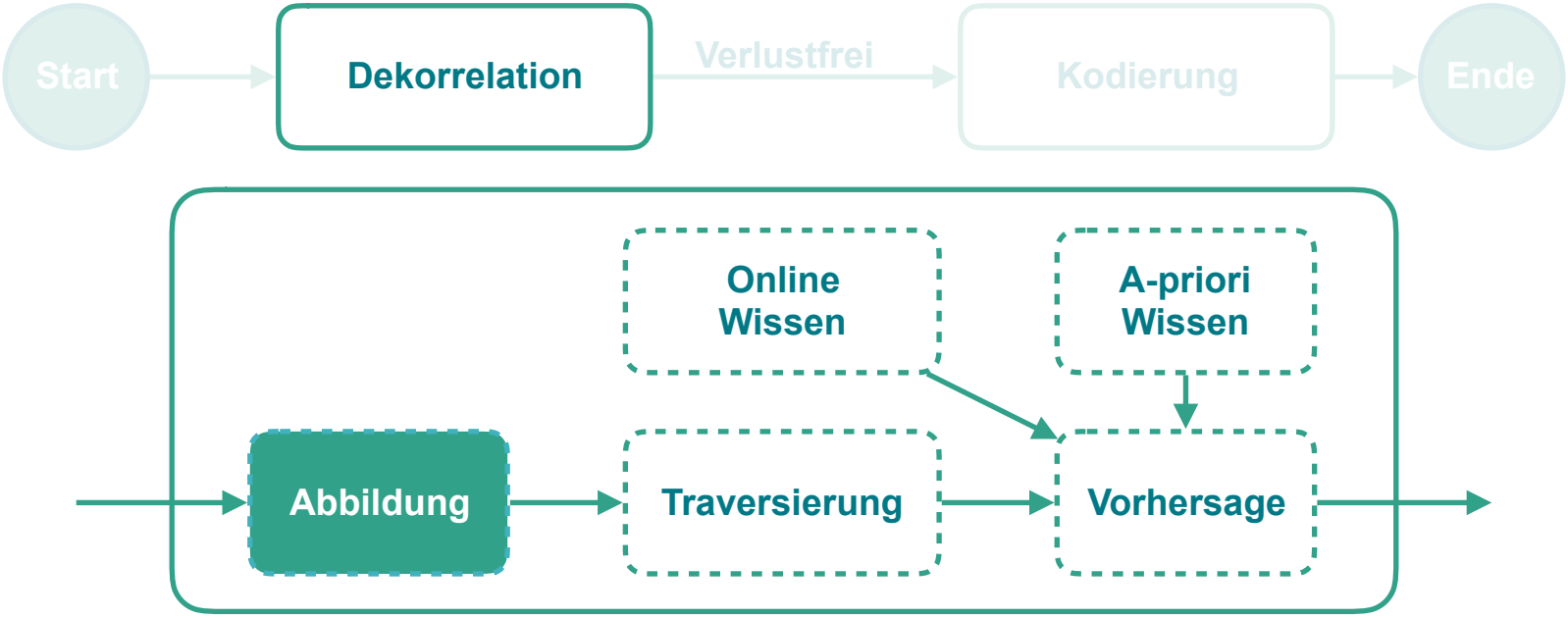
```
 Vorhersage (V): 1024.002
Wahrer Wert (W): 1023.991          LZC: 8 -> 24
```

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Pascal Zip (pzip)

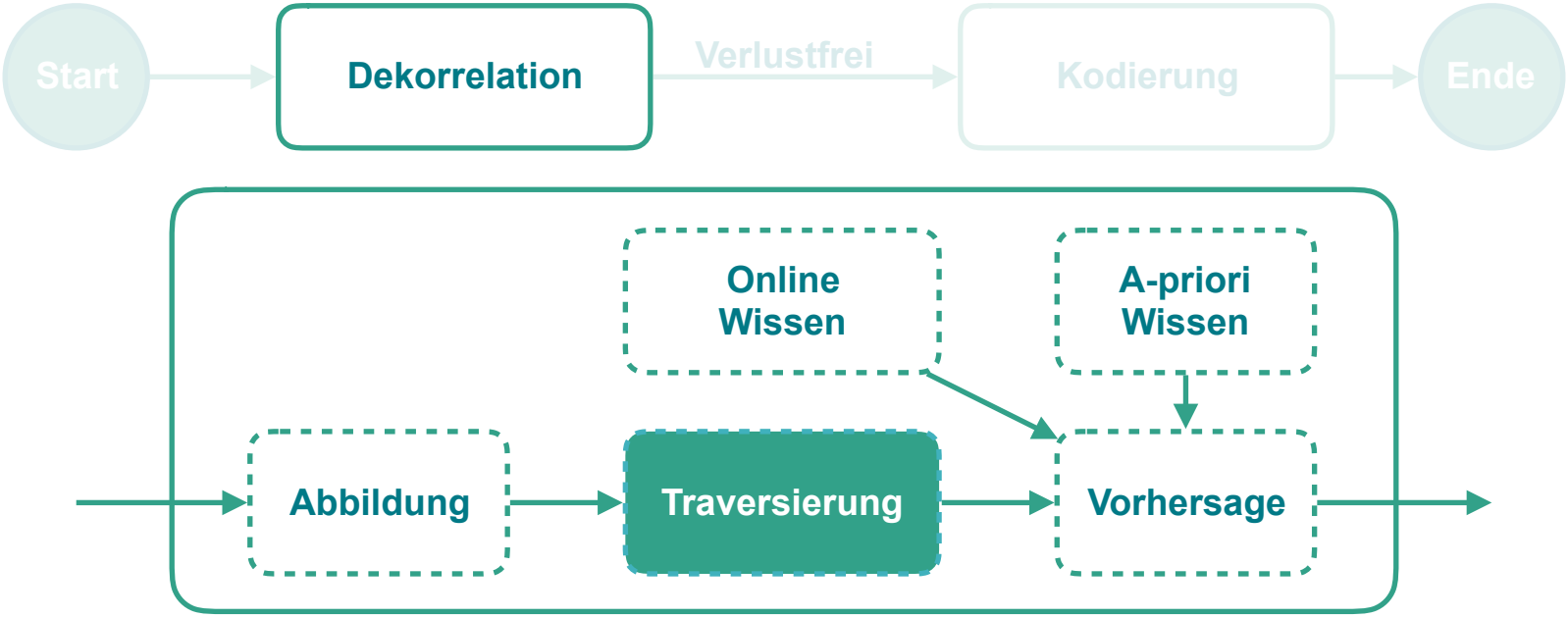
Pascal Zip (pzip)



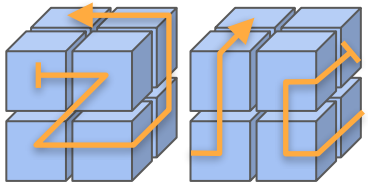
$$m : \mathbb{R} \rightarrow \mathbb{N}$$

```
256.321 -> 1132472599
255.931 -> 1132457558
    . . .
```

Pascal Zip (pzip)

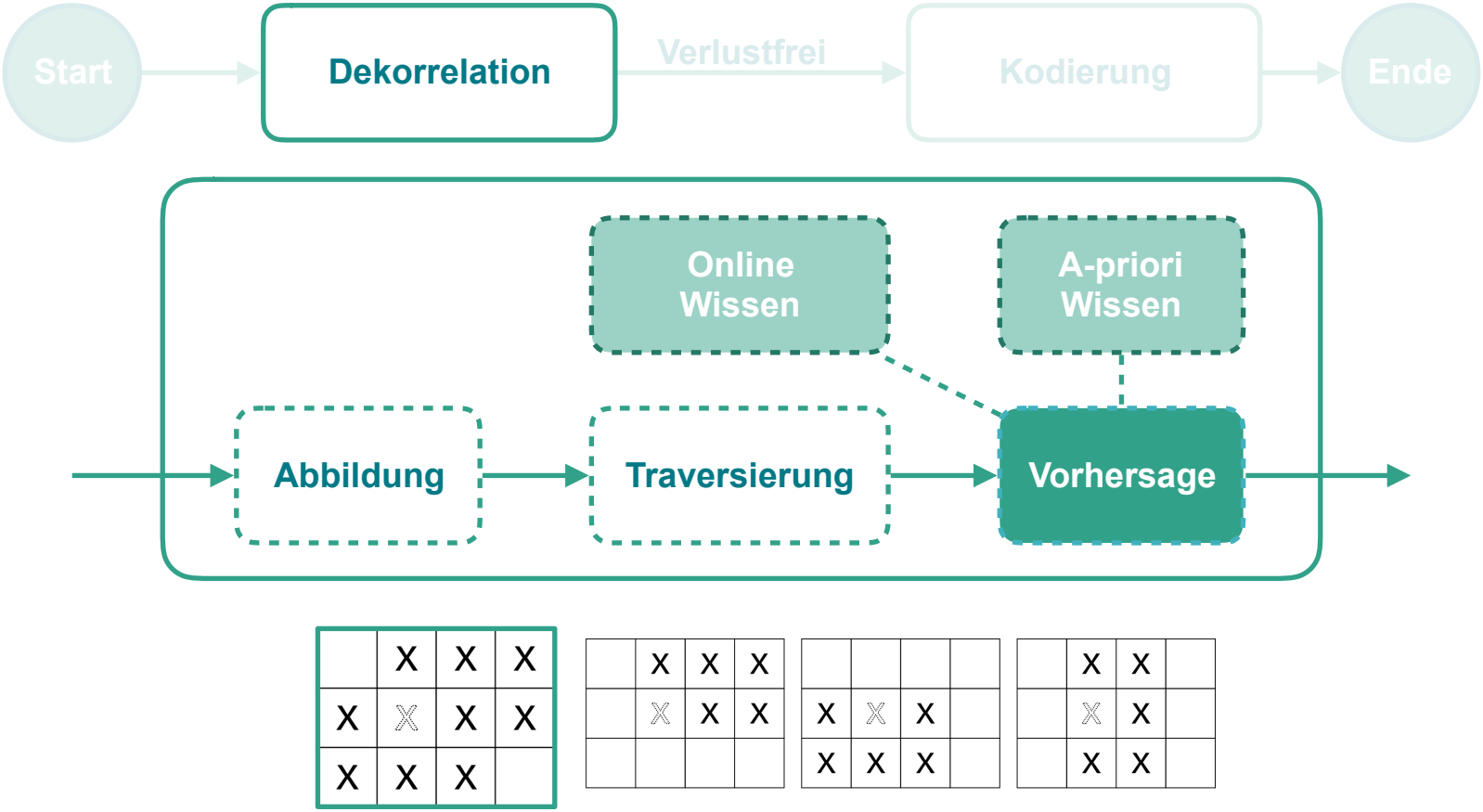


$$t : \mathbb{N} \rightarrow \mathbb{N}$$



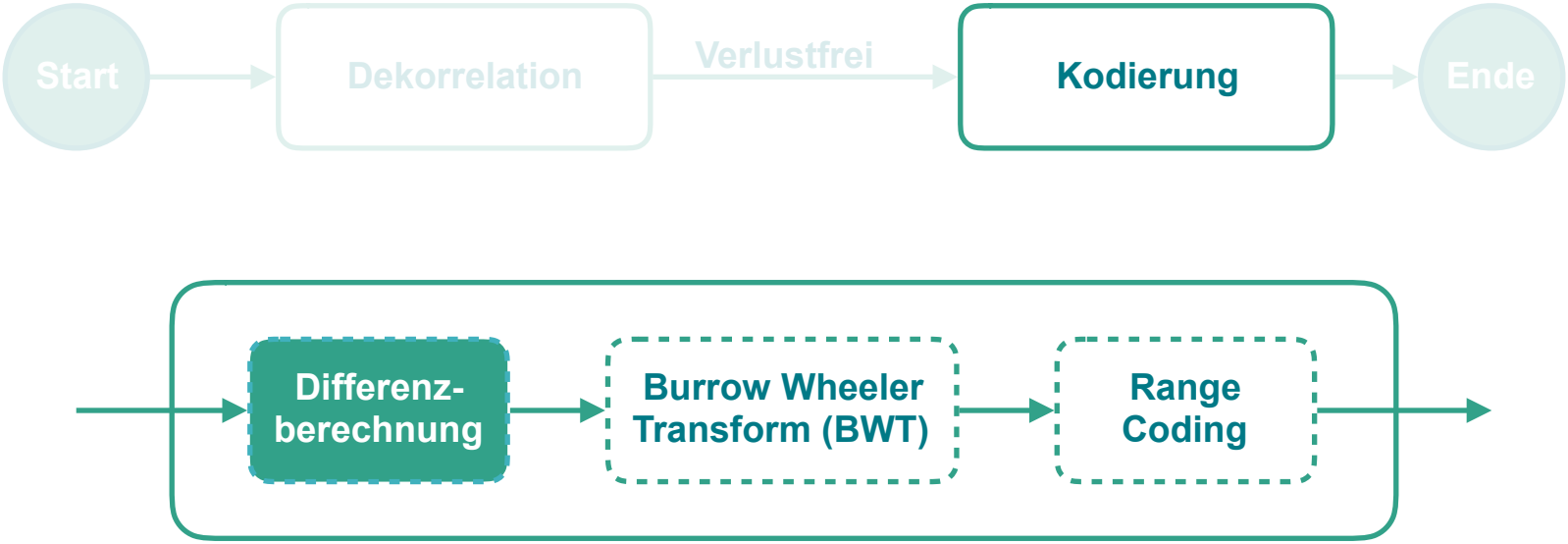
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**
Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression **GI INFORMATIK 2019**

Pascal Zip (pzip)



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output **EGU 2018**
Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using SARIMA **IEEE eScience 2017**
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**

Pascal Zip (pzip)

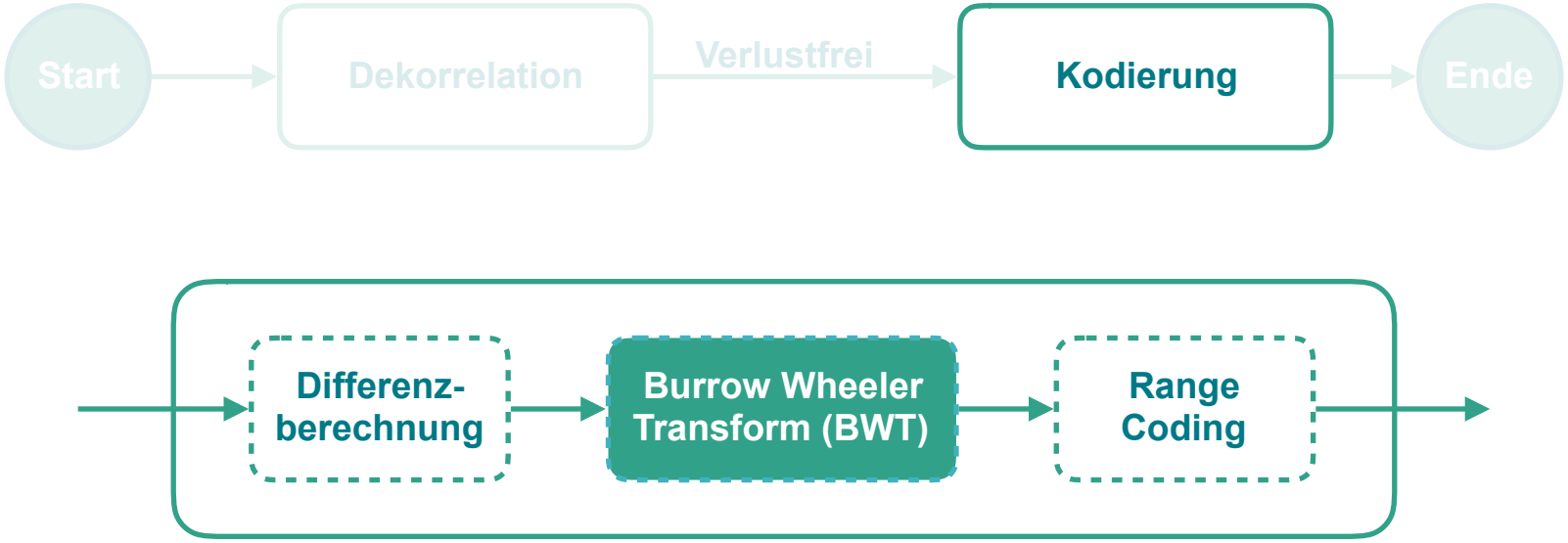



```
V + Shift = V'
W + Shift = W'
  V' ⊕ W' = R'
```

```
R' = 00000000000000111010111001000 1011
LZC = 14, FOC = 03, RES = 101110010001011
```

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms **IEEE eScience 2019**

Pascal Zip (pzip)



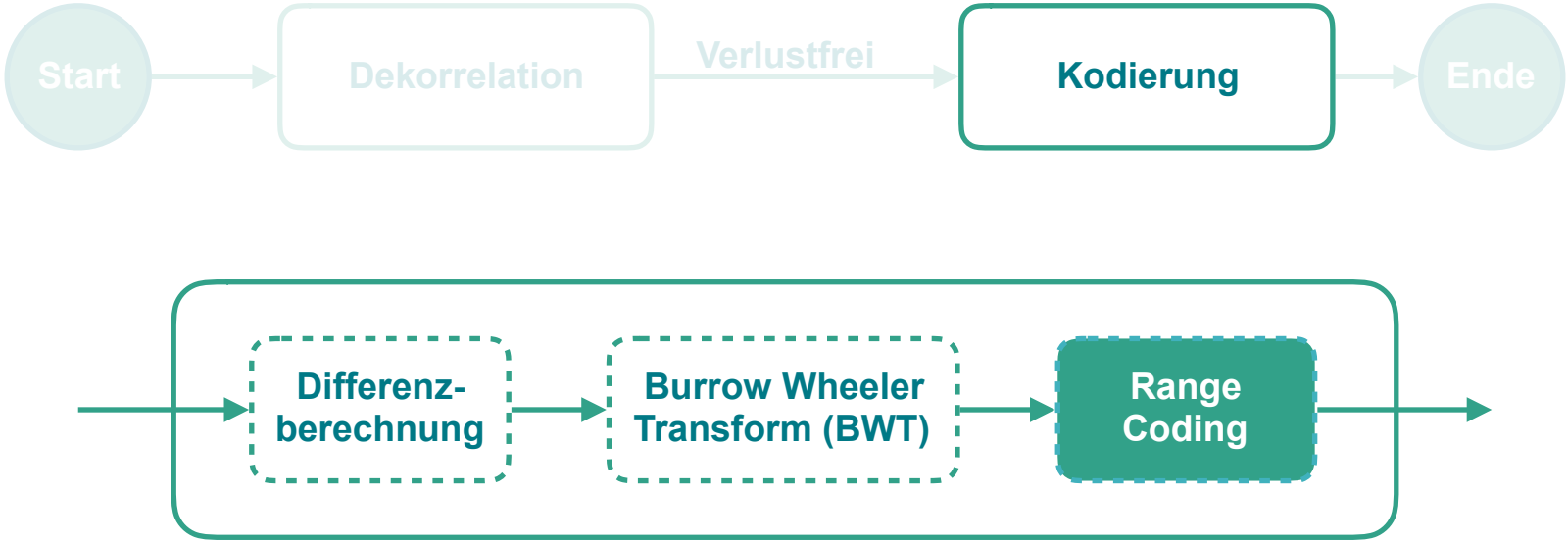
^BANANA$

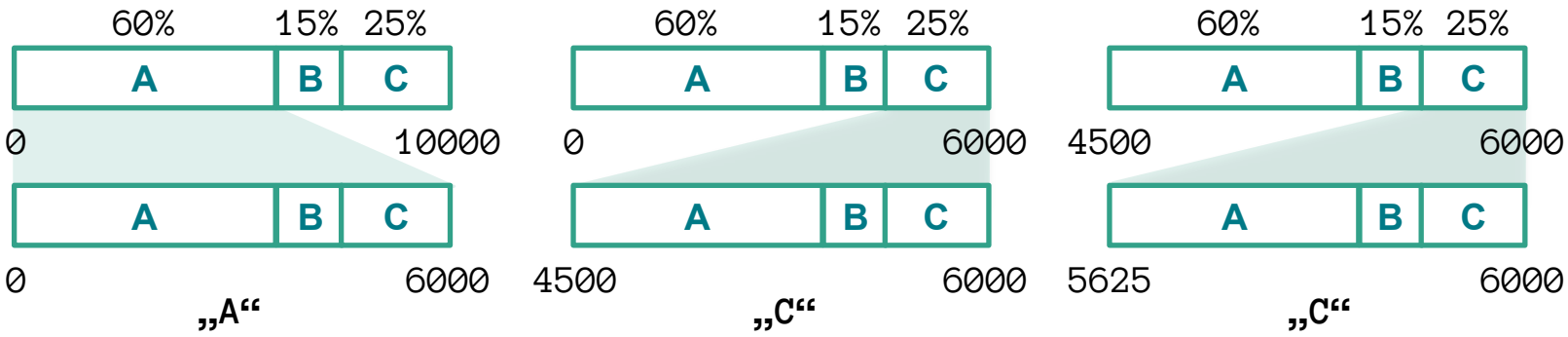
^BANANA$	**A**NANA$^B	ANANA$^**B**
$^BANANA	**A**NA$^BAN	ANA$^BA**N**
A$^BANAN	**A$**^BANAN	A$^BANA**N**
NA$^BANA	**B**ANANA$^	BANANA$^
ANA$^BAN	**N**ANA$^BA	NANA$^B**A**
NANA$^BA	**N**A$^BANA	NA$^BAN**A**
ANANA$^B	^BANANA$	^BANANA**$**
BANANA$^	**$**^BANANA	$^BANAN**A**

BNN^AA$A

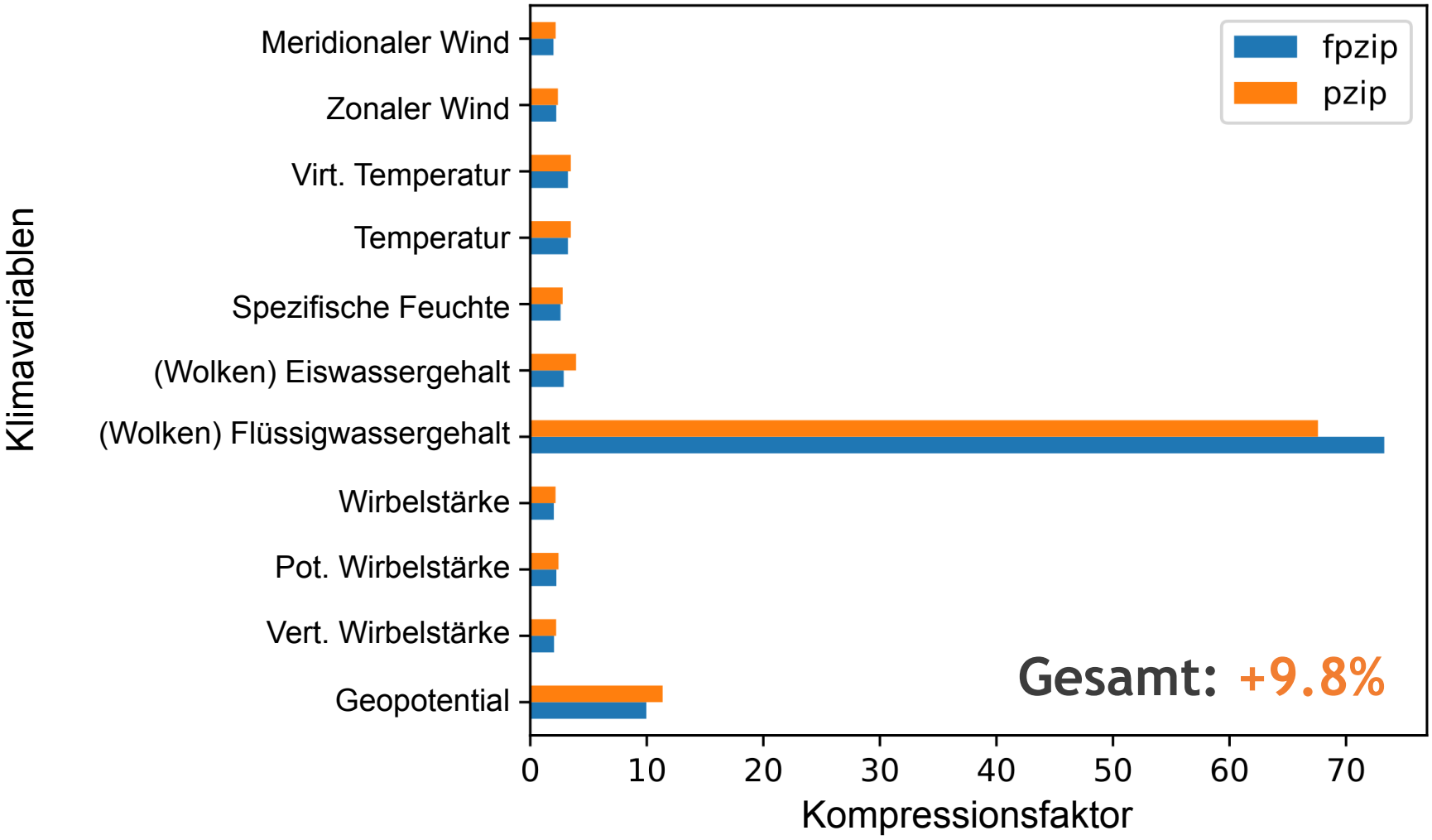
Quelle[3]

Pascal Zip (pzip)

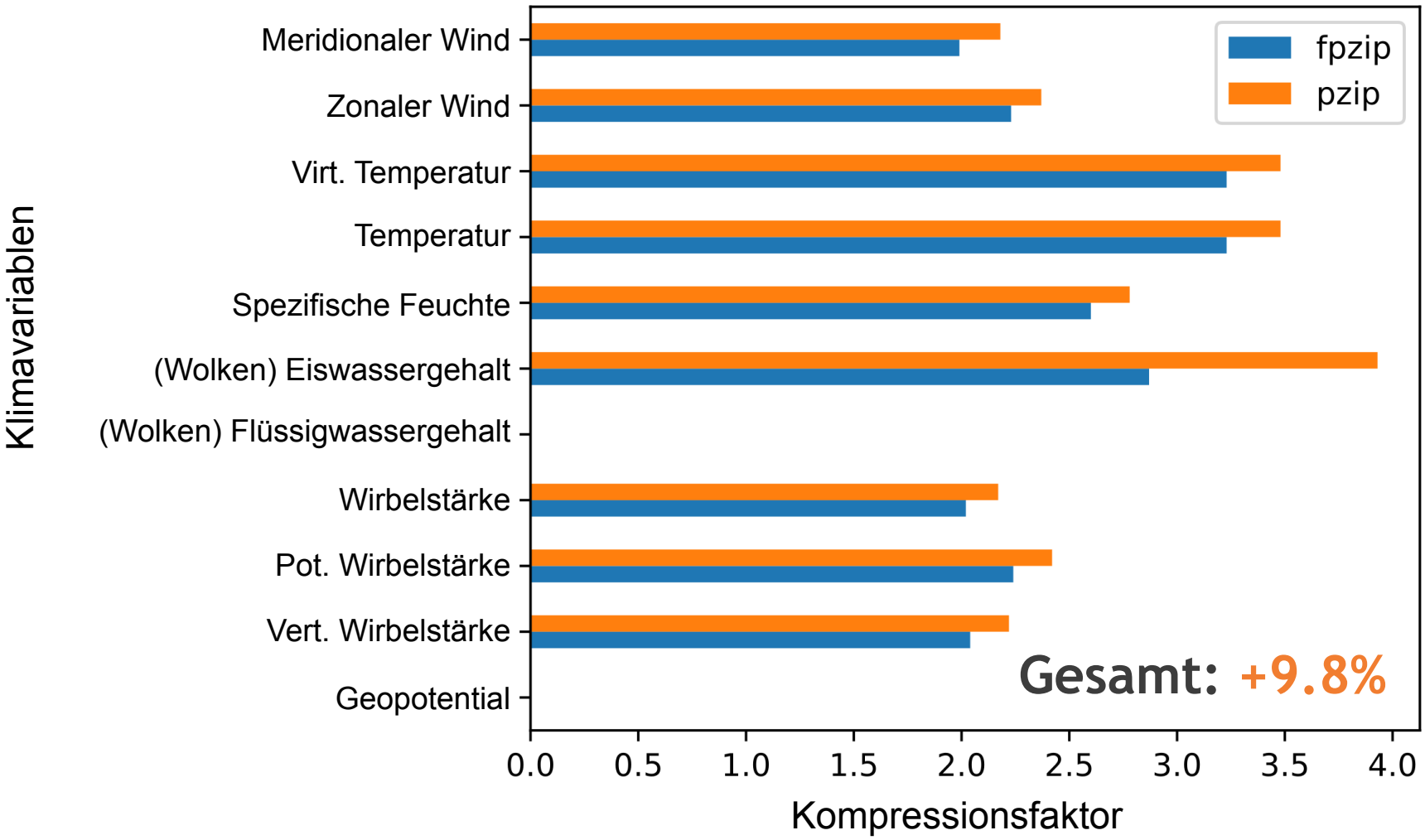




Kompressionsfaktor

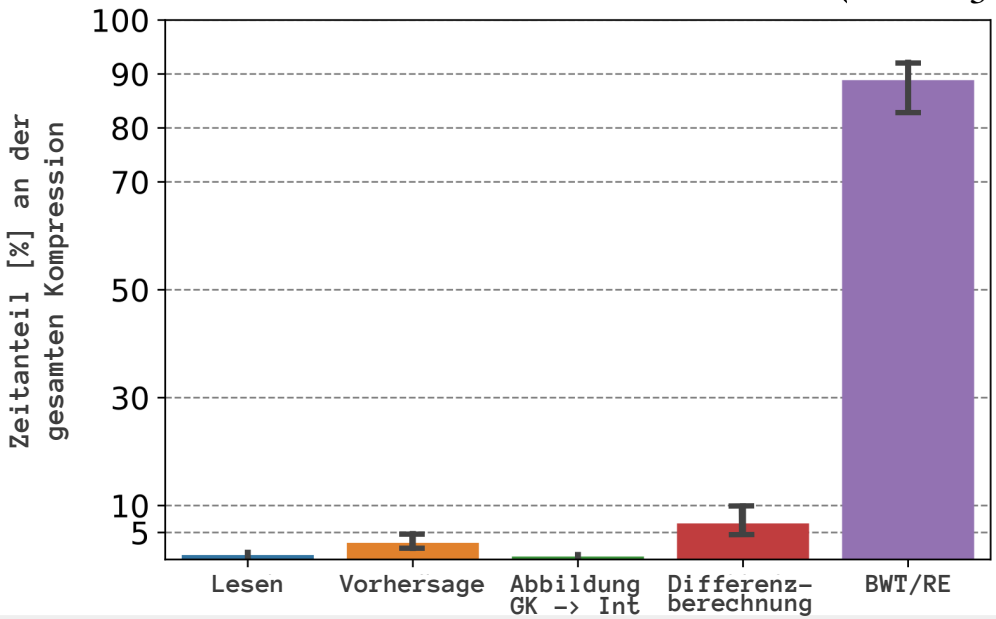


Kompressionsfaktor



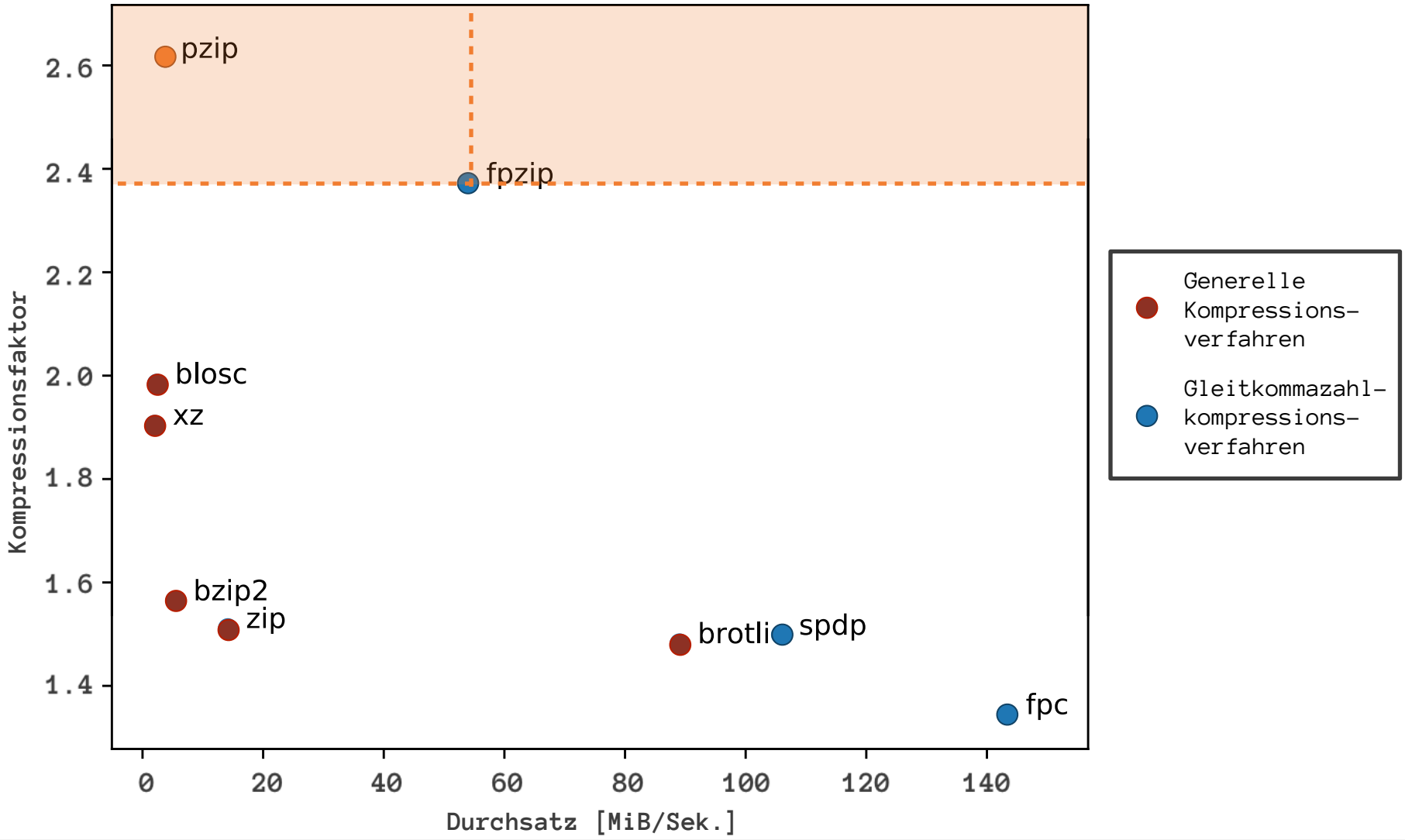
Durchsatz und Komplexität

- Engpass in der aktuellen Implementierung ist BWT/RE
- Laufzeit- und Speicherplatzkomplexität von fpzip $\mathcal{O}(n)$
- Laufzeitkomplexität $\mathcal{O}(n + 4 \cdot \tau \cdot n + n)$
- Speicherkomplexität $\mathcal{O}(\tau \cdot \left(1 + \frac{n}{d_3}\left(\frac{1}{d_2}\left(\frac{1}{d_1} + 1\right) + 1\right)\right) + n \log \sigma)$



τ = Nachbarschaft
$n = d_0 d_1 d_2 d_3$
σ = |Alphabet|

Kompressionsverfahren im Vergleich



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Reduktion

		~2.6	
ERA5:	10.89 PiB	⇨	4.19 PiB
IMK-ASF:	770 TiB	⇨	296 TiB

Open Source

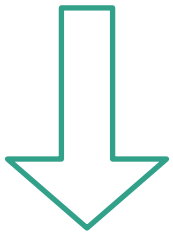
Alle Programmbeispiele und Daten sind öffentlich zugänglich (github.com, GPLv3)

Beitrag

Andere Kompressionsverfahren können einzelne Entwicklungen aus der Arbeit aufgreifen und einbauen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor erfüllt ✔



Verlustfreie Kompression von Klimadaten

Vielen Dank

Brücke

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften

- Varianz
- Mutual Information
- Entropie

2. Analyse des Stand der Technik

- Kompressionsfaktor/Durchsatz

3. Integration von vorhandenem Wissen

- Klimaindizes
- ARIMA Modell
- Verlustbehaftete Zeitreihenkompression

4. Schaffung und Integration von aktuellem Wissen

- Informationsräume/-kontexte
- Auswirkungen der Startposition
- Auswirkungen der Traversierung

5. Analyse und Optimierung der Kodierung

- Verschiebung vom Wertebereich (Details)
- Vergleich von Kodierungen

6. Schnelle Prototypenentwicklung

- Modifizierer & Objekte
- UML-Modell & Erw. Module

BS

UF

PR

Karlsruher Institut für Technologie

Backup slides

Backup slides

Categories for Compression Methods

Karlsruher Institut für Technologie

adaptive & non-adaptive

symmetric & asymmetric

single-pass & multi-pass

streaming & block-mode

universal & custom

Kompression von anderen Daten

Karlsruher Institut für Technologie

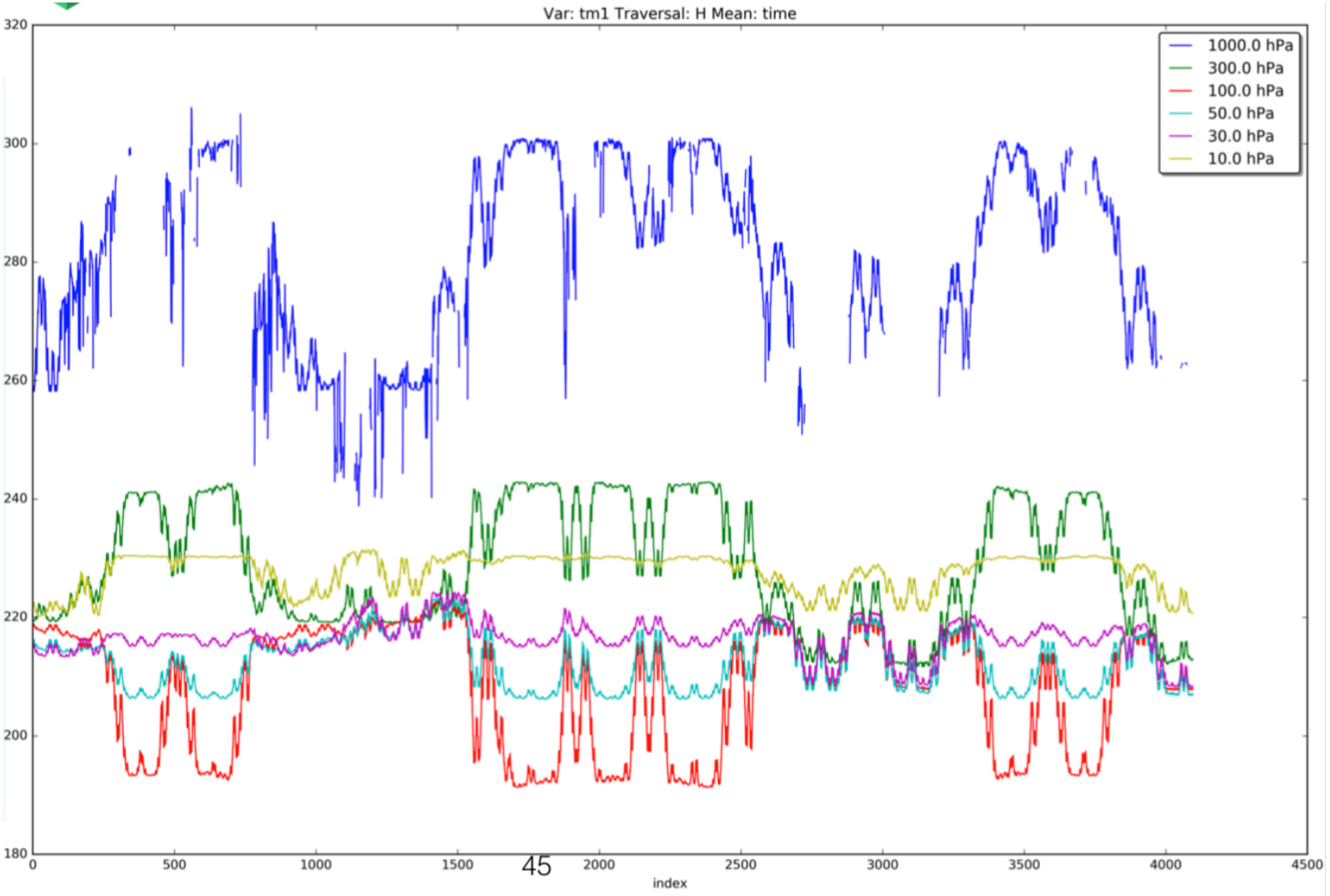
- Quantum Monte Carlo Dataset (QMC)
  - Test für die elektr. Struktur von Atome & Molekülen
  - Argonne National Laboratory
  - fpzip: 1.76, pzip: 1.92 **+9.3%**
- NYX Dataset (adaptive mesh)
  - „Cosmological Hydrodynamics Simulation Dataset“
  - Argonne National Laboratory
    - Darkmatter density: fpzip: 1.18, pzip: 1.40 **+18.6%**
    - Baryon density: fpzip: 1.60, pzip: 1.59 **-0.6%**
    - Velocity:
      - x —> fpzip: 1.76, pzip: 1.78 **+1.1%**
      - y —> fpzip: 1.69, pzip: 1.71 **+1.1%**
      - z —> fpzip: 1.64, pzip: 1.67 **+1.8%**

Karlsruher Institut für Technologie

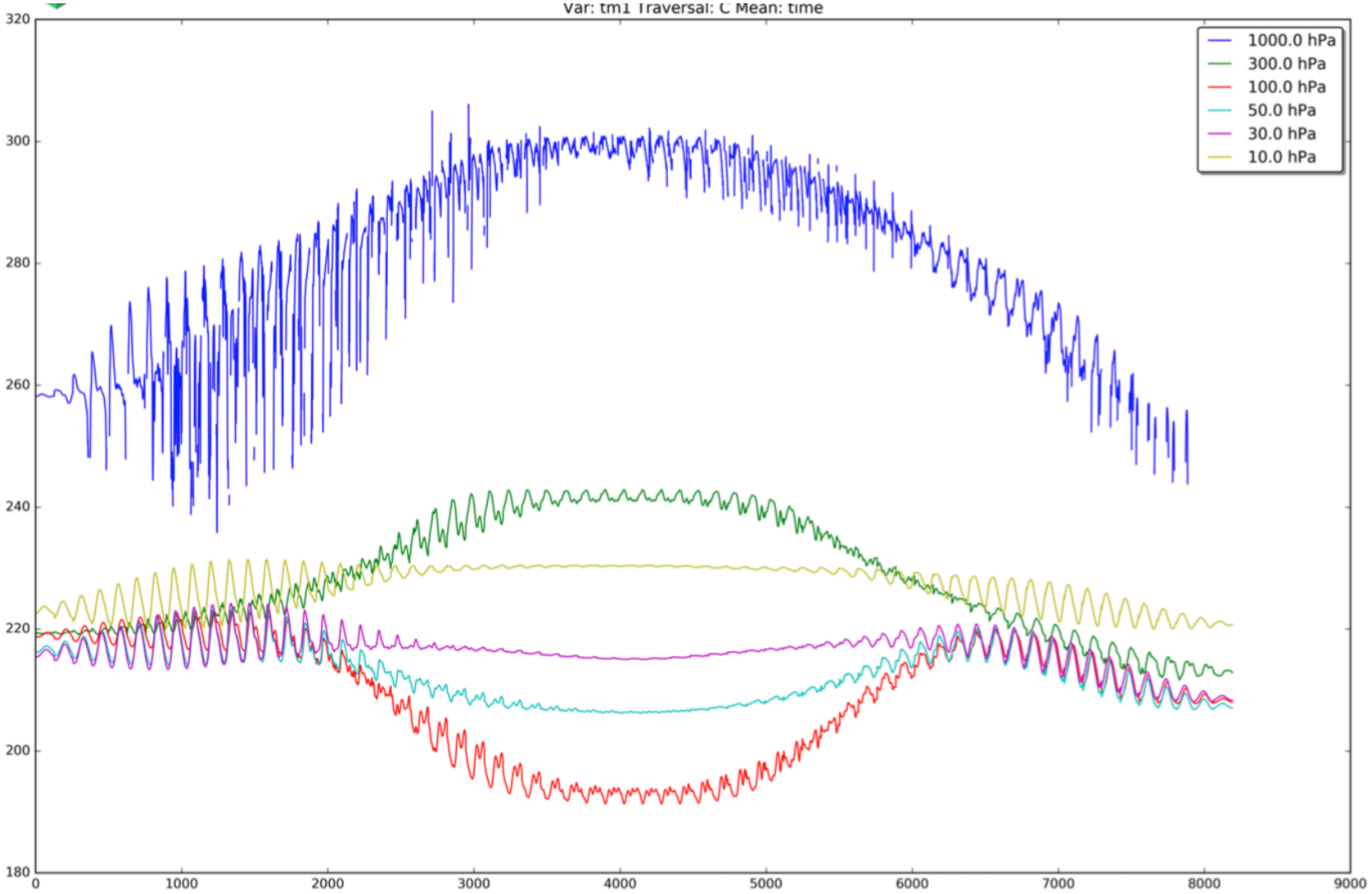
Hilbert curves

Hilbert

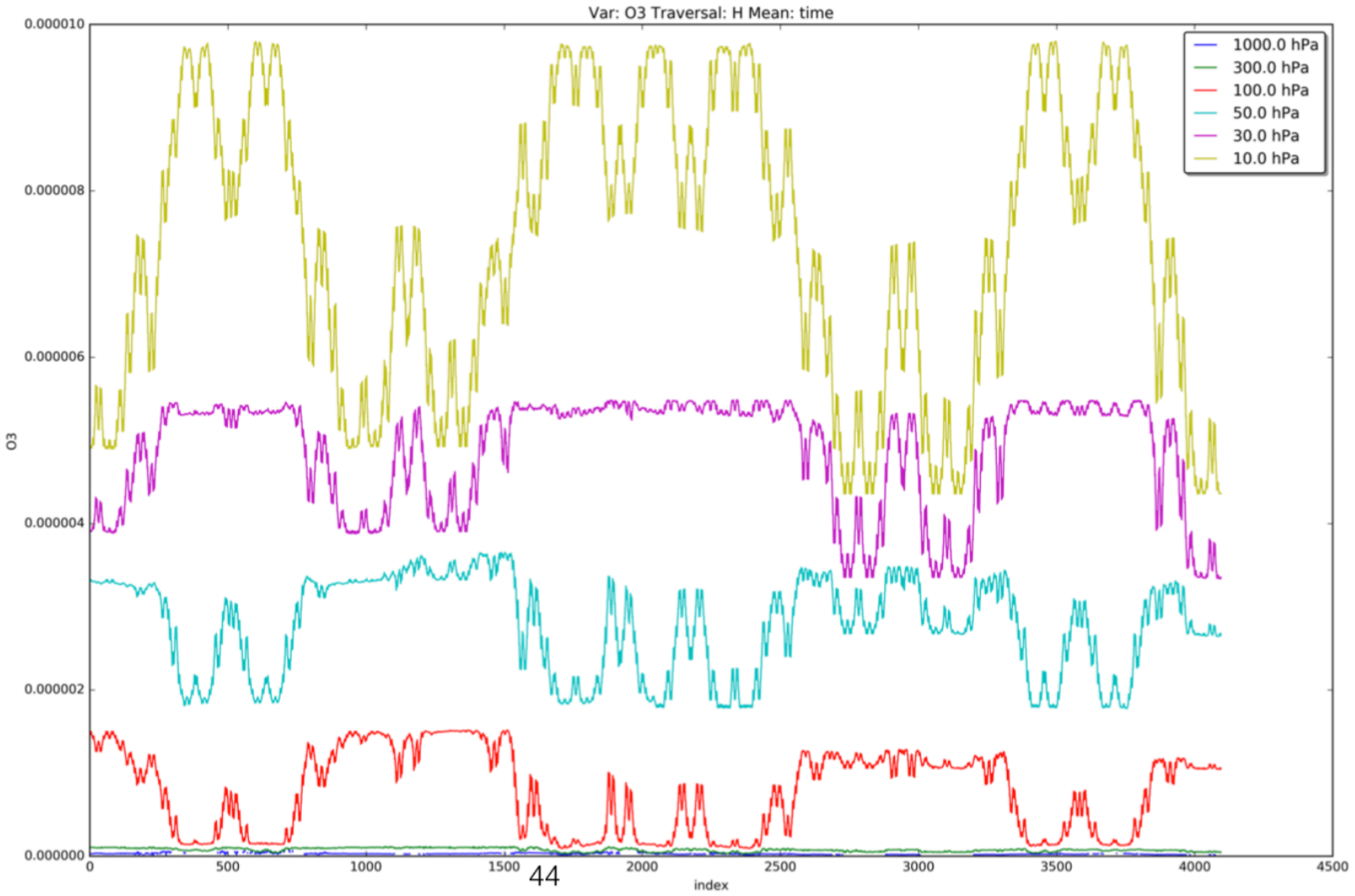
Hilbertkurven - Temperatur



C-Traversal - Temperatur

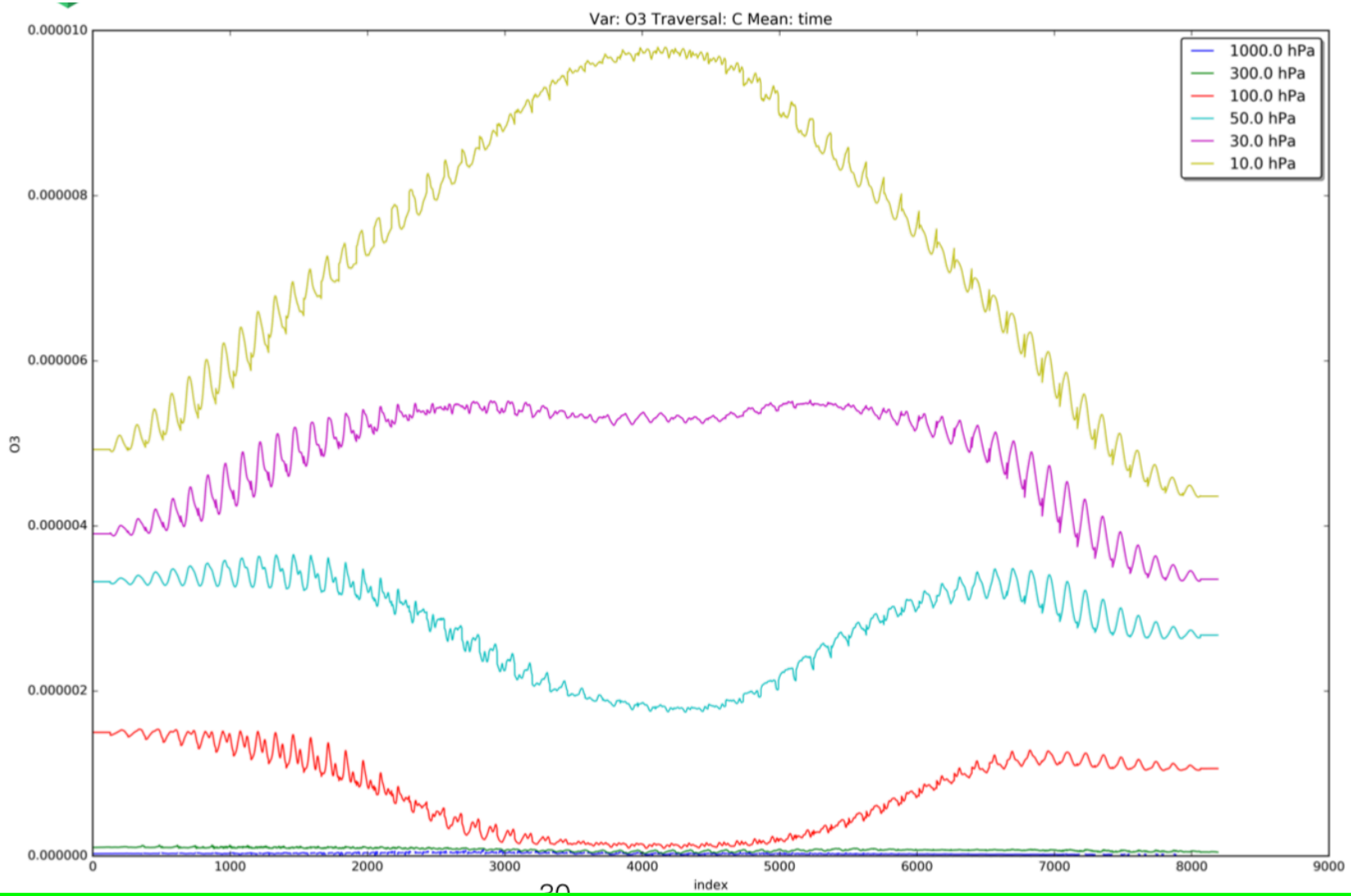


Hilbertkurven - Ozon



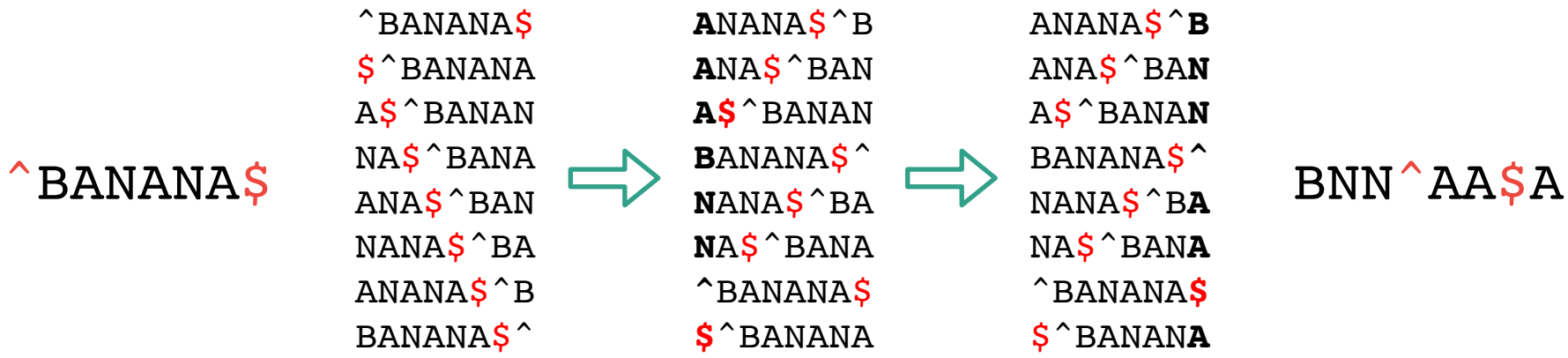
44

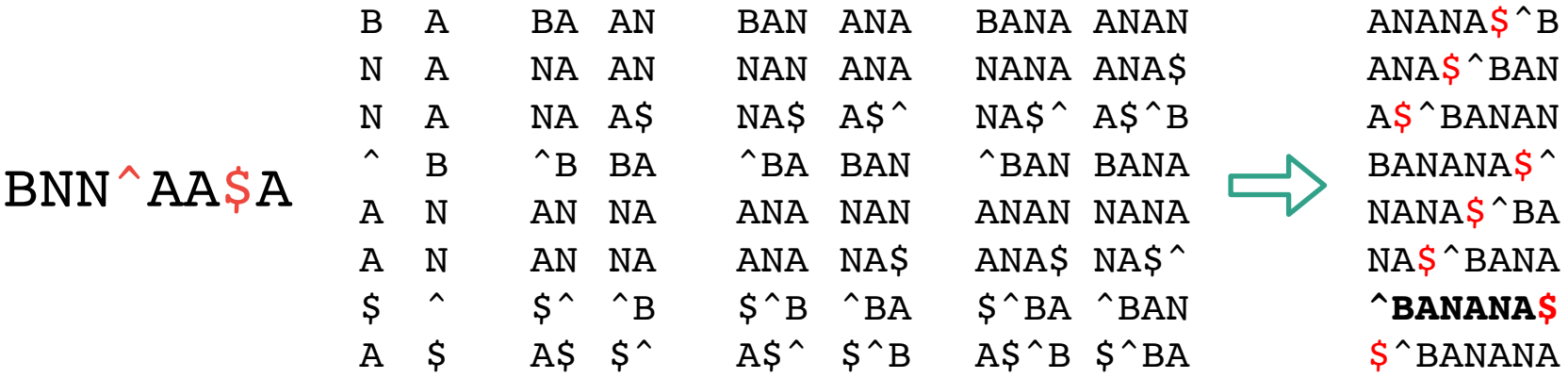
C-Traversal - Ozon



BWT Algorithmus

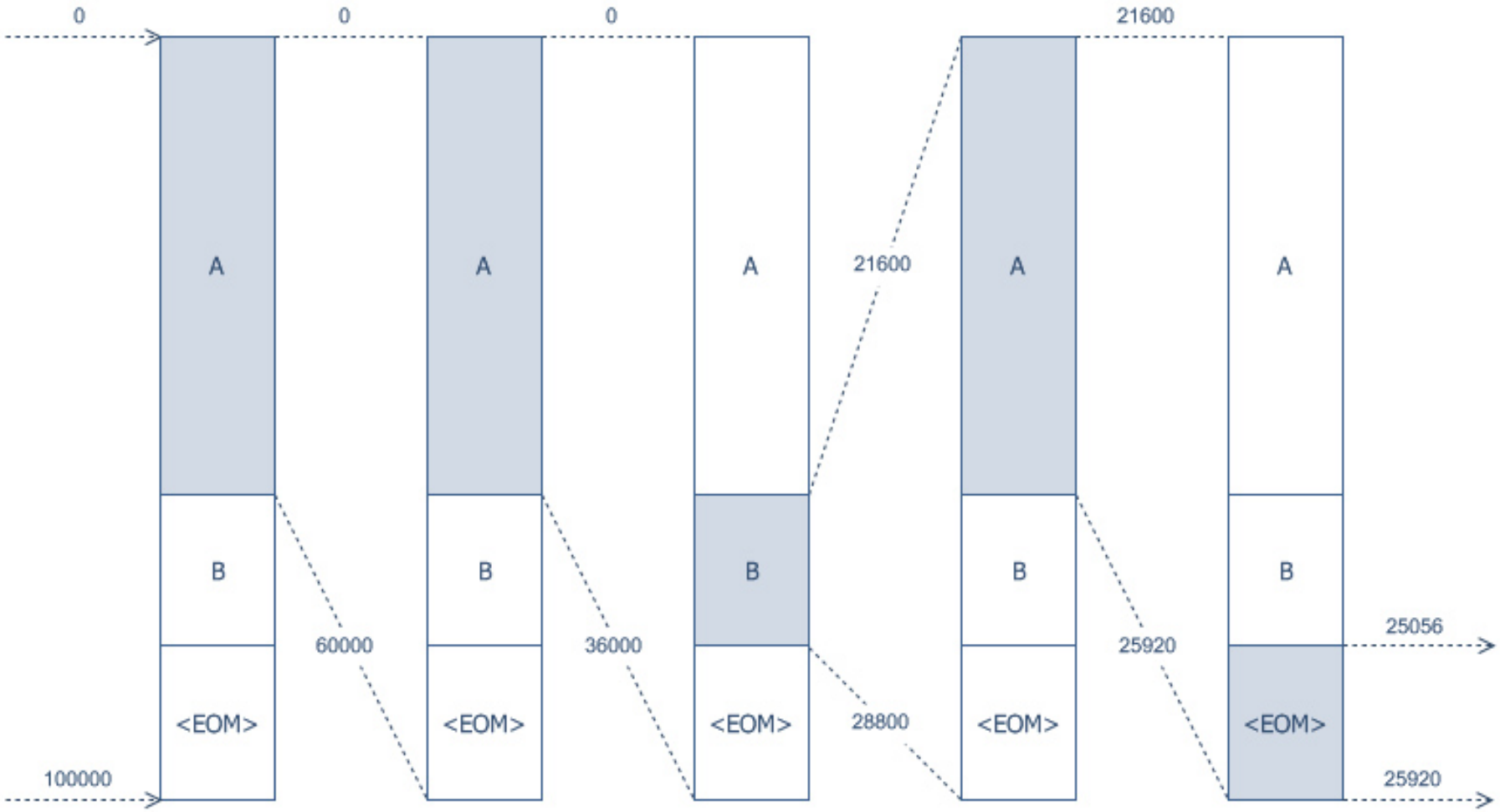
Karlsruher Institut für Technologie



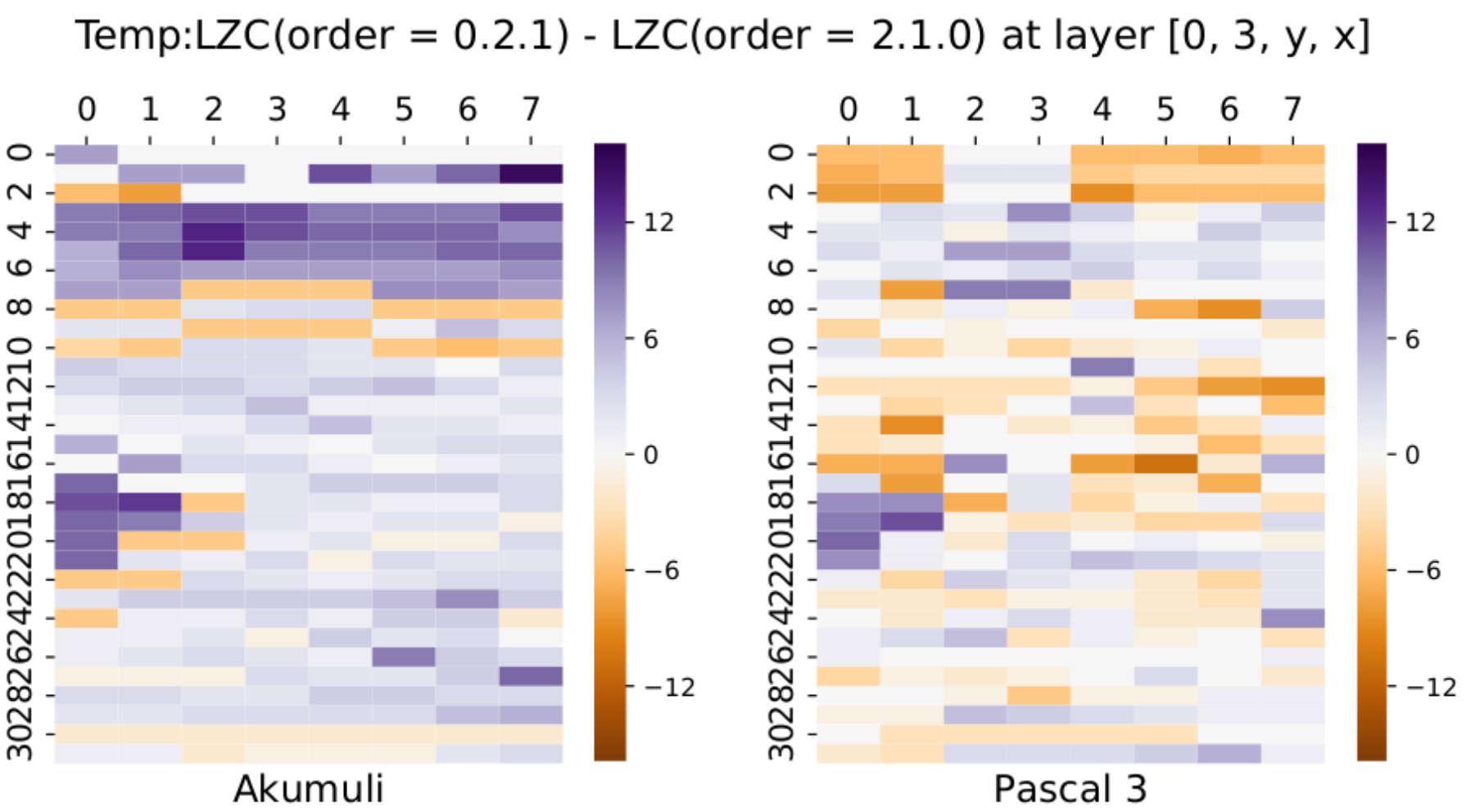


Source[4]

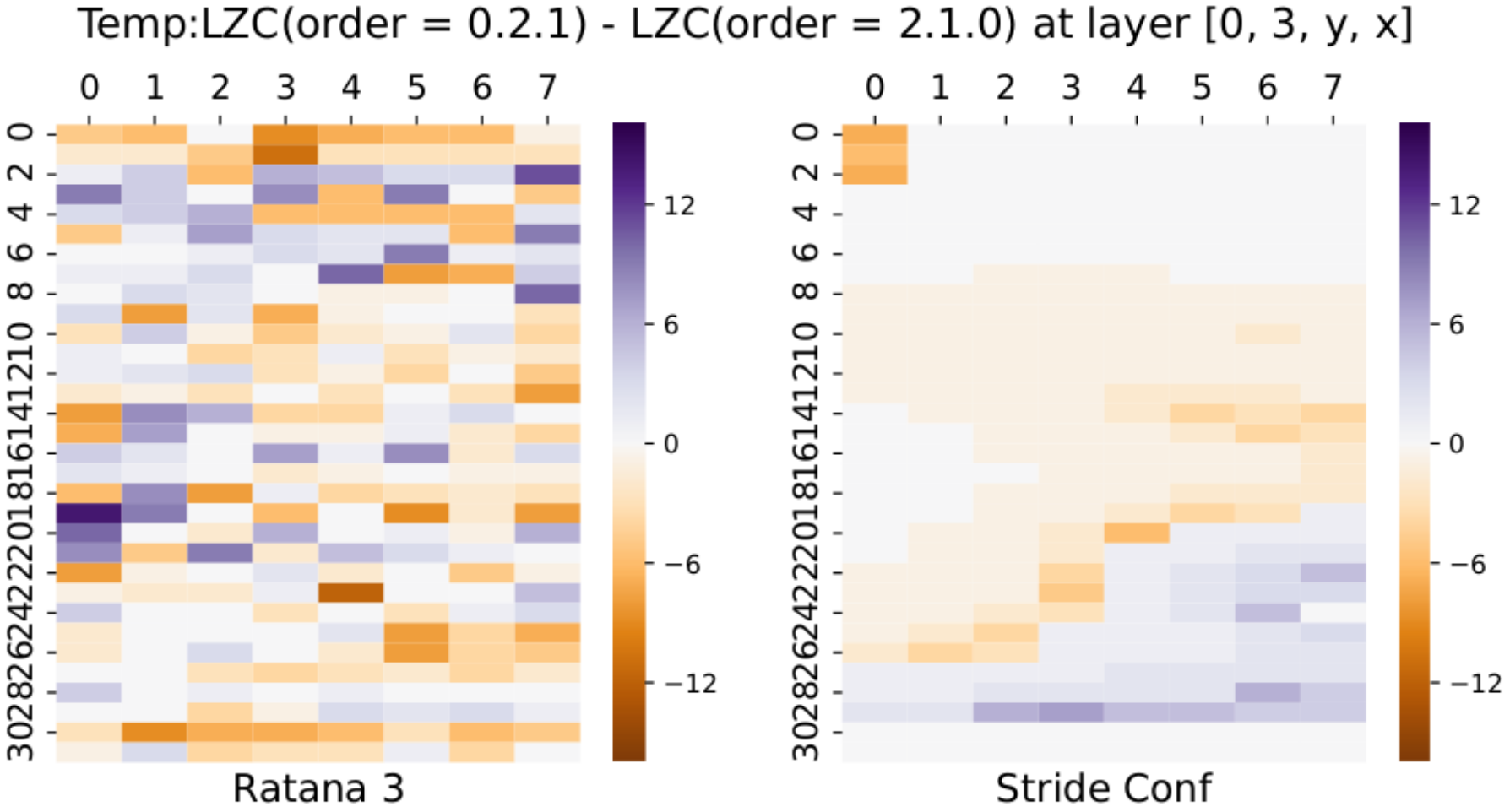
Range Encoding



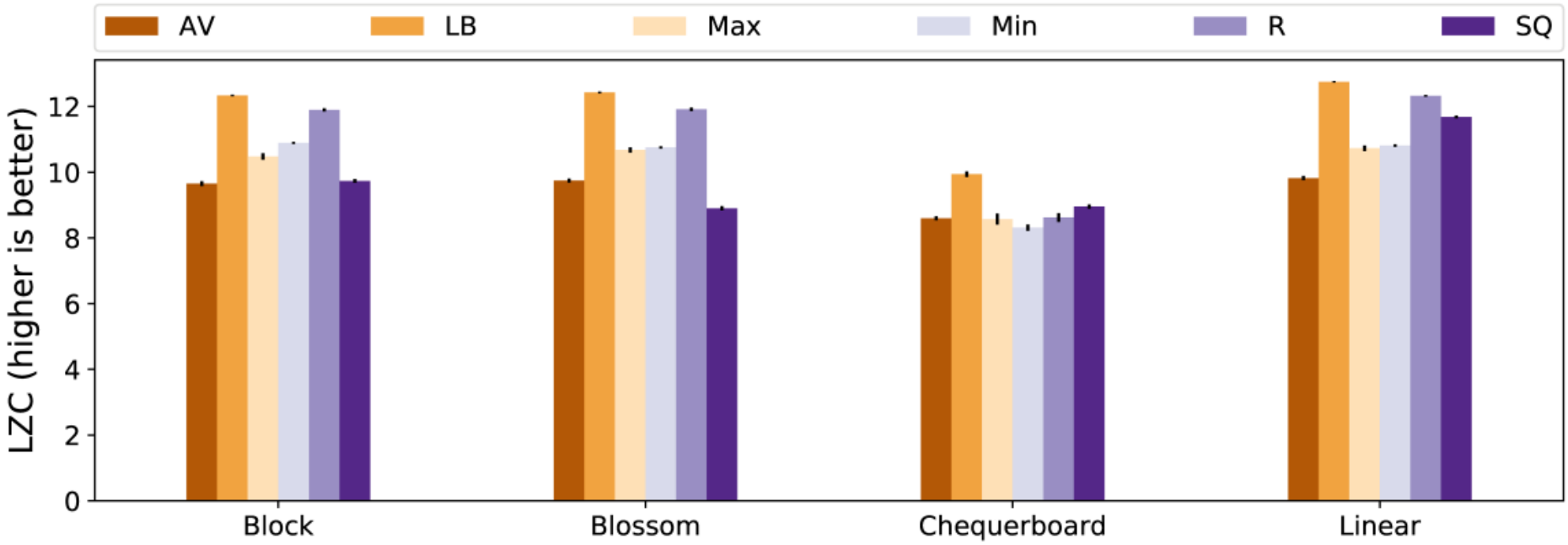
Traversal


Temp:LZC(order = 0.2.1) - LZC(order = 2.1.0) at layer [0, 3, y, x]
0 1 2 3 4 5 6 7
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
12
6
0
−6
−12
Akumuli
Pascal 3

Traversal


Temp:LZC(order = 0.2.1) - LZC(order = 2.1.0) at layer [0, 3, y, x]
0 1 2 3 4 5 6 7
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
12
6
0
−6
−12
Ratana 3
Stride Conf

Traversal



block

5	1	3	10
4	0	2	9
8	6	7	11

blossom

8	1	5	10
4	0	2	9
7	3	6	11

chequerboard

0	6	1	7
8	2	9	3
4	10	5	11

linear

0	1	2	3
4	5	6	7
8	9	10	11

ERA5 Daten

- 10.89 PiB
  - 1440 x 721 Grid (31 km)
  - 137 Level (bis 80 km)
  - Stündliche Daten, seit 01.01.1979
  - 16 Bit Integer (Poor Man Compression: Faktor und Offset)
  - 120 Variables

Karlsruher Institut für Technologie

Ungenutzte Folien

Ungenutzte Folien

Pascal Zip (pzip)

Vorhersagebasiertes Kompressionsverfahren

Methode

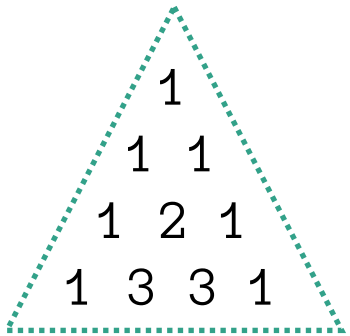
Für jeden einzelnen Datenpunkt wird (basierend auf vorhergehenden Werten) eine Vorhersage gegeben und die Differenz zum wahren Wert (Residuum) gespeichert

Name

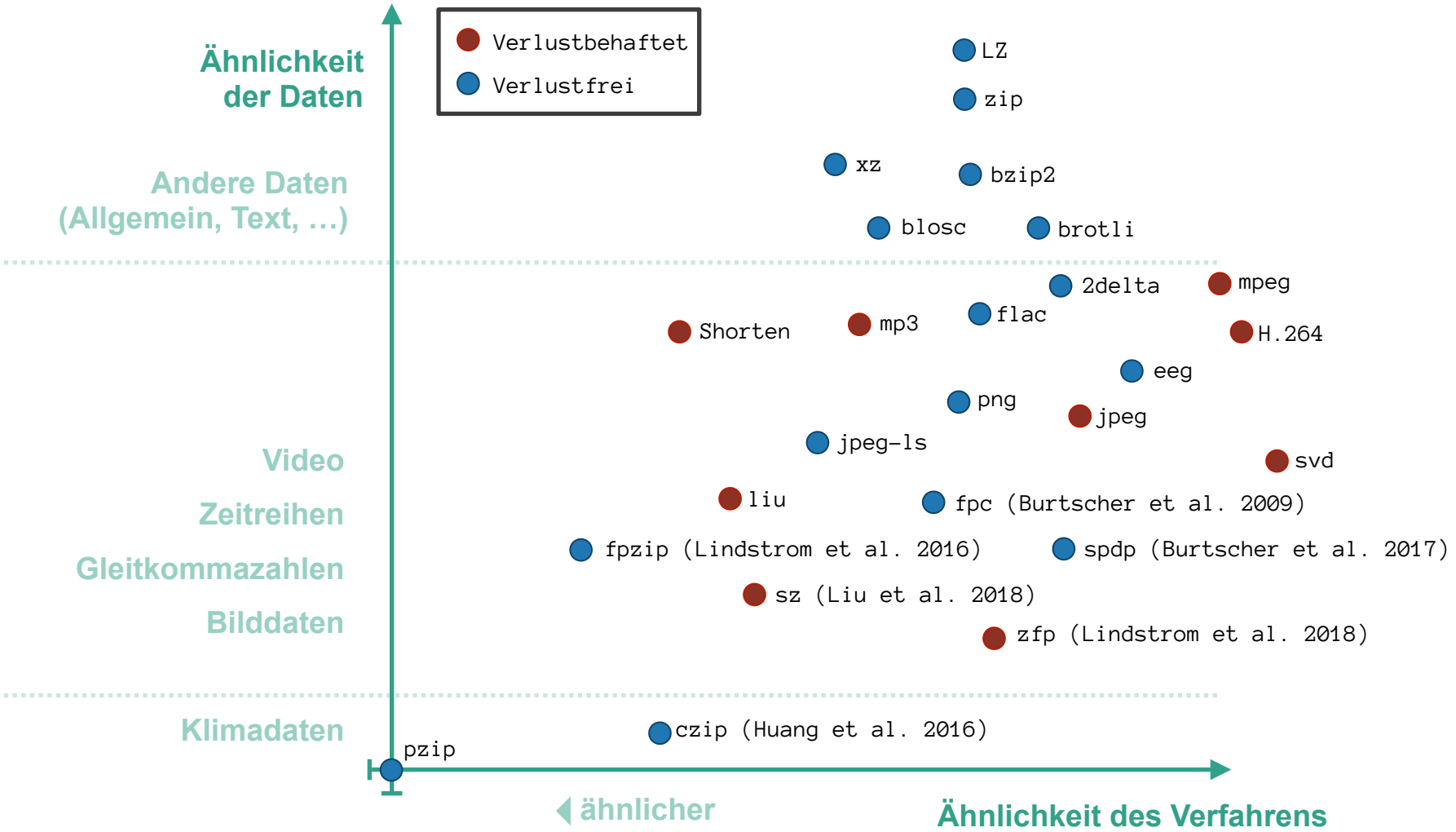
Die Koeffizienten der Vorhersage sind identisch mit dem Zahlen von Pascals Dreieck

1	−3	3	−1
−2	6	−6	2
1	−3	3	?

−1	2	−1
2	−4	2
−1	2	?



Unterscheidung von anderen Kompressionsverfahren



Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik

Wissens-generierung

3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen

Wissens-integration (zur Benutzung in der Dekorrelation)

5. Analyse und Optimierung der Kodierung

Optimierung

Schnelle Prototypen-entwicklung

Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung

1. Identifikation von Abhängigkeiten und Zusammenhängen zwischen und innerhalb der Variablen
   a. Entropie
   b. Varianz
   c. Mutual Information

2. Anwendung von aktuellen Kompressionsmethoden
   a. Kompressionsfaktor
   b. Durchsatz [Bytes/Sek.]

Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung

1. Identifikation von Abhängigkeiten und Zusammenhängen zwischen und innerhalb der Variablen
   a. Entropie
   b. Varianz
   c. Mutual Information

2. Anwendung von aktuellen Kompressionsmethoden
   a. Kompressionsfaktor
   b. Durchsatz [Bytes/Sek.]

Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured [..] EGU 2018

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms. IEEE eScience 2019

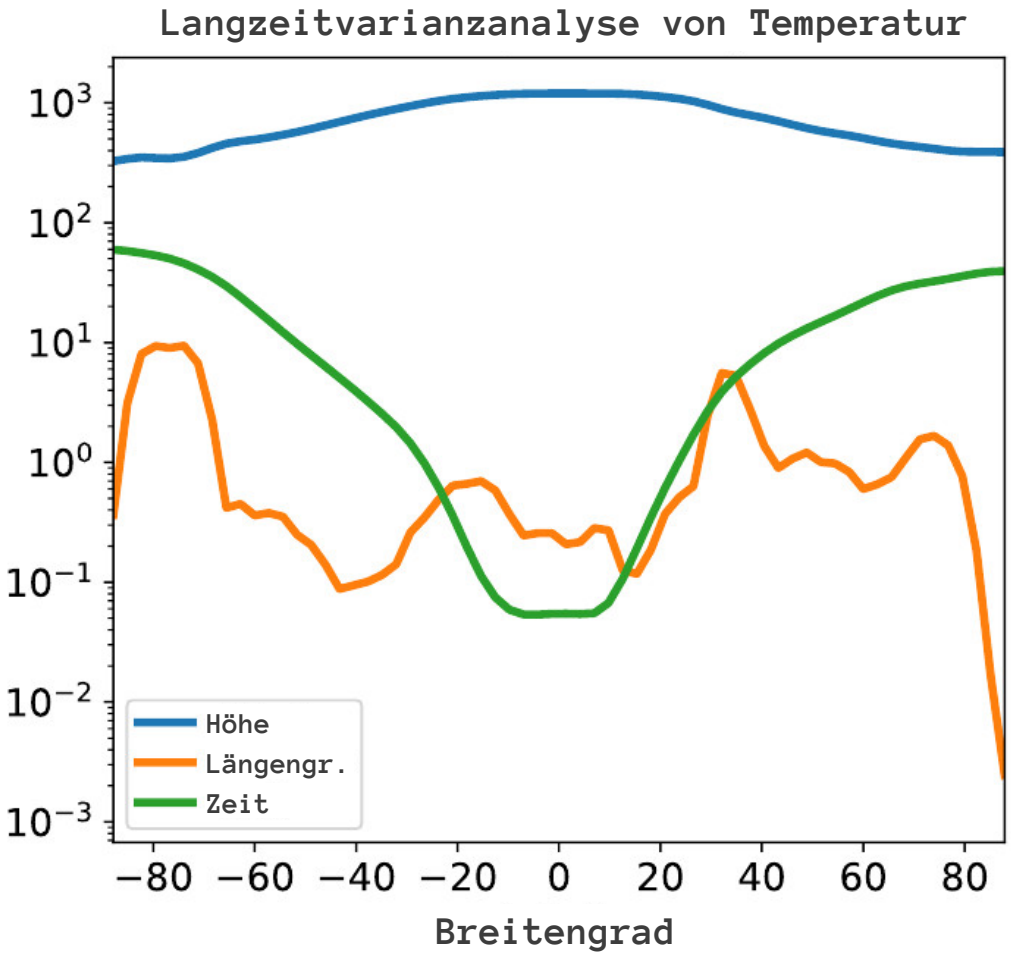
Kerzenmacher et al. (2018). QBO influence on the ozone distribution in the extra-tropical stratosphere. EGU 2018

Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung



Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung

(3) Einbau von Wissen in den Kompressor bzw. Dekompressor
   a. aus den Erfahrungen von (1)
   b. etablierte Indikatoren aus Forschungsfeld

(4) Identifikation von Informationen aus dem zu komprimierenden Datensatz

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using SARIMA
IEEE eScience 2017

Cayoglu et al. (2018c). Concept and Analysis of Information Spaces to improve Prediction-Based Comp.
IEEE Big Data 2018

Cayoglu et al. (2019a). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Entwicklung eines Kompressionsverfahrens

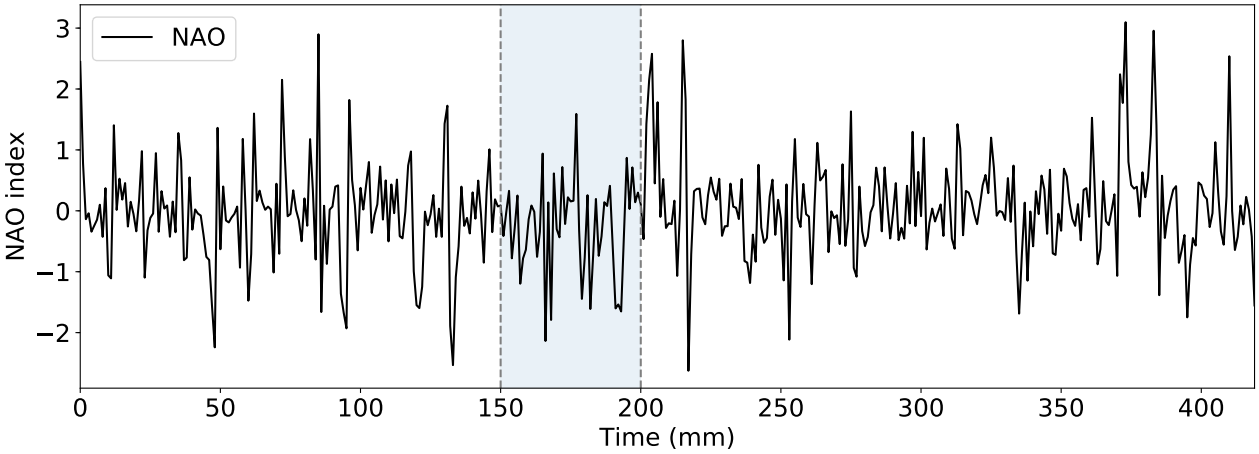
Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung

(3) Einbau von Wissen in den Kompressor bzw. Dekompressor

a. aus den Erfahrungen von (1)
b. etablierte Indikatoren aus Forschungsfeld



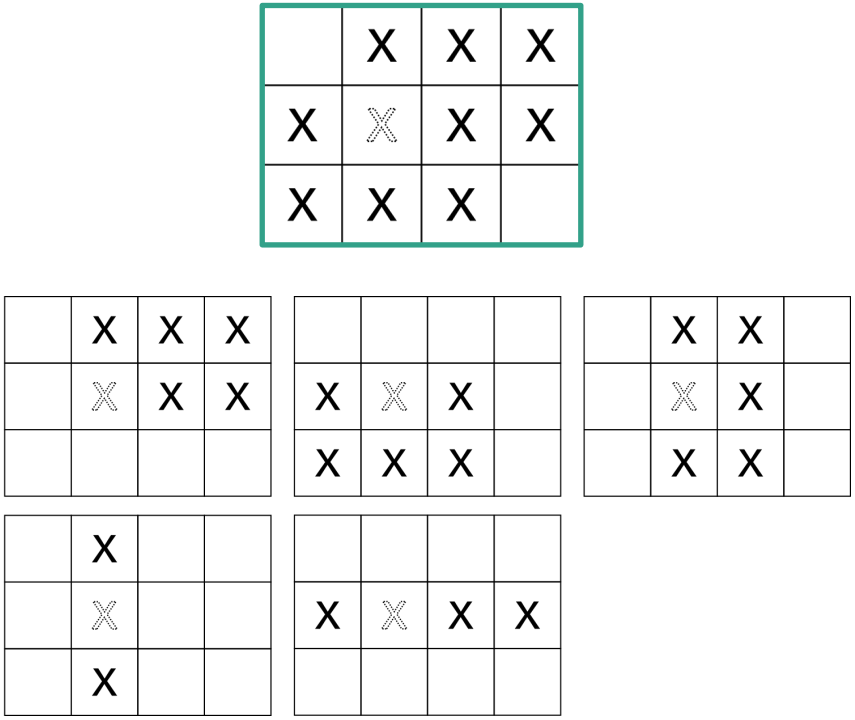
Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung

Schnelle Prototypen-entwicklung

4 Identifikation von Informationen aus dem zu komprimierenden Datensatz



Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung
Schnelle Prototypen-entwicklung

(5) Bewertung des Informationsgehalts in den dekorrelierten Daten und Anpassung der Kodierung

◌ Unterstützung bei der iterativen Entwicklung

a. Ensemble Prediktoren
b. Qualitätssicherheit
c. Parallele Kompression
d. Zufällige Untermengenwahl

Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung
Schnelle Prototypen-entwicklung

5 Bewertung des Informationsgehalts in den dekorrelierten Daten und Anpassung der Kodierung

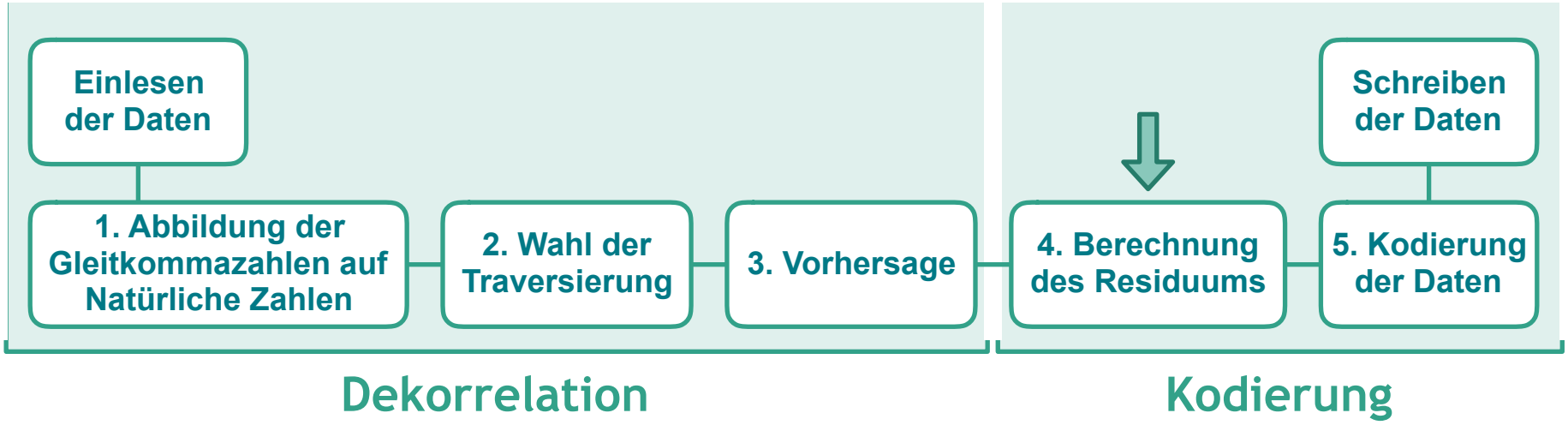
Entwicklung eines Kompressionsverfahrens

Karlsruher Institut für Technologie

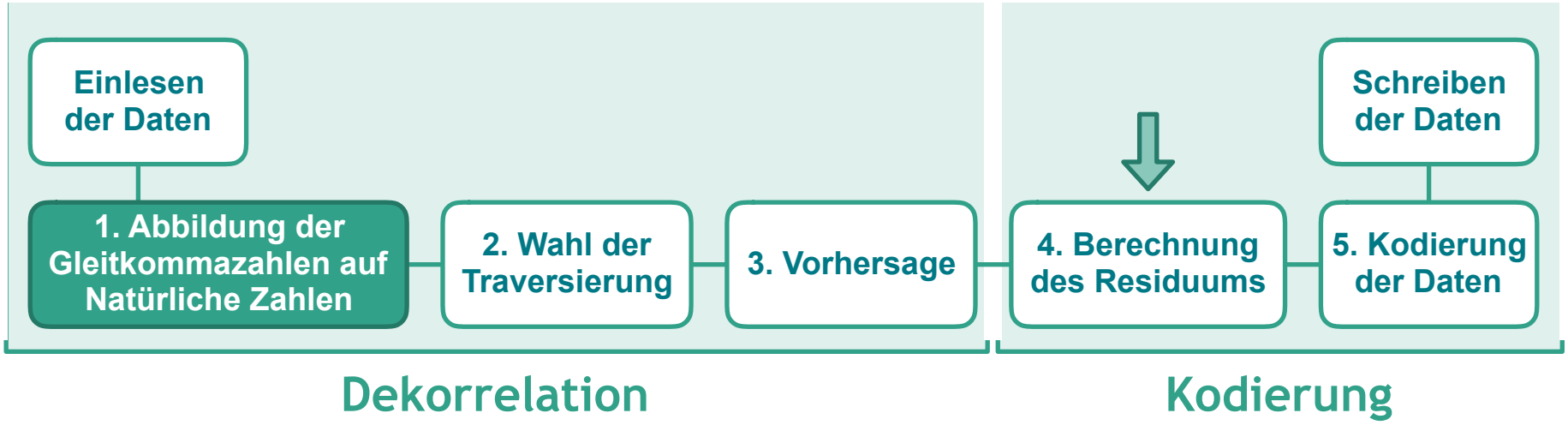
1. Analyse der Dateneigenschaften
2. Analyse des Stand der Technik
3. Integration von vorhandenem Wissen
4. Schaffung und Integration von aktuellem Wissen
5. Analyse und Optimierung der Kodierung
6. Schnelle Prototypen-entwicklung

- Unterstützung bei der iterativen Entwicklung
  a. Ensemble Prediktoren
  b. Qualitätssicherheit
  c. Parallele Kompression
  d. Zufällige Untermengenwahl

Vorhersagebasierende Kompression



Vorhersagebasierende Kompression

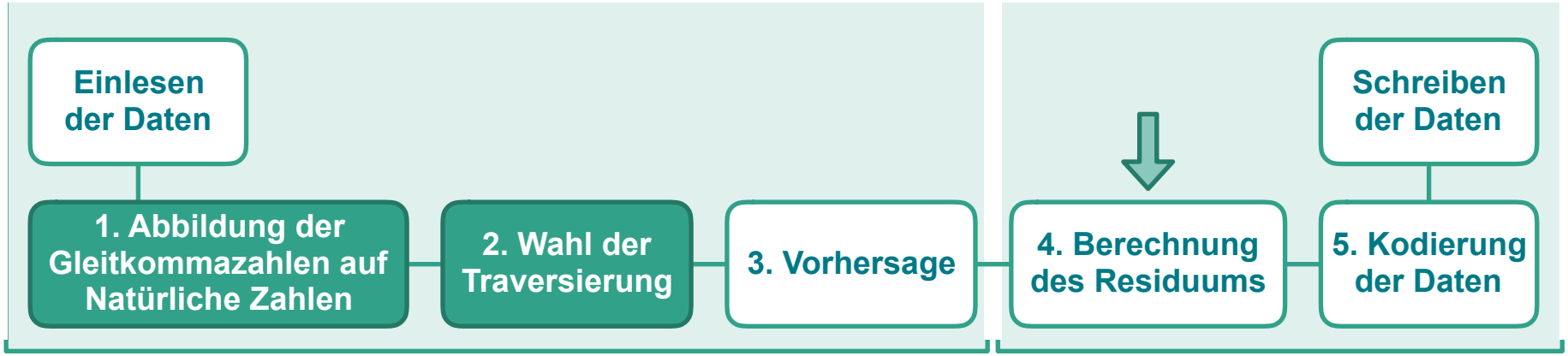


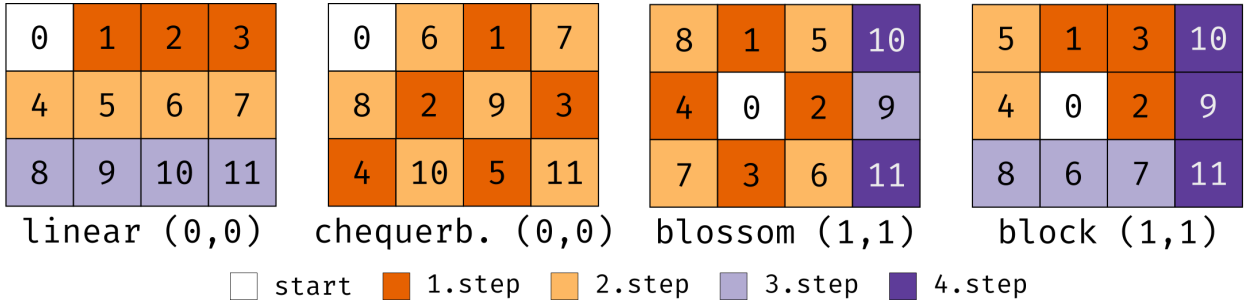
$$f : \mathbb{R} \rightarrow \mathbb{N}$$

```
Zahl 1: 256.321 -> 1132472599
Zahl 2: 255.931 -> 1132457558

Zahl n:  . . .  ->  . . .
```

Vorhersagebasierende Kompression

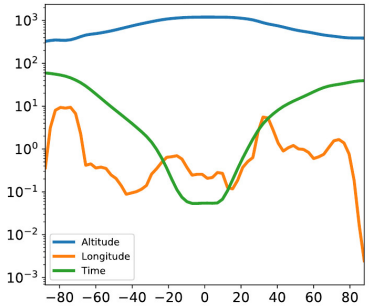


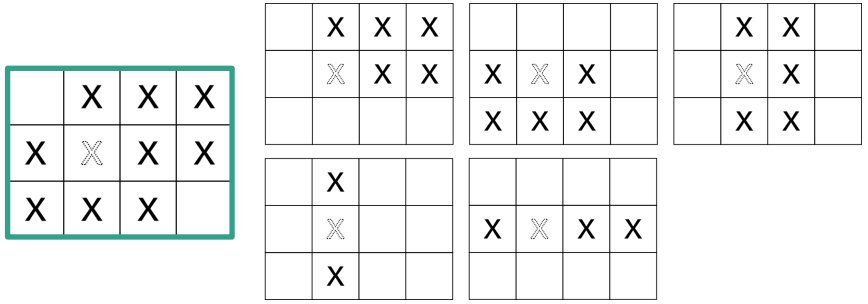


Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**
Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression **GI INFORMATIK 2019**

Vorhersagebasierende Kompression

Karlsruher Institut für Technologie

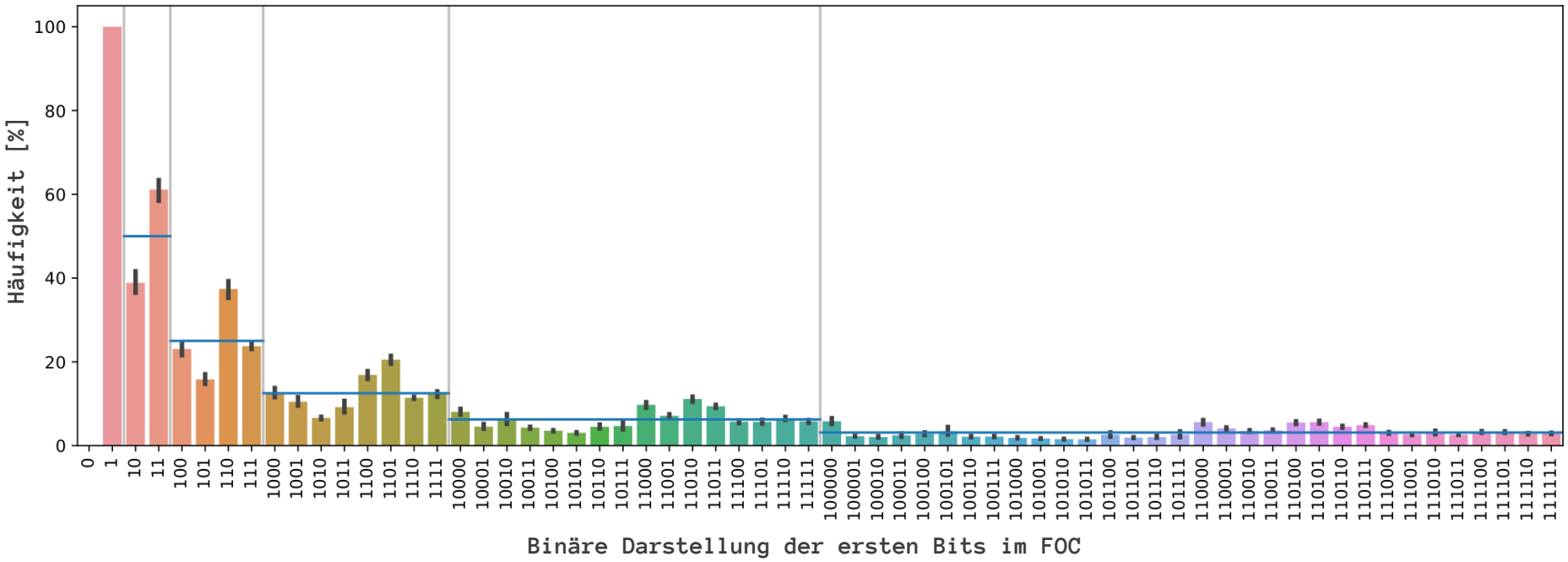




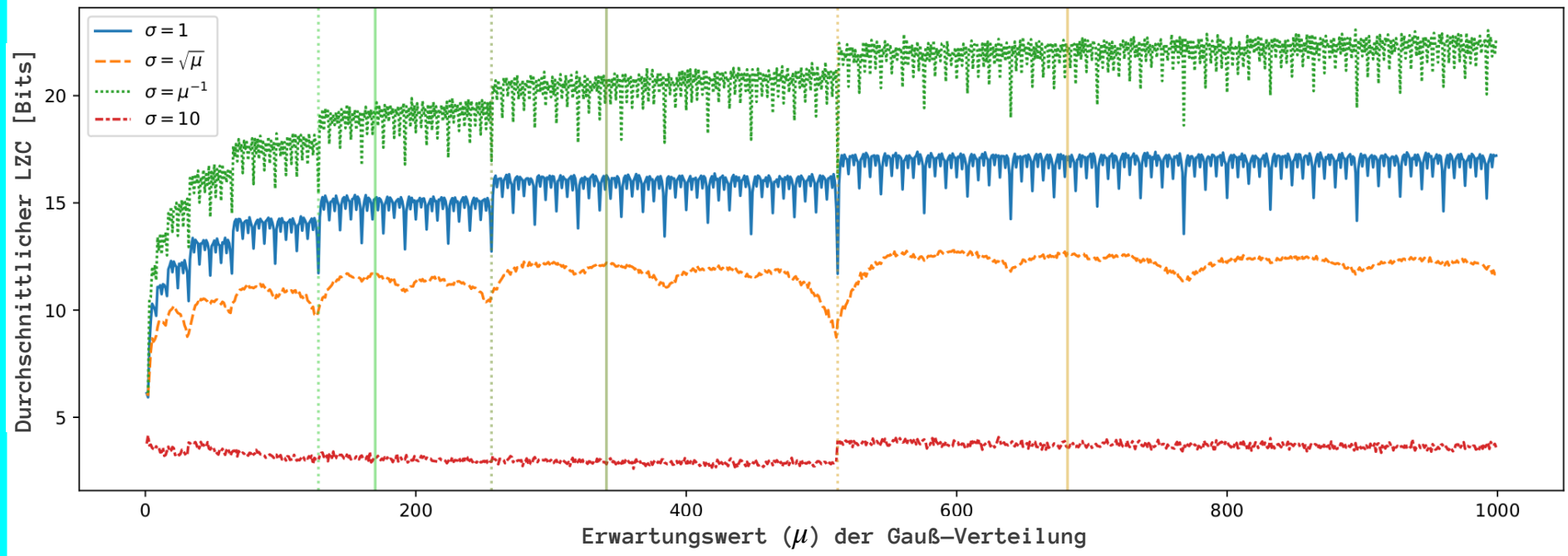
Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output **EGU 2018**
Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using SARIMA **IEEE eScience 2017**
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**

Wie häufig tritt das Phänomen auf?

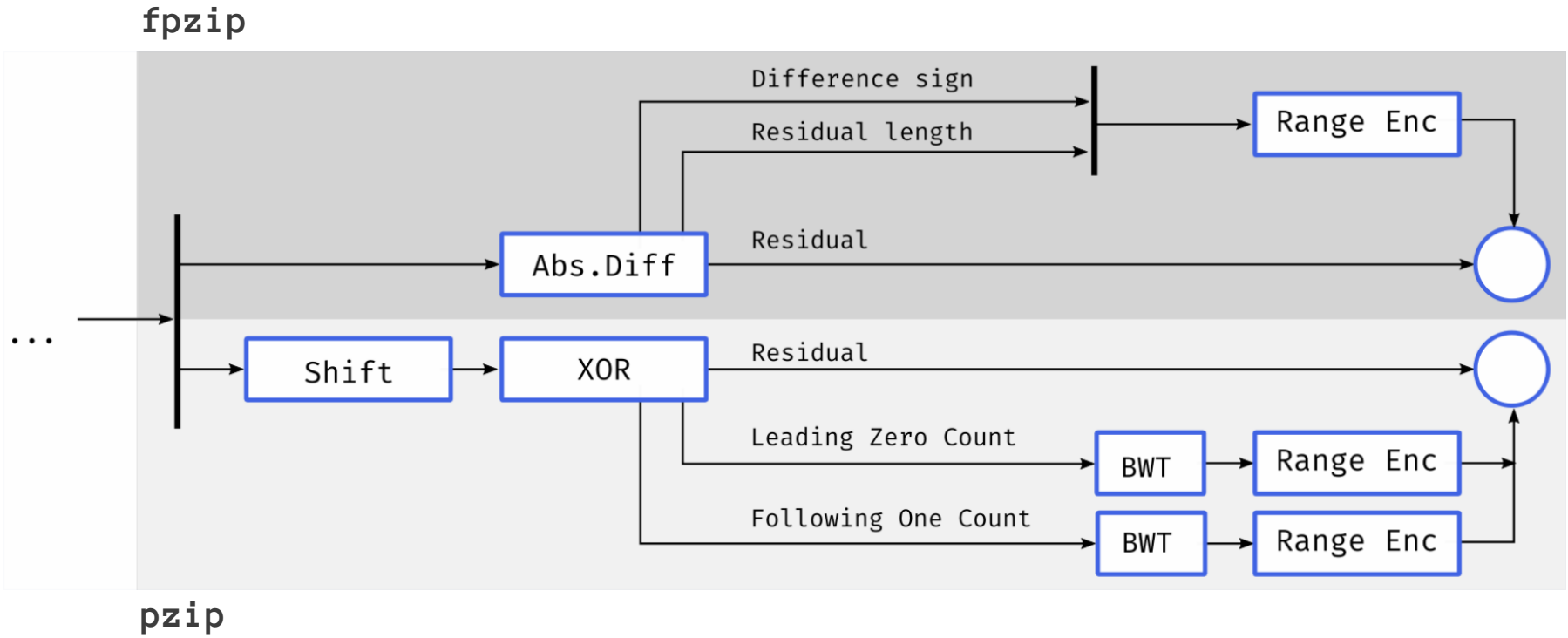
Analyse der ersten sechs Bits



Weitere Testreihen...

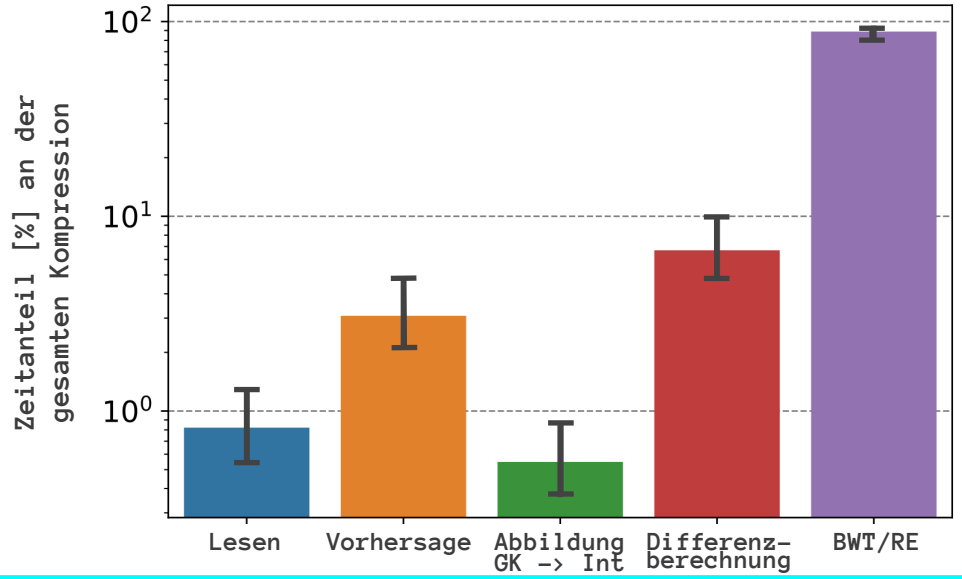


fpzip vs pzip (Kodierung)



Durchsatz und Komplexität

- Engpass in der aktuellen Implementierung ist BWT
- Laufzeit- und Speicherplatzkomplexität von fpzip $\mathcal{O}(n)$
- Laufzeitkomplexität $\mathcal{O}(n + 4 \cdot \tau \cdot n + n)$
- Speicherkomplexität $\mathcal{O}(\tau \cdot \left(1 + \frac{n}{d_3}\left(\frac{1}{d_2}\left(\frac{1}{d_1} + 1\right) + 1\right)\right) + n \log \sigma)$



τ = Nachbarschaft
$n = d_0 d_1 d_2 d_3$
σ = |Alphabet|

Quellen

[1] https://www.britannica.com/science/global-warming/Theoretical-climate-models
[2] D. Tao, S. Di, Z. Chen, and F. Cappello, “Significantly Improving Lossy Compression for
Scientific Data Sets Based on Multidimensional Prediction and Error-Controlled Quantization,”
in 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2017,
pp. 1129–1139.
[3] https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform
[4] https://en.wikipedia.org/wiki/File:Range_encoding.jpg

Karlsruher Institut für Technologie

Professorenrunde

Professorenrunde

Agenda

Karlsruher Institut für Technologie

- **Motivation**
  - Compression of Environmental Data
- **Climate Data**
- **Compression 101**
- **Development of a Compression Method**
  - Contributions
- **Details: Data Coding**

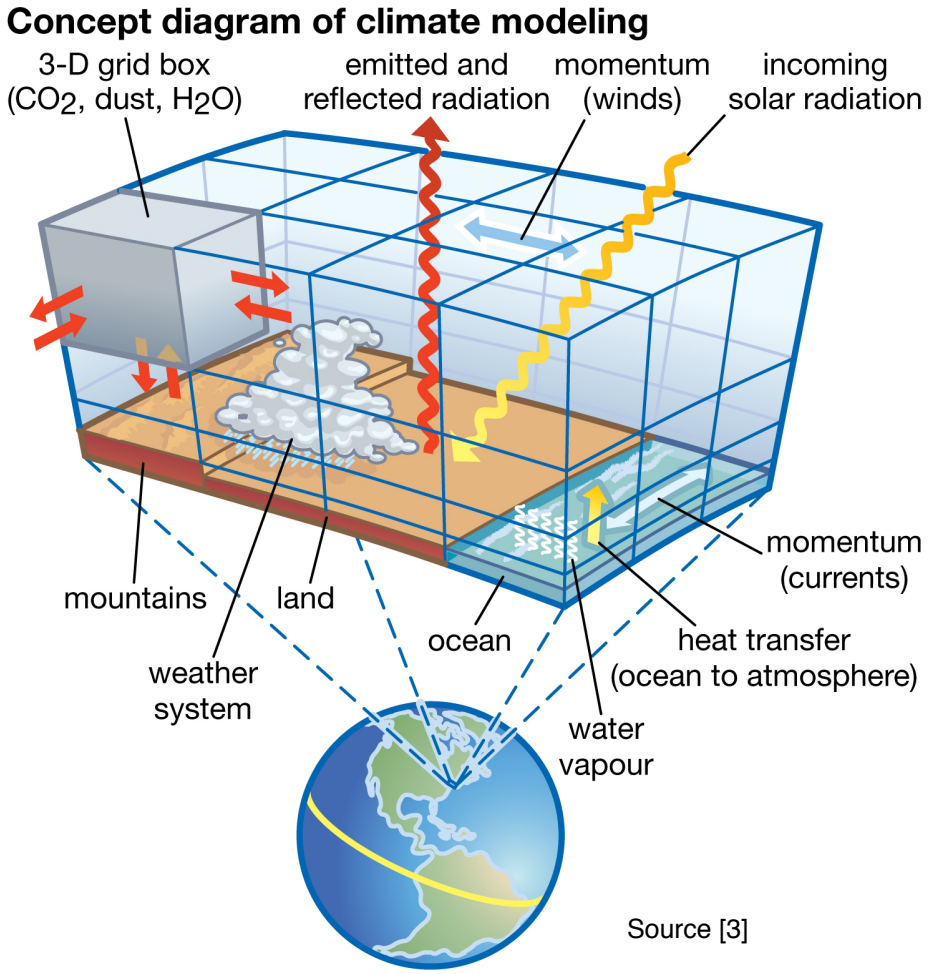
Motivation für die Kompression von Klimadaten

Karlsruher Institut für Technologie

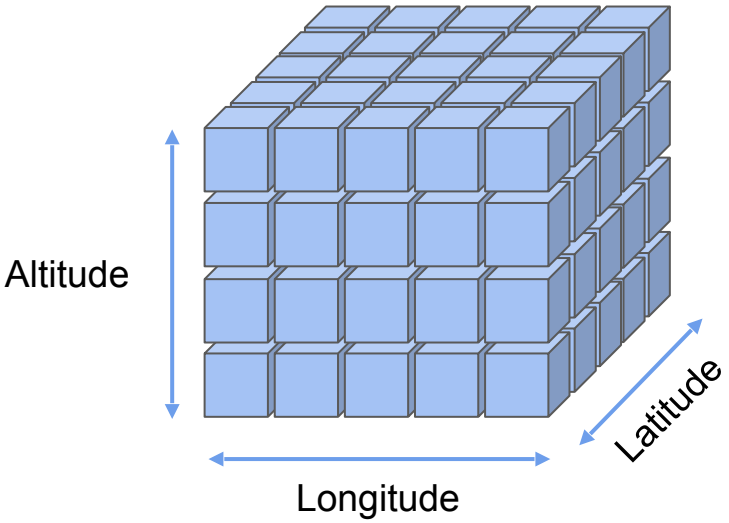
- **Problem**
  Data Volume. The simulation output of high-resolution Earth System Models is too big.
- **Current solution**
  - Reduce temporal scale
  - Limit variable output
- **ERA5**: Current reanalysis dataset used for initialisation and validation of simulations uses for each variable 2.23 TiB p.a. (~50 years, ~120 variables).

ECMWF [1]

Climate Models and Data Structure

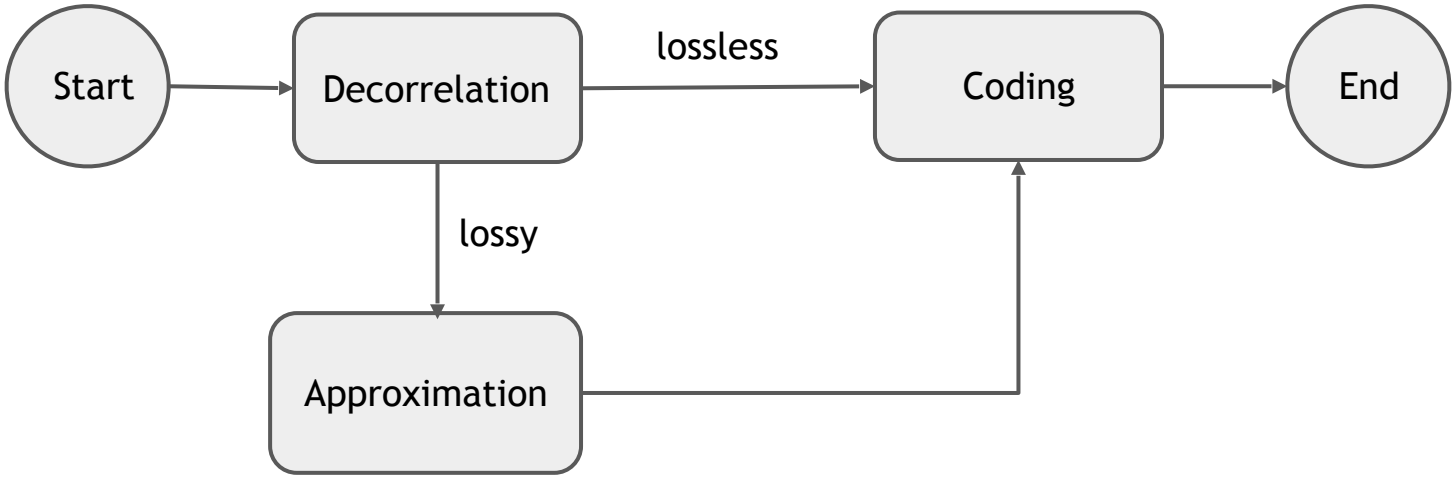


Source [3]



Compression 101

- **Decorrelation**: Removing redundancy from the data e.g. probability models like context-based dictionaries (q > u)
- **Approximation**: Evaluating information content and define a threshold for information loss e.g. pi = 3
- **Coding**: Representing information in a compact form (actual compressing of bytes) e.g. Huffman coding

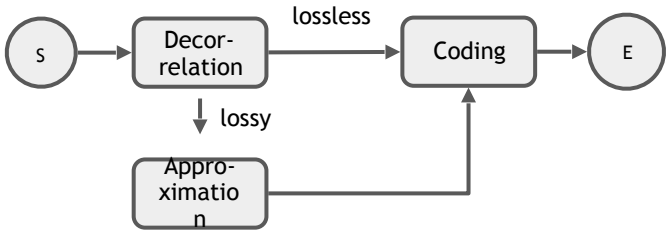


based on [2]

Development of a Compression Method for specific data (e.g. environmental data)

Karlsruher Institut für Technologie

1. **Understanding the structure** and intrinsics of the data
2. Analysing **available compression techniques** for strengths and weaknesses
3. **Integrating existing knowledge** about the interactions of variables
4. **Identifying new patterns and relationships** within and between variables for the current data
5. **Optimizing coding method** to write on disk
6. Building a **framework to perform rapid testing** of new compression algorithms

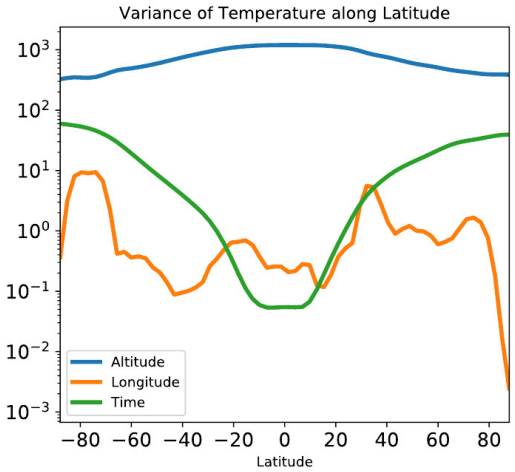


based on [2]

1. & 2. Analysis of redundancy in the data and state-of-the-art compression methods

Karlsruher Institut für Technologie

- **Identification of redundancy** in the data
  a. Variance Analysis
  b. Entropy Analysis
  c. Mutual Information Analysis
- **Application of compression methods** on the data for optimization
  a. Compression Factor/Ratio [w/o dimension]
  b. Throughput [Bytes/sec]



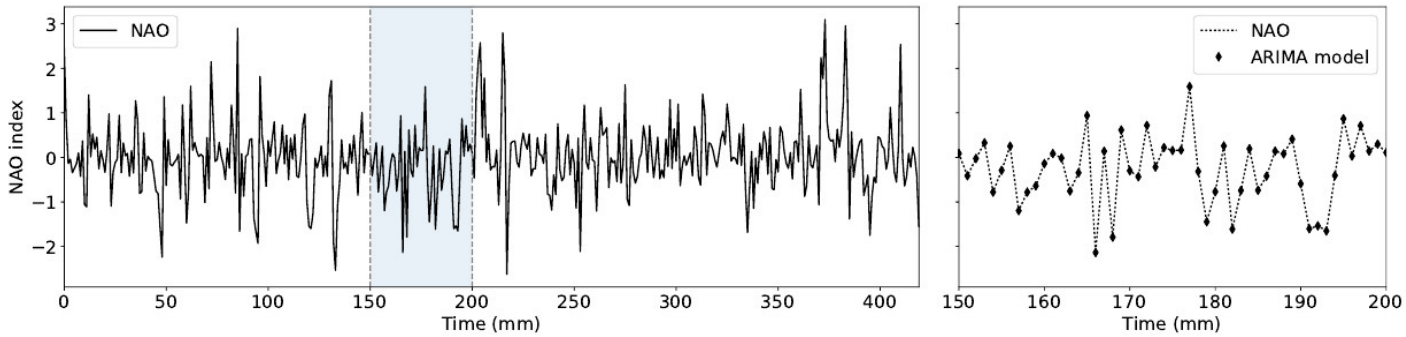
Cayoglu, U., Braesicke, P., Kerzenmacher, T., Meyer, J., and Streit, A. (2018a). Towards an optimised environmental data compression method for structured model output. In EGU General Assembly Conference Abstracts, volume 20, page 8609. https://meetingorganizer.copernicus.org/EGU2018/EGU2018-8609.pdf

Cayoglu, U., Tristram, F., Meyer, J., Schröter, J., Kerzenmacher, T., Braesicke, P., and Streit, A. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms. In IEEE 15th International Conference on e-Science (e-Science). ISBN: 978-1-7281-2451-3, DOI: 10.1109/eScience.2019.00032

3. Integrating existing knowledge about the interactions of variables

Karlsruher Institut für Technologie

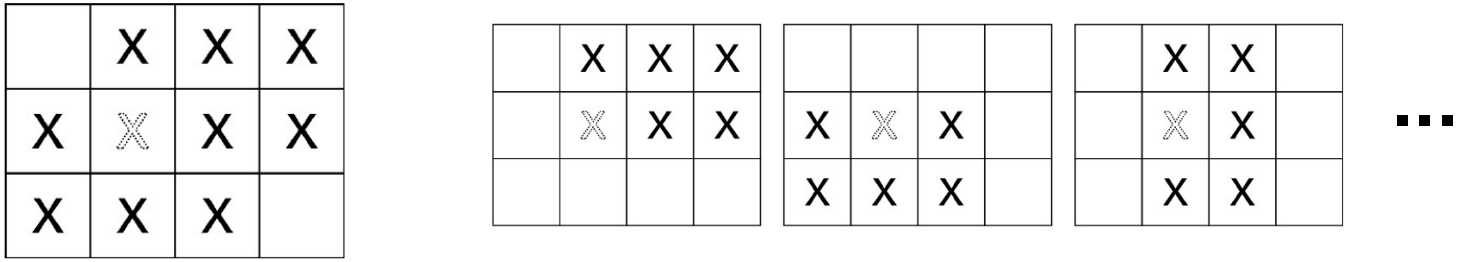
- **Add a-priori information in encoder/decoder** to identify redundant information in dataset.
- **Use of established climate indices** for saving of interconnections between variables (i.e. ENSO34, QBO30/50, and NAO).
- **Development of a novel compression algorithm** to save these time-series indices using minimal space requirements using statistical compression methods.



Cayoglu, U., Braesicke, P., Kerzenmacher, T., Meyer, J., and Streit, A. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models. In 2017 IEEE 13th International Conference on e-Science (e-Science), pages 315–324. DOI: 10.1109/eScience.2017.45. [best paper award]

4. Identifying new patterns and relationships within and between variables for the current data

- **Identify information** in current dataset.
- **Use information from several dimensions** to predict the development of the variable using neighbouring values.



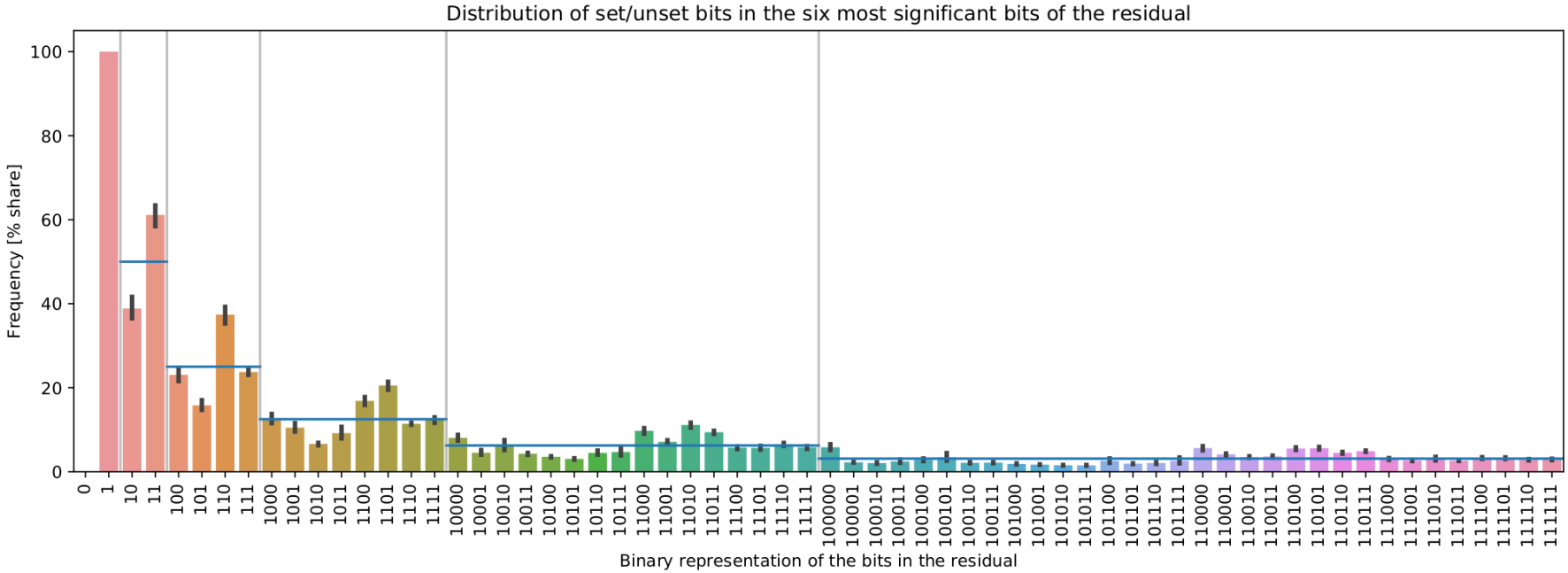
Cayoglu, U., Tristram, F., Meyer, J., Kerzenmacher, T., Braesicke, P., and Streit, A. (2019a). On Advancement of Information Spaces to Improve Prediction-Based Compression. In David, K., Geihs, K., Lange, M., and Stumme, G., editors, INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft, pages 271–272, Bonn. Gesellschaft für Informatik e.V., ISBN: 978-3-88579-688-6, ISSN: 1617-5468, DOI: 10.18420/inf2019_39

Cayoglu, U., Tristram, F., Meyer, J., Kerzenmacher, T., Braesicke, P., and Streit, A. (2018c). Concept and Analysis of Information Spaces to improve Prediction-Based Compression. In 2018 IEEE International Conference on Big Data (Big Data), pages 3392–3401. DOI: 10.1109/BigData.2018.8622313

5. Optimizing coding method to write on disk

Karlsruher Institut für Technologie

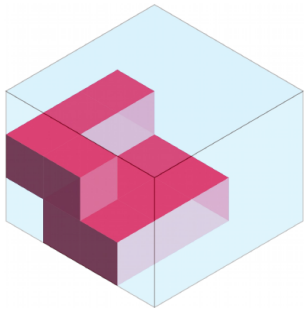
- **Identify information** in current dataset.
- **Development of a novel coding algorithm** to save the data in a more compact form on disk.



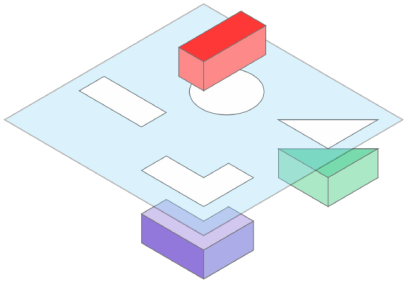
Cayoglu, U., Tristram, F., Meyer, J., Schröter, J., Kerzenmacher, T., Braesicke, P., and Streit, A. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms. In IEEE 15th International Conference on e-Science (e-Science). ISBN: 978-1-7281-2451-3, DOI: 10.1109/eScience.2019.00032

6. Building a framework to perform rapid testing of new compression algorithms

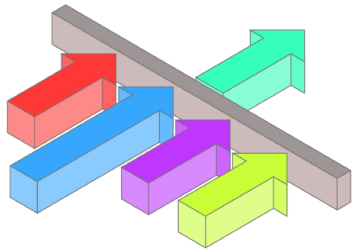
- **Framework** for the development of custom compression methods.
- **Strict interface** using objects and modifier
- **Support for additional modules:**
  a. Ensemble predictors
  b. Quality assessment
  c. Parallel compression
  d. Random subsetting



Random subsetting



Strict interfaces



Ensemble predictors

Cayoglu, U., Schröter, J., Meyer, J., Streit, A., and Braesicke, P. (2018b). A Modular Software Framework for Compression of Structured Climate Data. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '18, pages 556–559, New York, NY, USA. ACM, ISBN: 978-1-4503-5889-7, DOI: 10.1145/3274895.3274897

Details: Data Coding

Details

Principles of Prediction-based Compression of Floating-Point Data.

Karlsruher Institut für Technologie

By using knowledge about the data like its **structure and variability** one can get an estimation for each value. The **better the prediction, the less storage space** is needed for saving the values.

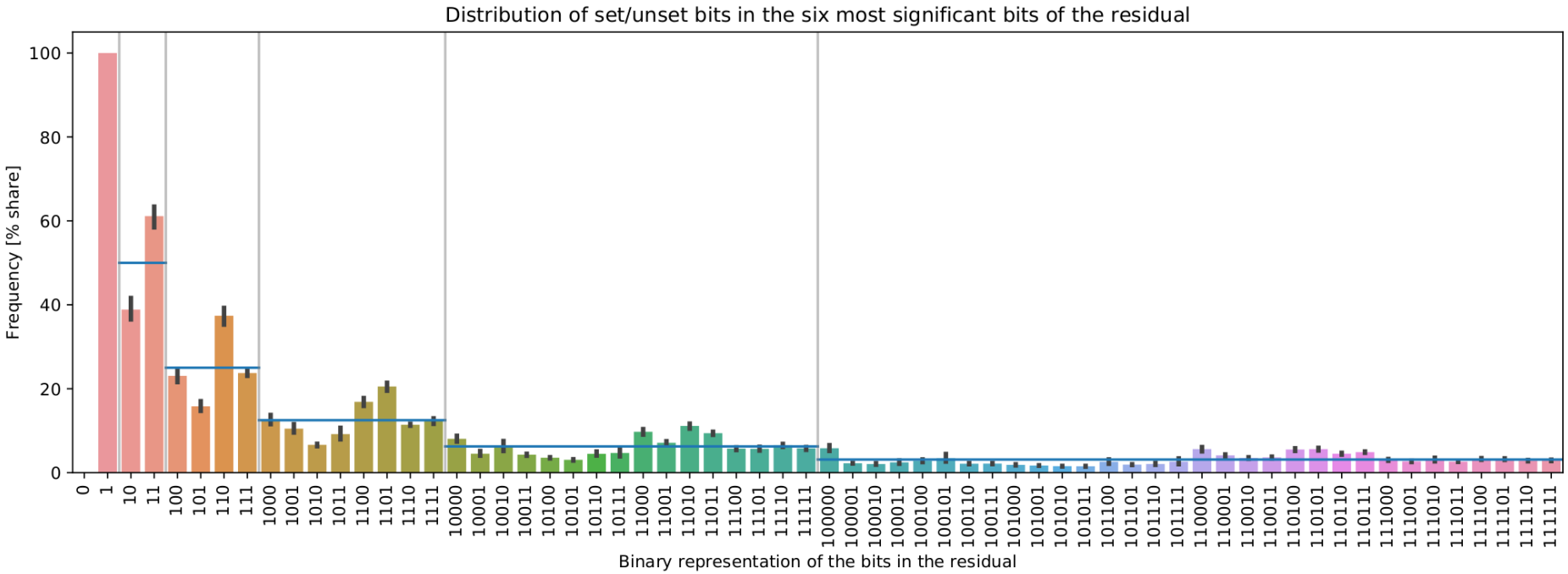
Principles of Prediction-based Compression of Floating-Point Data.

Karlsruher Institut für Technologie

By using knowledge about the data like its **structure and variability** one can get an estimation for each value. The **better the prediction, the less storage space** is needed for saving the values.

```
11110111111000101011011 1  // prediction
11110111111001001101110 1  // truth
00000000000001100110101 0  // prediction^truth
             **0========  // LZC = 13, FOC = 2
```

Currently our focus is on the last step of the chain: encoding of the data.

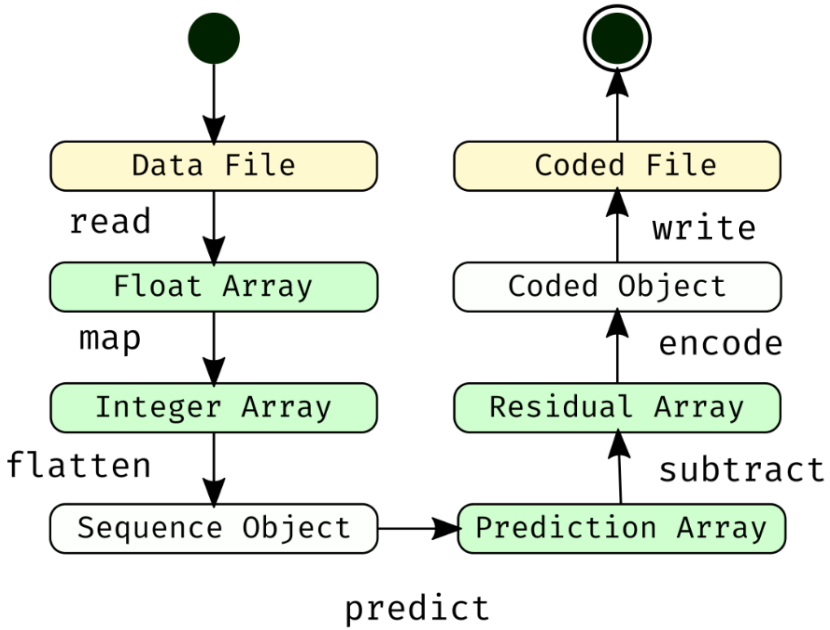


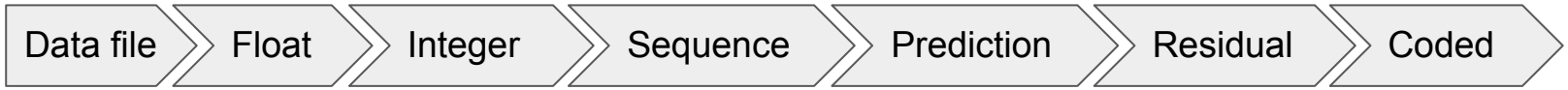

```
bin(p=256.321) = 01000011100000000010100100010111
bin(t=255.931) = 01000011011111111111011001010110
         p ⊕ t = 00000000111111111110001110100001
index 1  3  5  7  9  11 13 15 17 19 21 23 25 27 29 31
```

Steps in developing a compression algorithm for gridded floating point data.

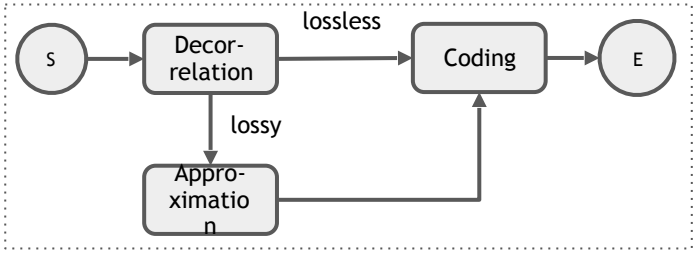
Karlsruher Institut für Technologie

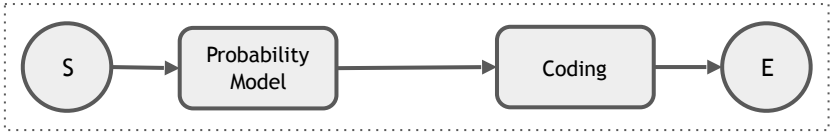
1. Reading the data
2. Mapping floating-point data to integer data
3. Defining traversal sequence
4. Predicting future values
5. Calculating the residual
6. Coding of the residuals
7. Write data on disk

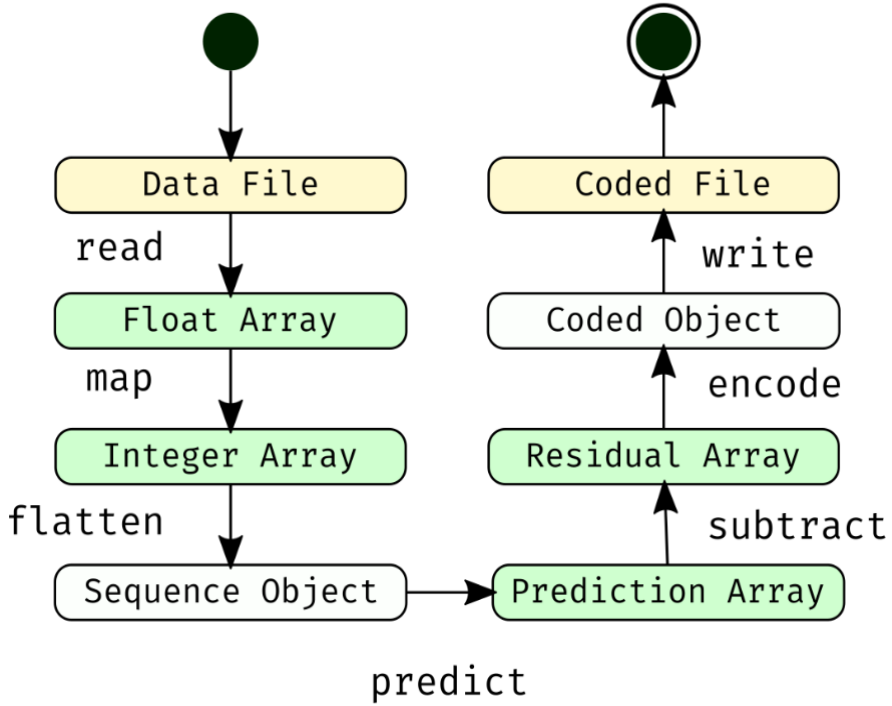




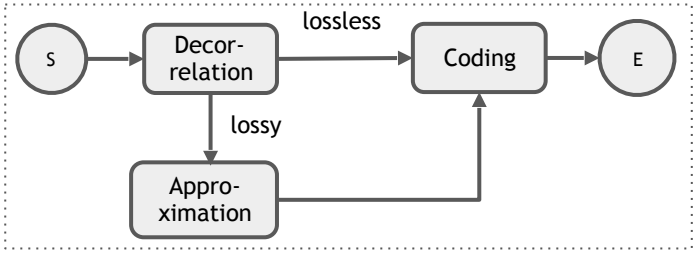
The last three steps are influenced by the encoding method.

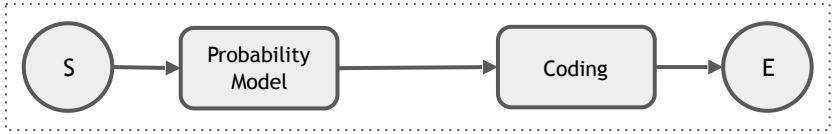






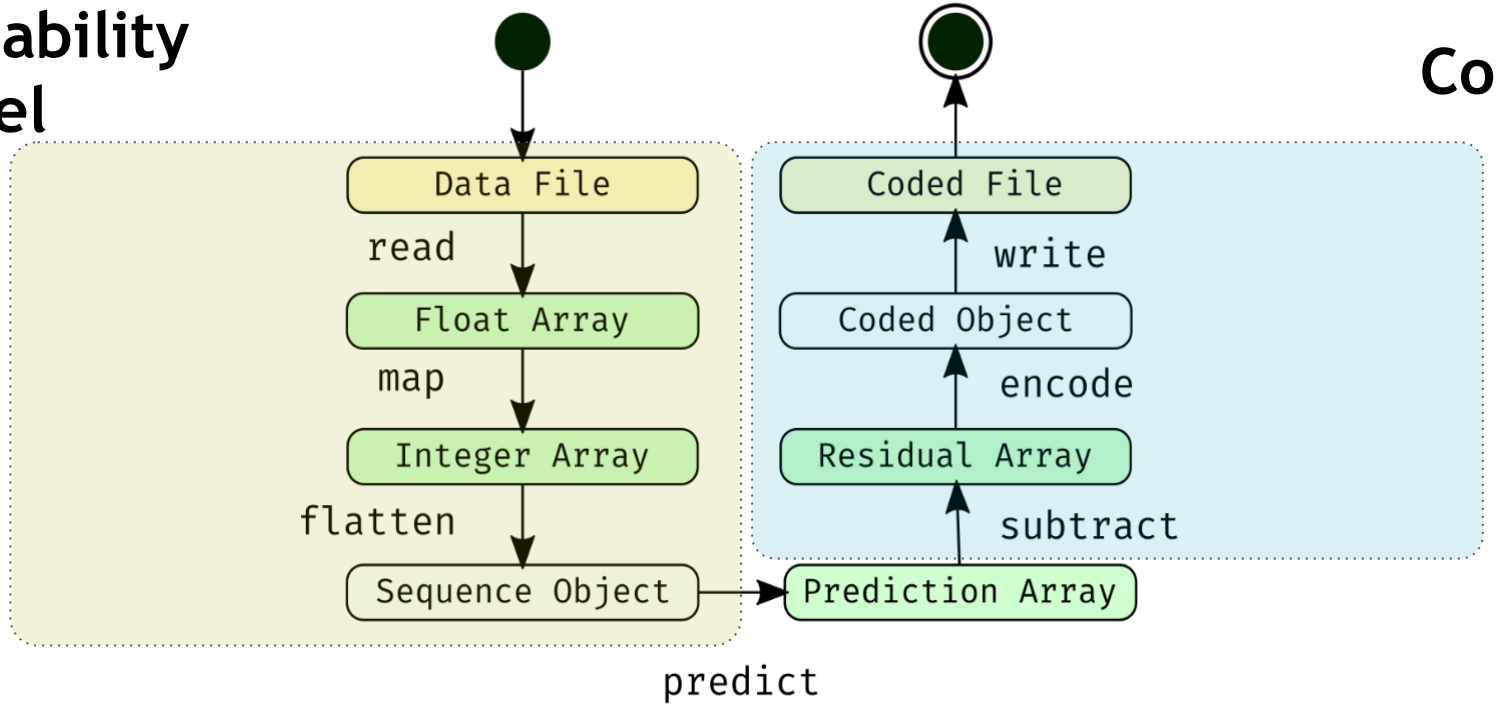
The last three steps are influenced by the encoding method.





Probability model

Coding



How to calculate the difference between two floating-point values

Karlsruher Institut für Technologie

$$\text{diff}_{xor}(p,t) = p \oplus t$$
$$\text{diff}_{abs}(p,t) = \mid p - t \mid$$

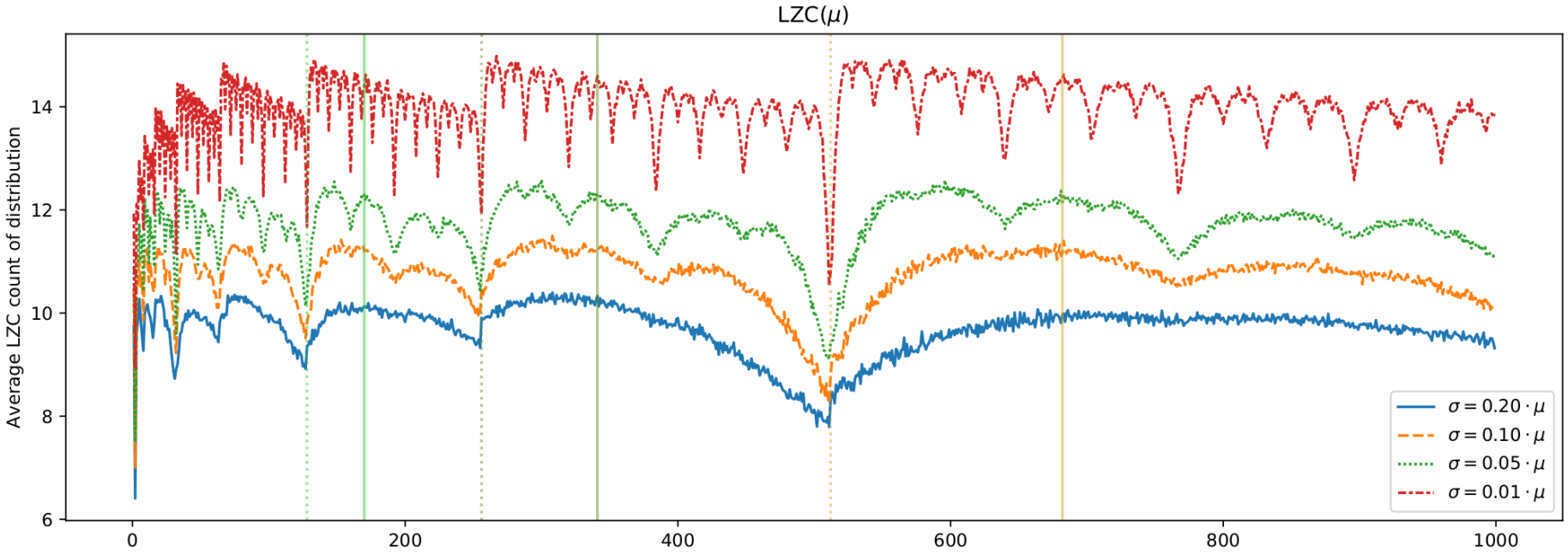
pred: 256.321
true: 255.931
diff: 0.390
diff (xor): 16,762,689

pred: 847,390.837
true: 847,794.417
diff: 403.580
diff (xor): 6,458

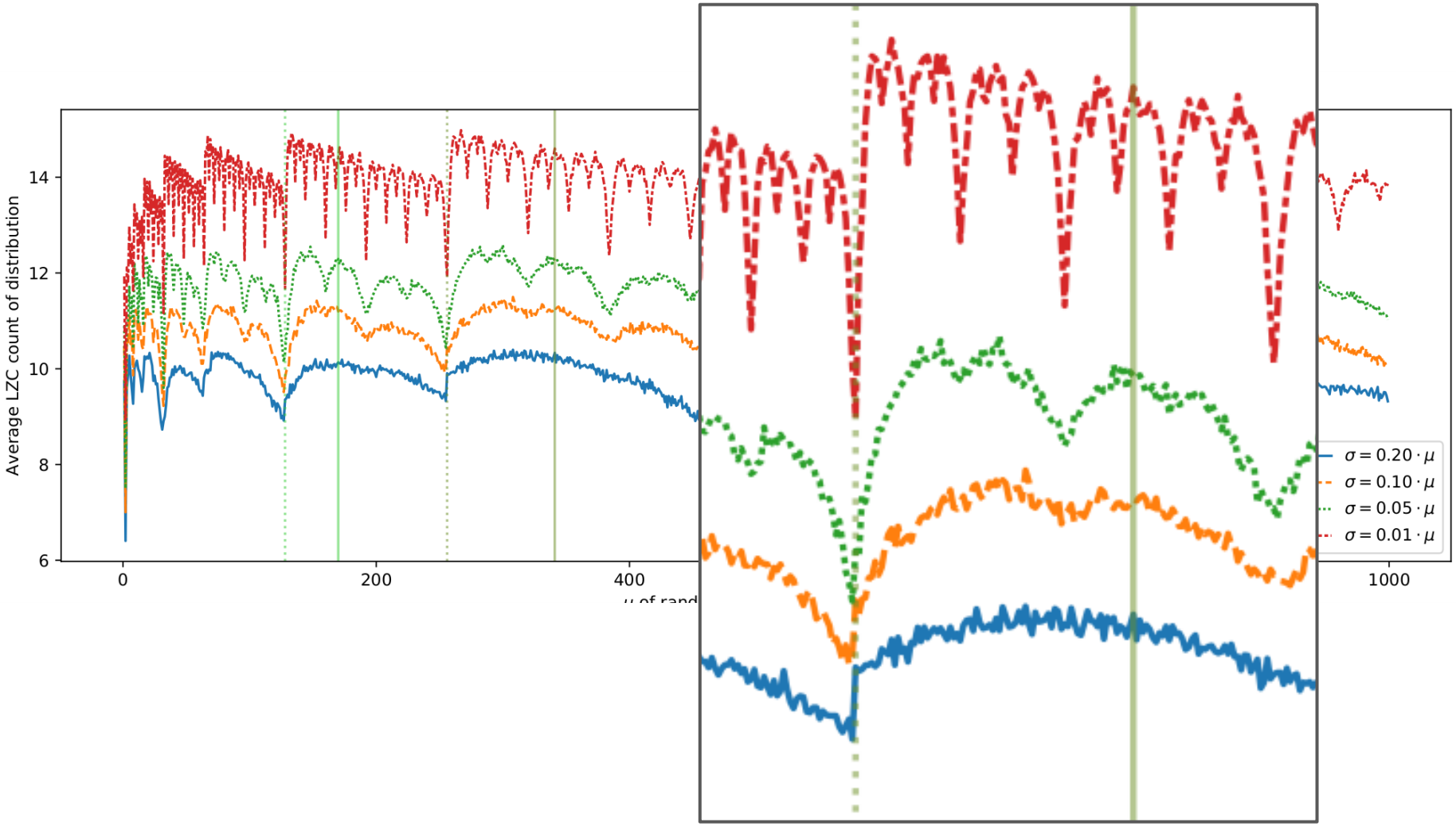
```
bin(p=256.321) = 01000011100000000010100100010111
bin(t=255.931) = 01000011011111111110111001010110
         p ⊕ t = 00000000111111111000111010000001
index 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
```

LZC **FOC**

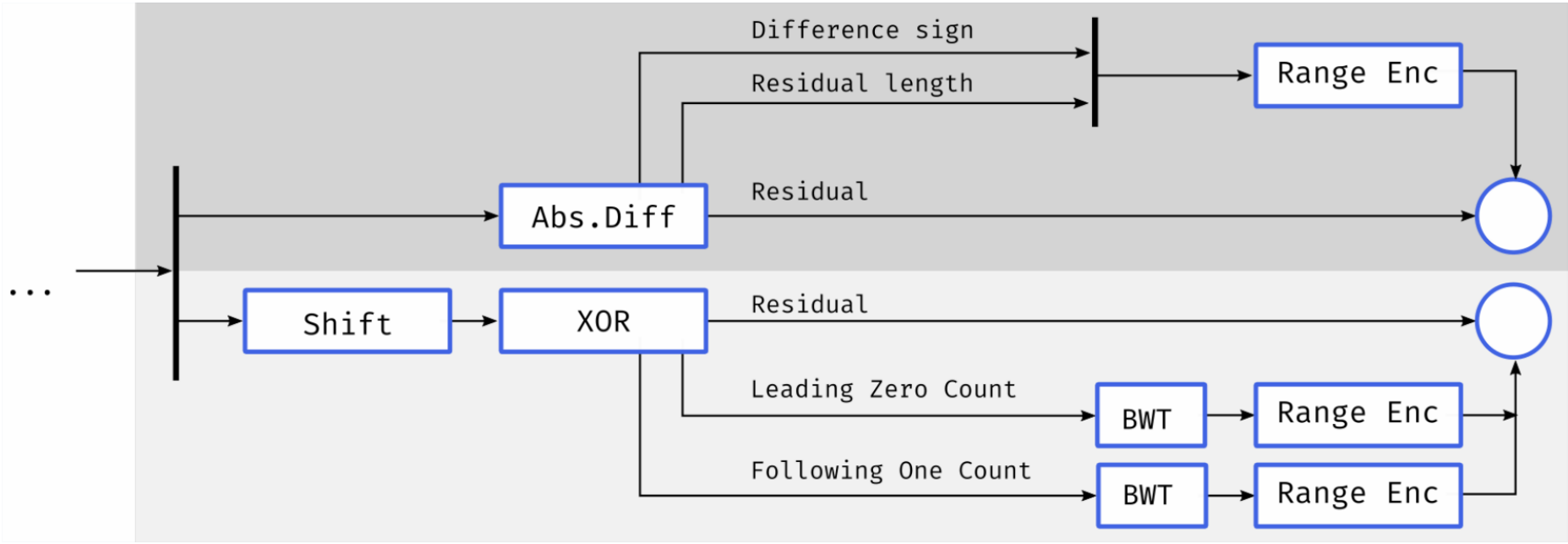
Average LZC for different normal distributed random datasets



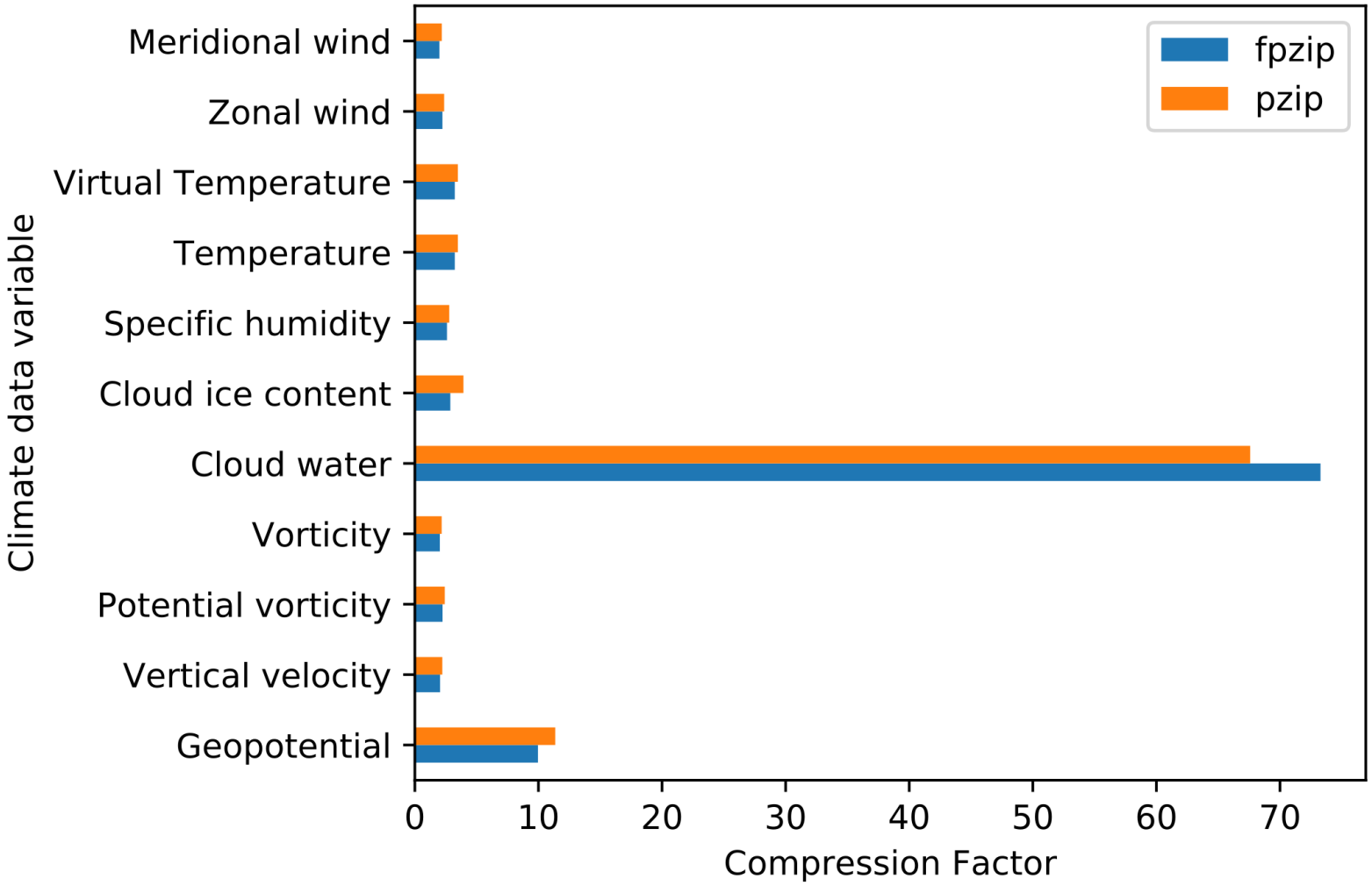
Average LZC for different normal distributed random datasets



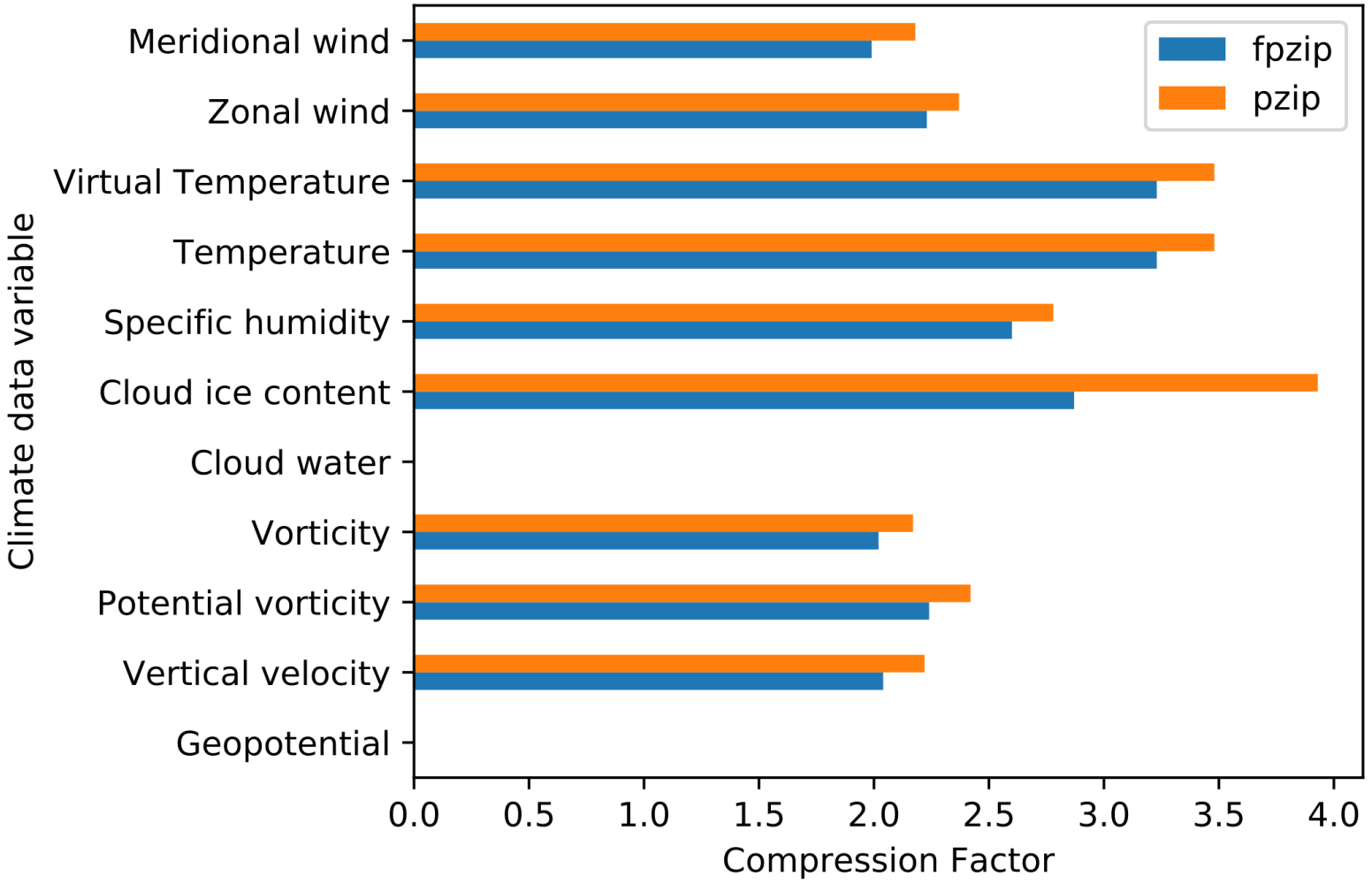
Final scheme of compression method



Final results of compression factor using our proposed approach (1)



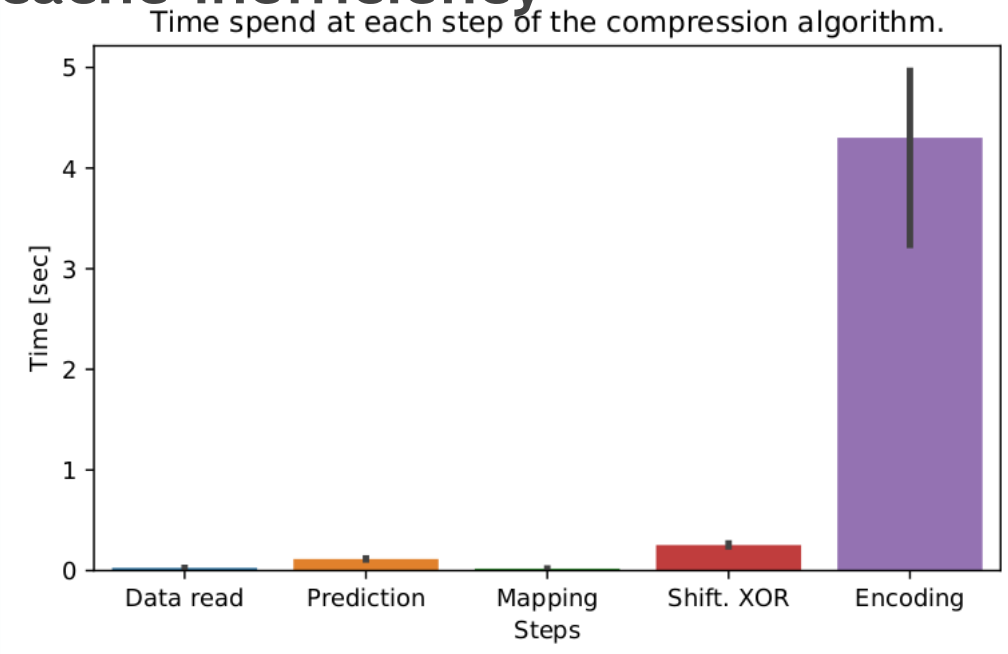
Final results of compression factor using our proposed approach (2)



Pitfalls and future work regarding our proposed approach

Karlsruher Institut für Technologie

- 5x slower than fpzip
- Memory space needed by BWT
- Big-O complexity of fpzip and pzip is same > O(n)
- Might be due to L1-L3 cache inefficiency



Karlsruher Institut für Technologie

Contributions & Publications

Contributions

Publications

Contributions

Karlsruher Institut für Technologie

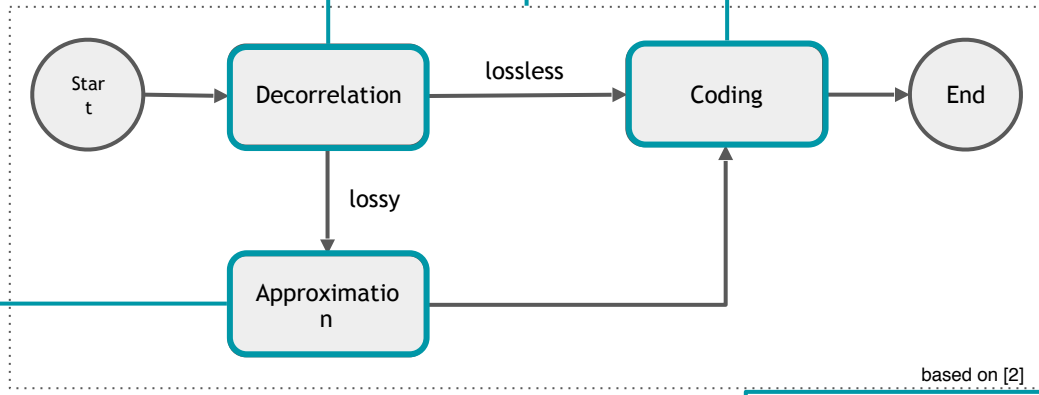
1. **Analysis methods** for identification of redundant information in data
2. **Comparison of lossless compression algorithms** for structured floating-point data
3. **Novel data coding scheme** for prediction-based lossless compression methods
4. Development of a **novel lossy compression** algorithm for time-series data
5. Introduction of **Information Spaces to use information across all dimensions** for data prediction
6. A **modular framework for testing** and quality assessment of compression algorithms

Publications

Karlsruher Institut für Technologie

Data Encoding in Lossless Prediction-Based Compression Algorithms (IEEE eScience 2019)

Concept and Analysis of Information Spaces to improve Prediction-Based Compression (IEEE Big Data 2018)



Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models (IEEE eScience 2017)

Towards an optimised environmental data compression method for structured model output (EGU 2018)

A Modular Software Framework for Compression of Structured Climate Data (ACM SIGSPATIAL 2018)

On Advancement of Information Spaces to Improve Prediction-Based Compression (GI INFORMATIK 2019)

QBO influence on the ozone distribution in the extra-tropical stratosphere (EGU 2018)

Open Source Repositories

Karlsruher Institut für Technologie

- **Lossy compression of time-series data**
  - https://github.com/ucyo/adaptive-lossy-compression
- **Identification of new patterns (Information Spaces)**
  - https://github.com/ucyo/informationspaces
- **Framework for testing**
  - https://github.com/ucyo/cframework
- **Data Coding**
  - https://github.com/ucyo/xor-and-residual-calculation
- **Data Analysis**
  - https://github.com/ucyo/climate-data-analysis

Karlsruher Institut für Technologie

Thank you!

Cayoglu@kit.edu

Questions!?

Sources

[1] https://www.wekeo.eu/themes/charabimba/images/logo-ecmwf.png
[2] D. Tao, S. Di, Z. Chen, and F. Cappello, “Significantly Improving Lossy Compression for Scientific Data Sets Based on Multidimensional Prediction and Error-Controlled Quantization,” in 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2017, pp. 1129–1139.
[3] https://www.britannica.com/science/global-warming/Theoretical-climate-models

Karlsruher Institut für Technologie

1. & 2. Analysis of redundancy in the data and state-of-the-art compression methods

Steps 1 & 2

Karlsruher Institut für Technologie

Earth's atmosphere is a chaotic system. The goal of these analyses are to gain insights into the interactions of the variables.
Proposed analysis methods are:

- **Variance analysis** for the development of the variables along dimensions
- **Entropy analyses** using Shannon and Sample Entropy
- **Mutual information analysis** for interactions between variables

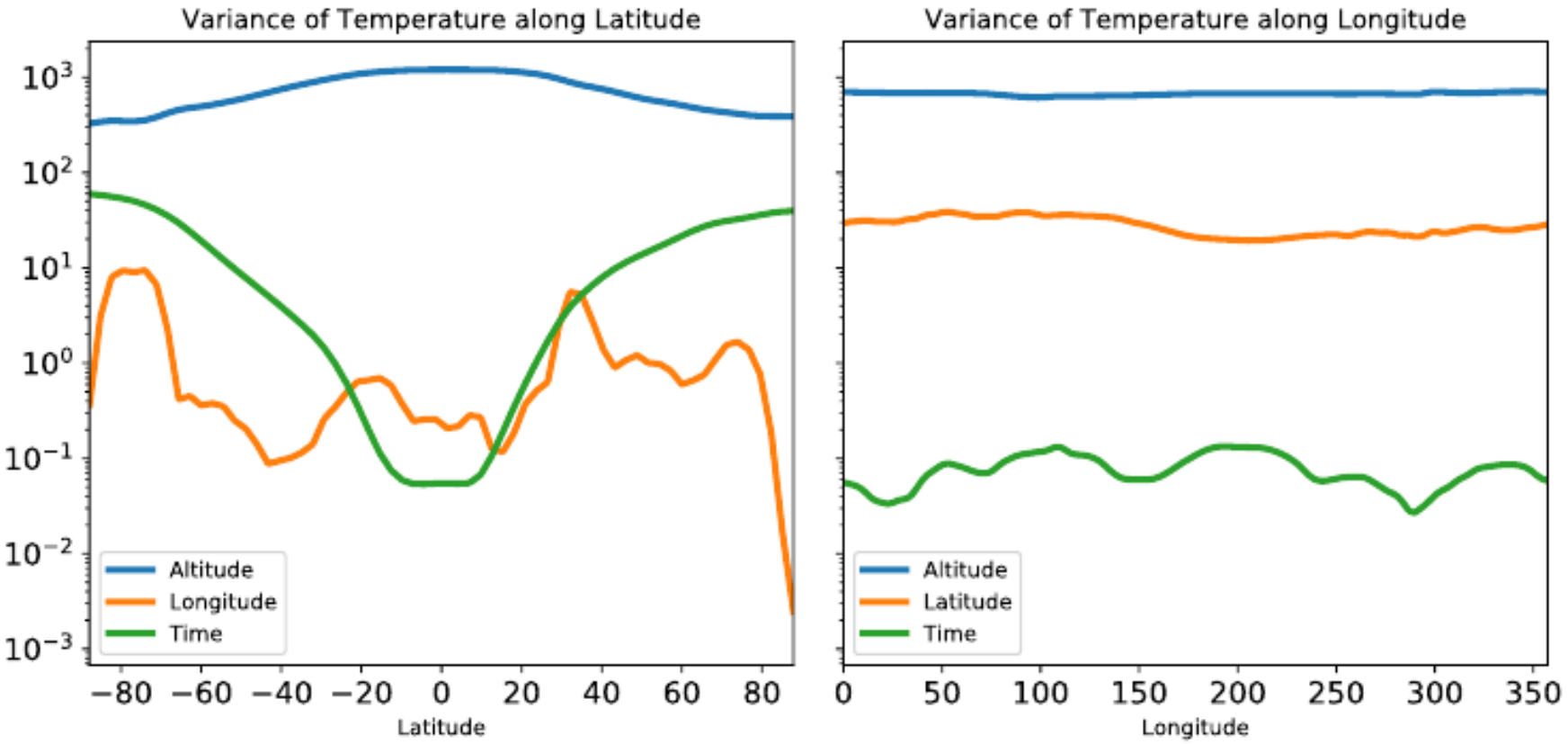
Finally, **compare available compression** algorithms for compression factor and throughput.

Experimental Setup

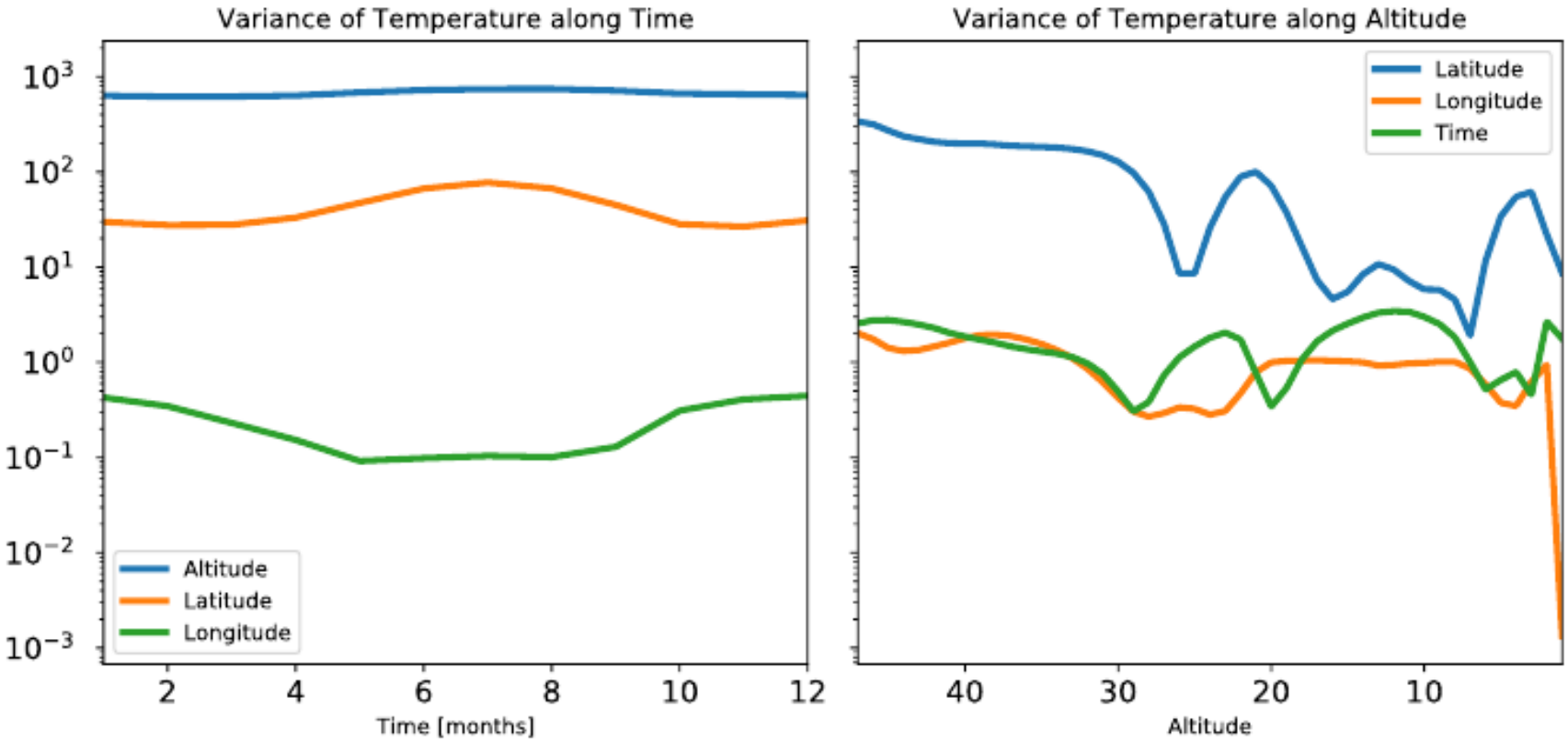
Climate simulation run using EMAC climate model

- **Spatial resolution** 128 x 64 x 47 (Longitude x Latitude x Altitude)
- **Temporal resolution** Three different datasets
  - One month (January 2013) w/ 74 time steps (every 10h)
  - One year (2013) w/ 365 time steps (every 24h)
  - Whole simulation (2000-2013) w/ 168 time steps (each month)
- **Data variables** as single precision floating-point data
  - Temperature, zonal and meridional wind, and specific humidity

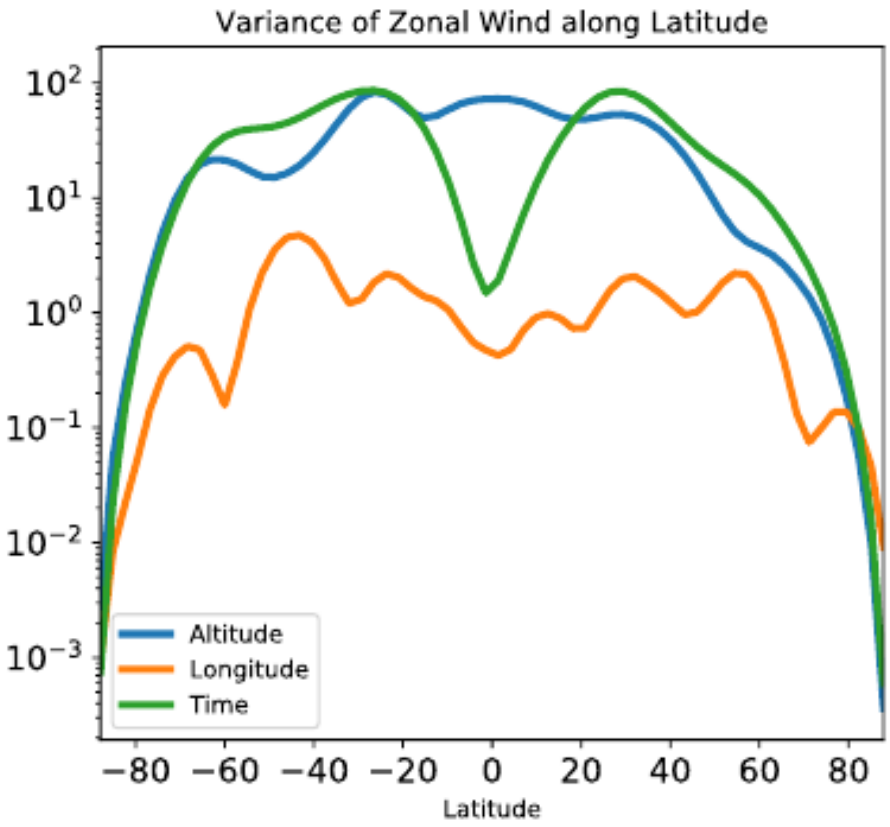
Variance Analysis - (Temperature)

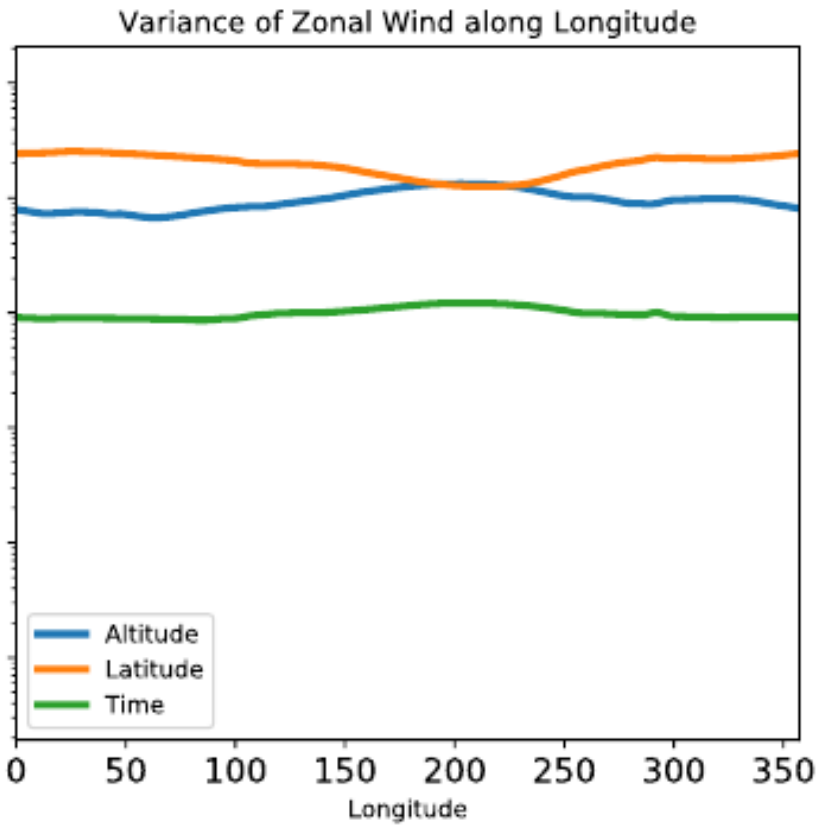


Variance Analysis - (Temperature)

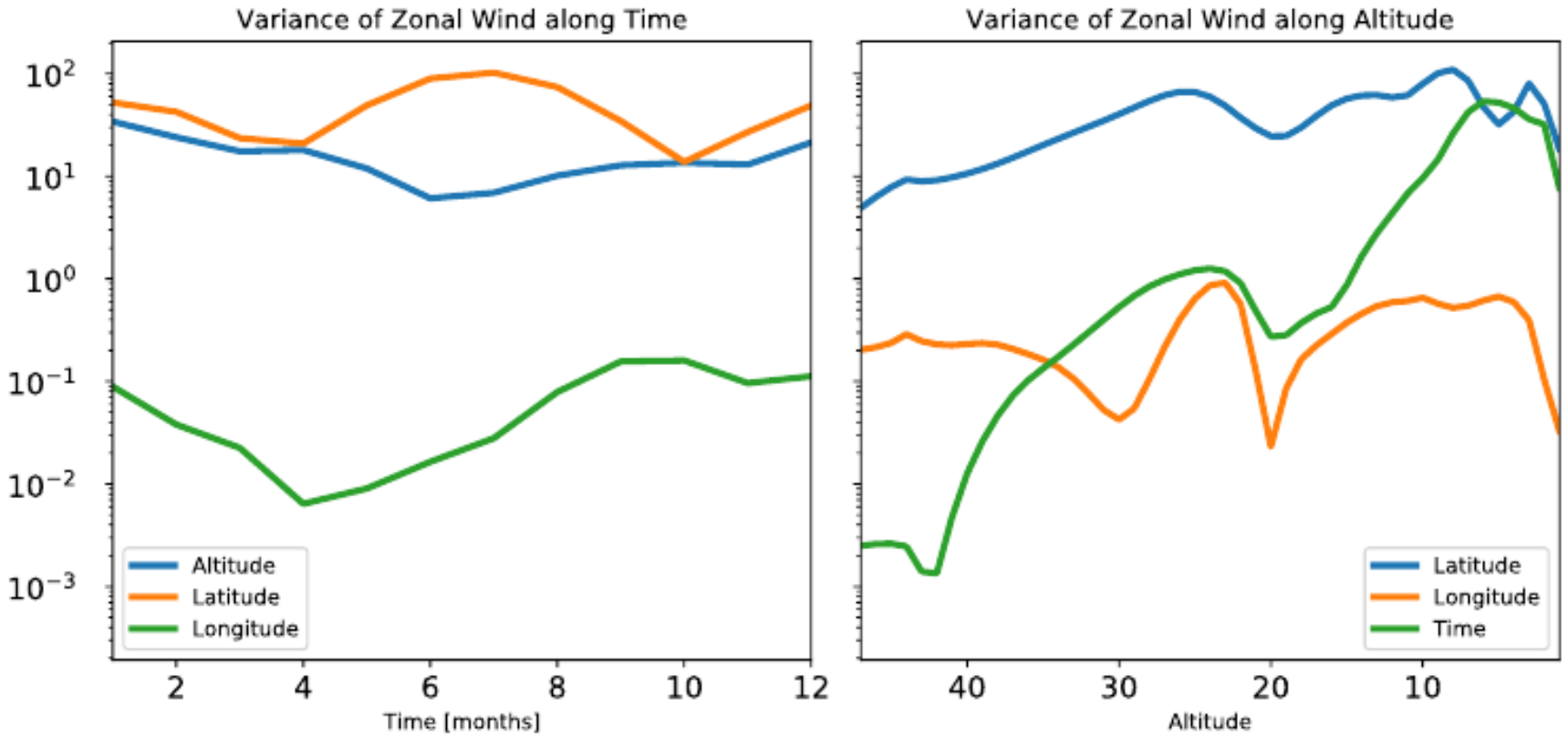


Variance Analysis - (Zonal wind)

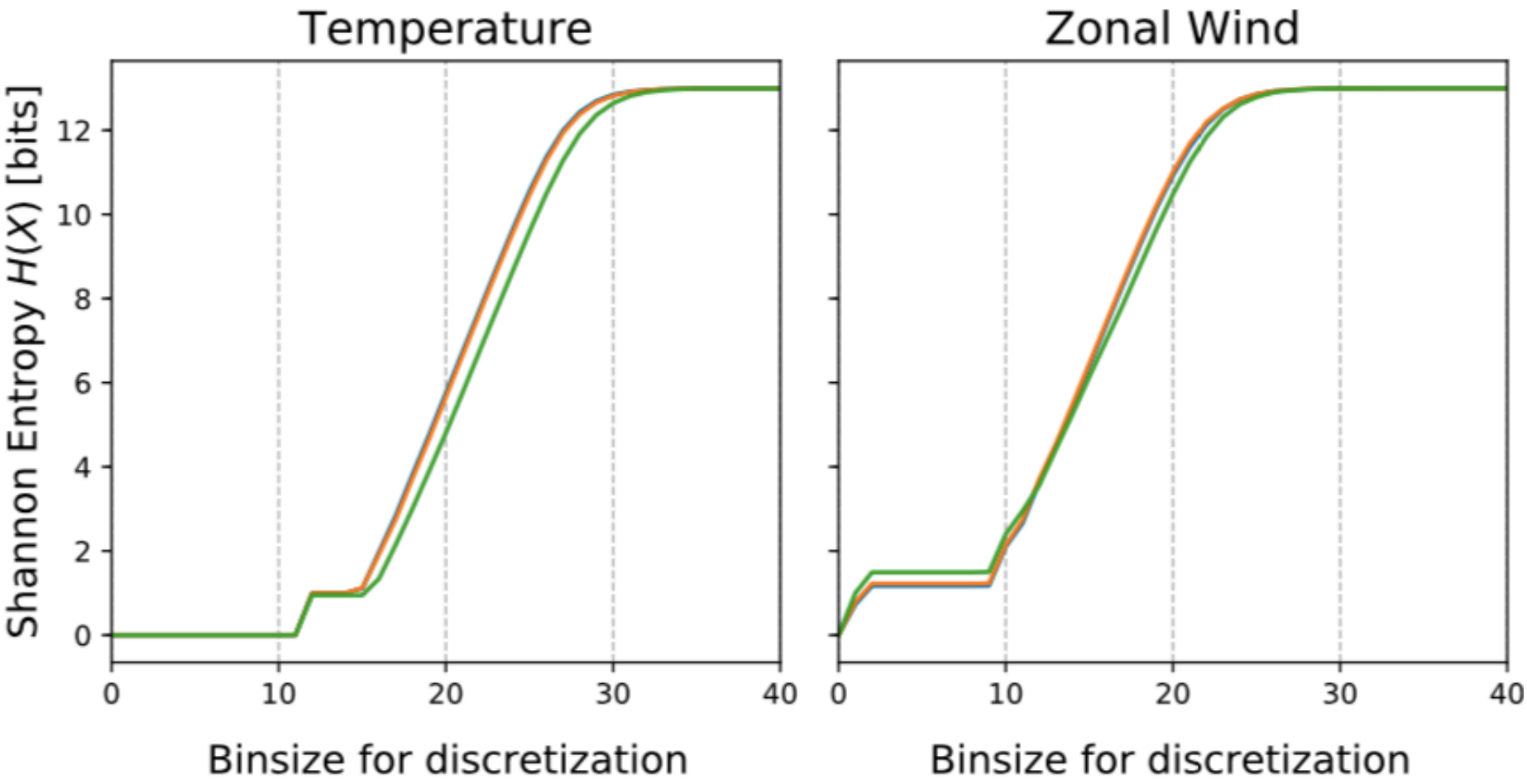




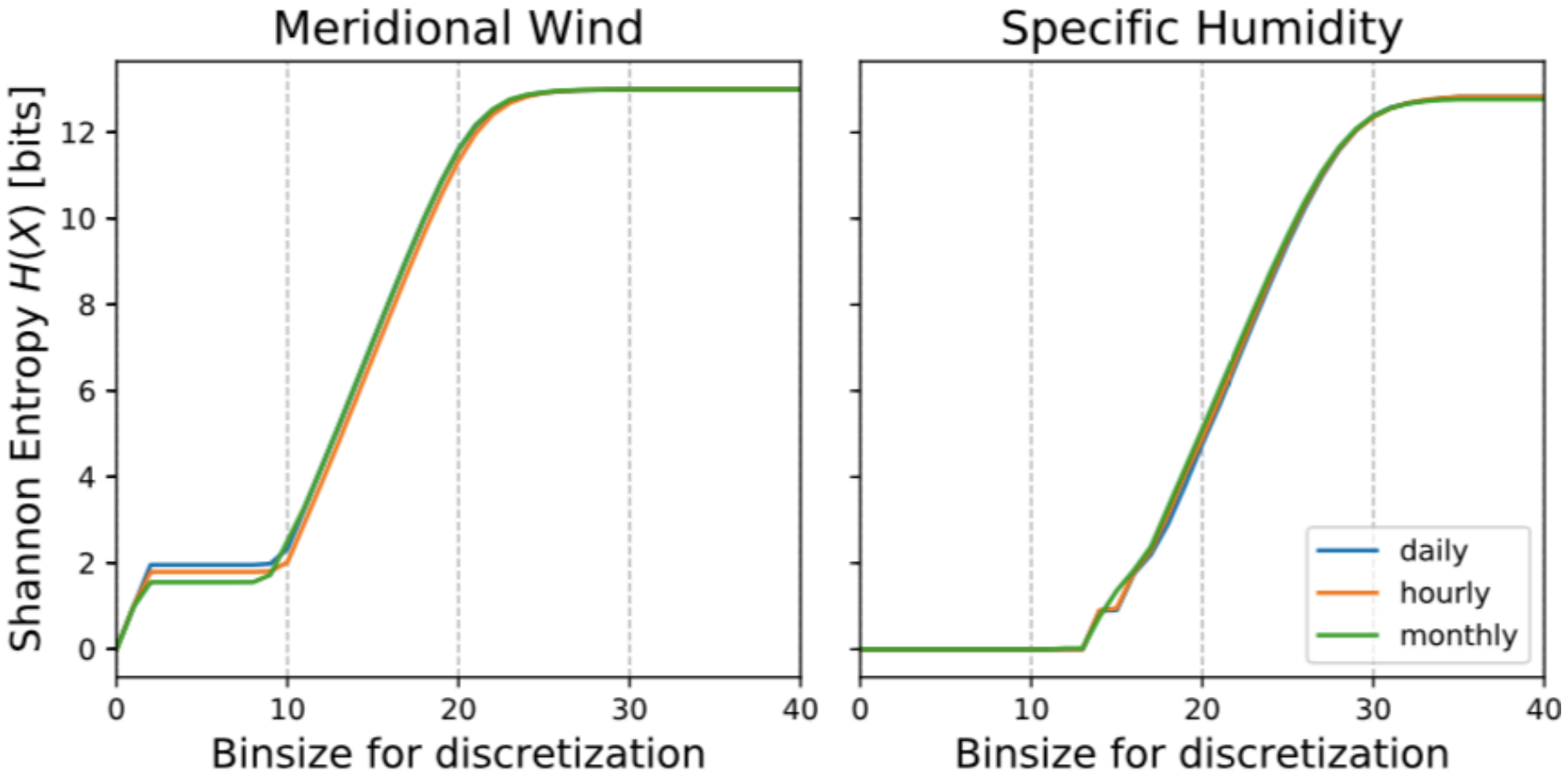
Variance Analysis - (Zonal wind)



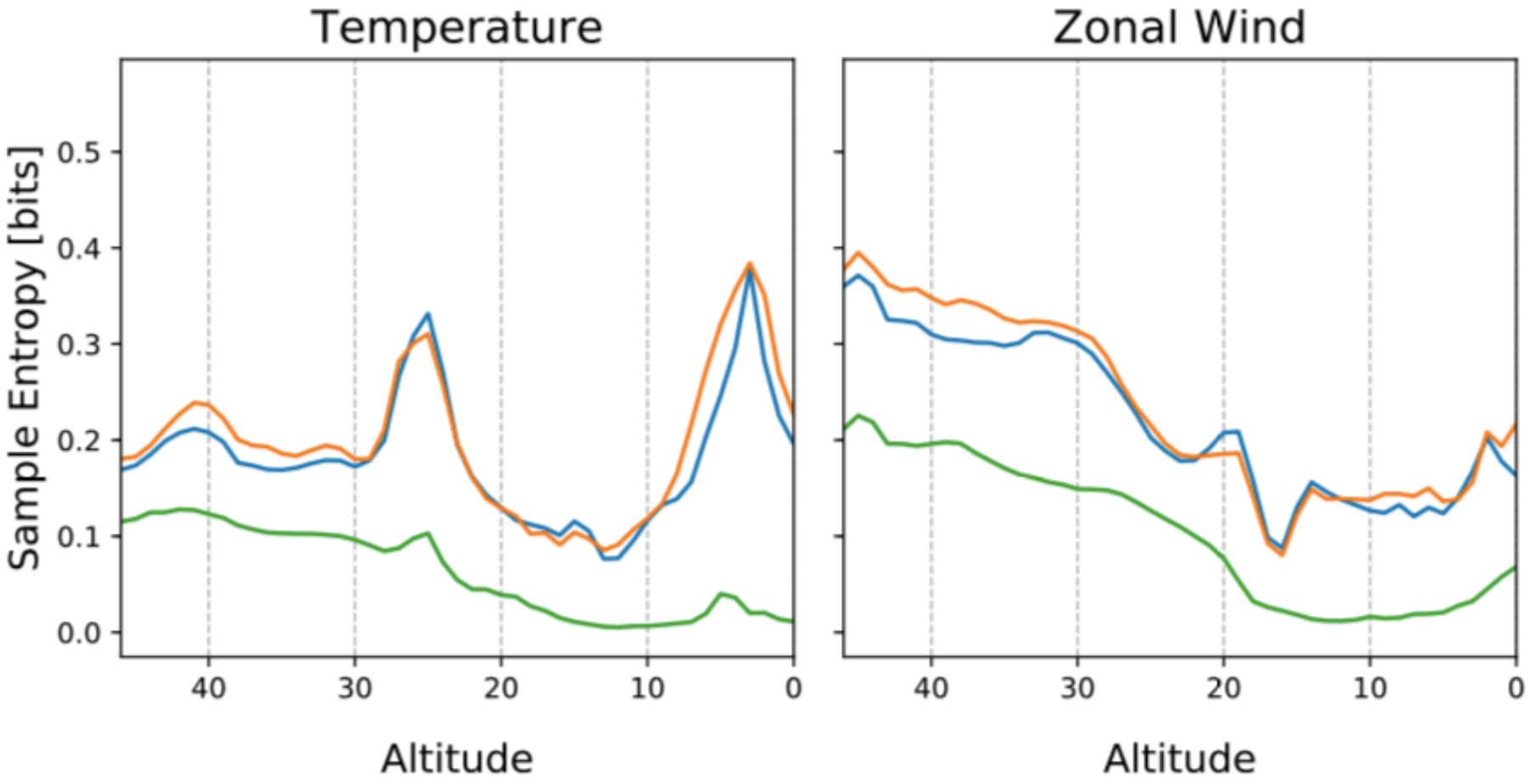
Shannon Entropy



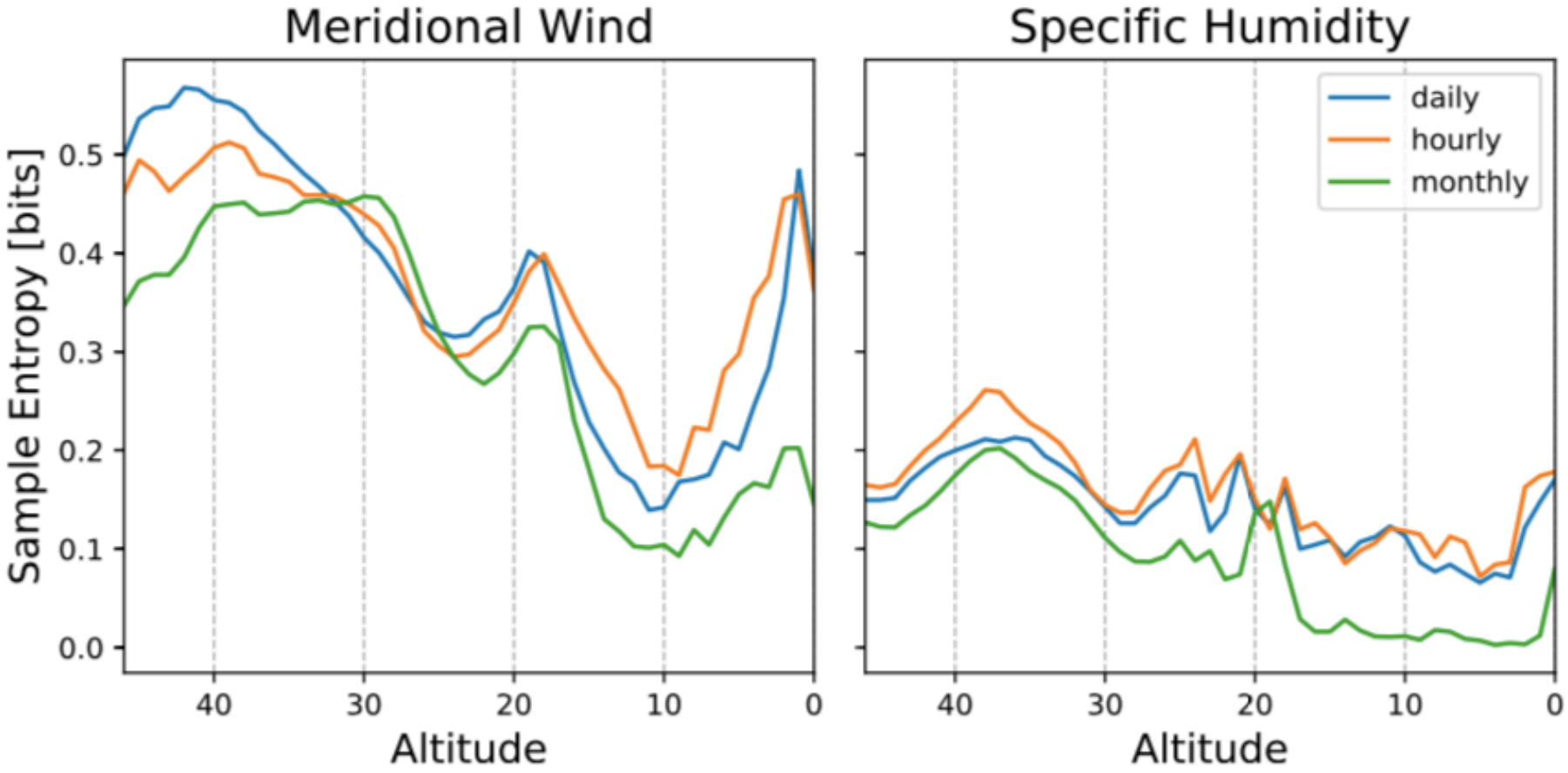
Shannon Entropy



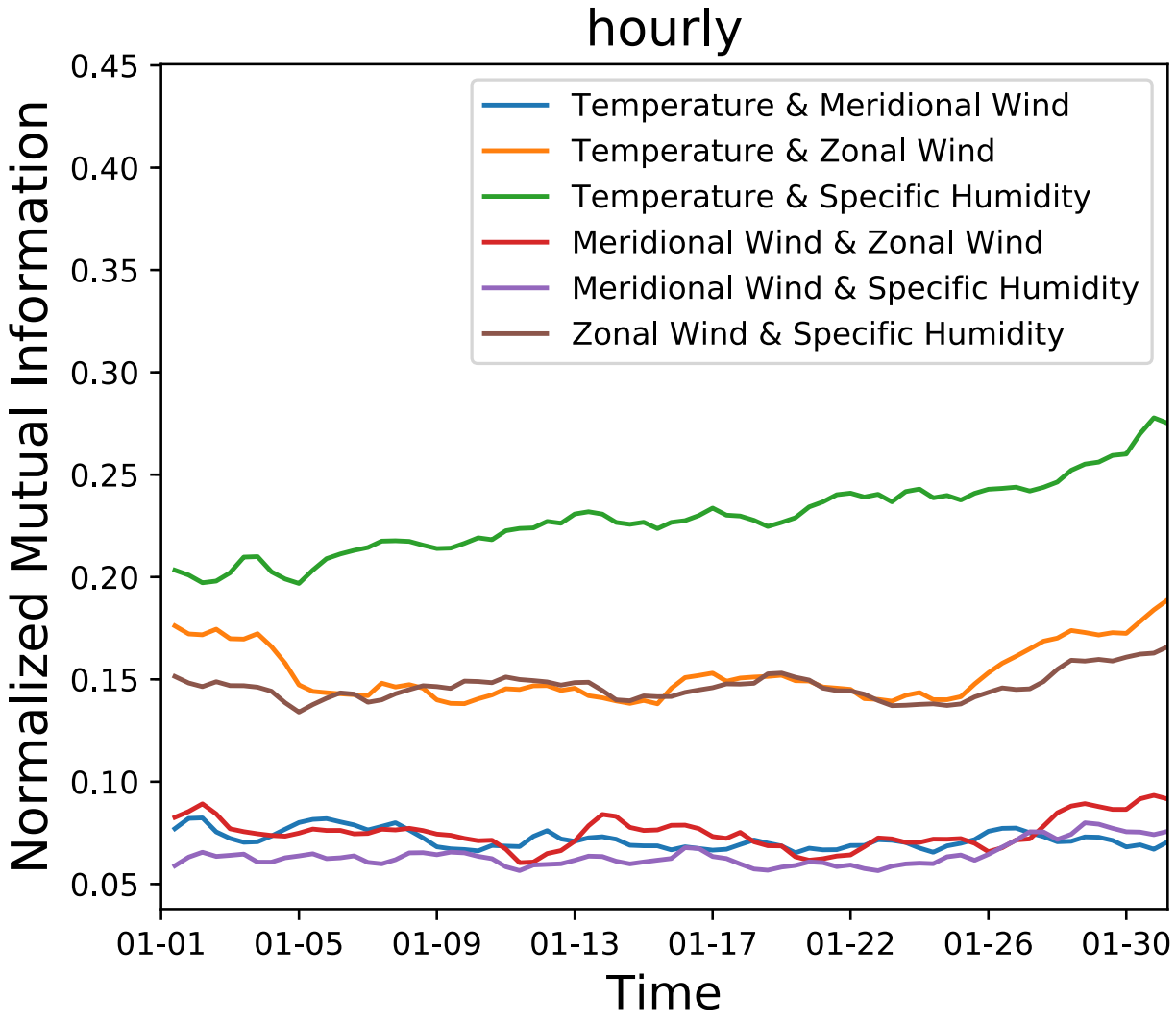
Sample Entropy



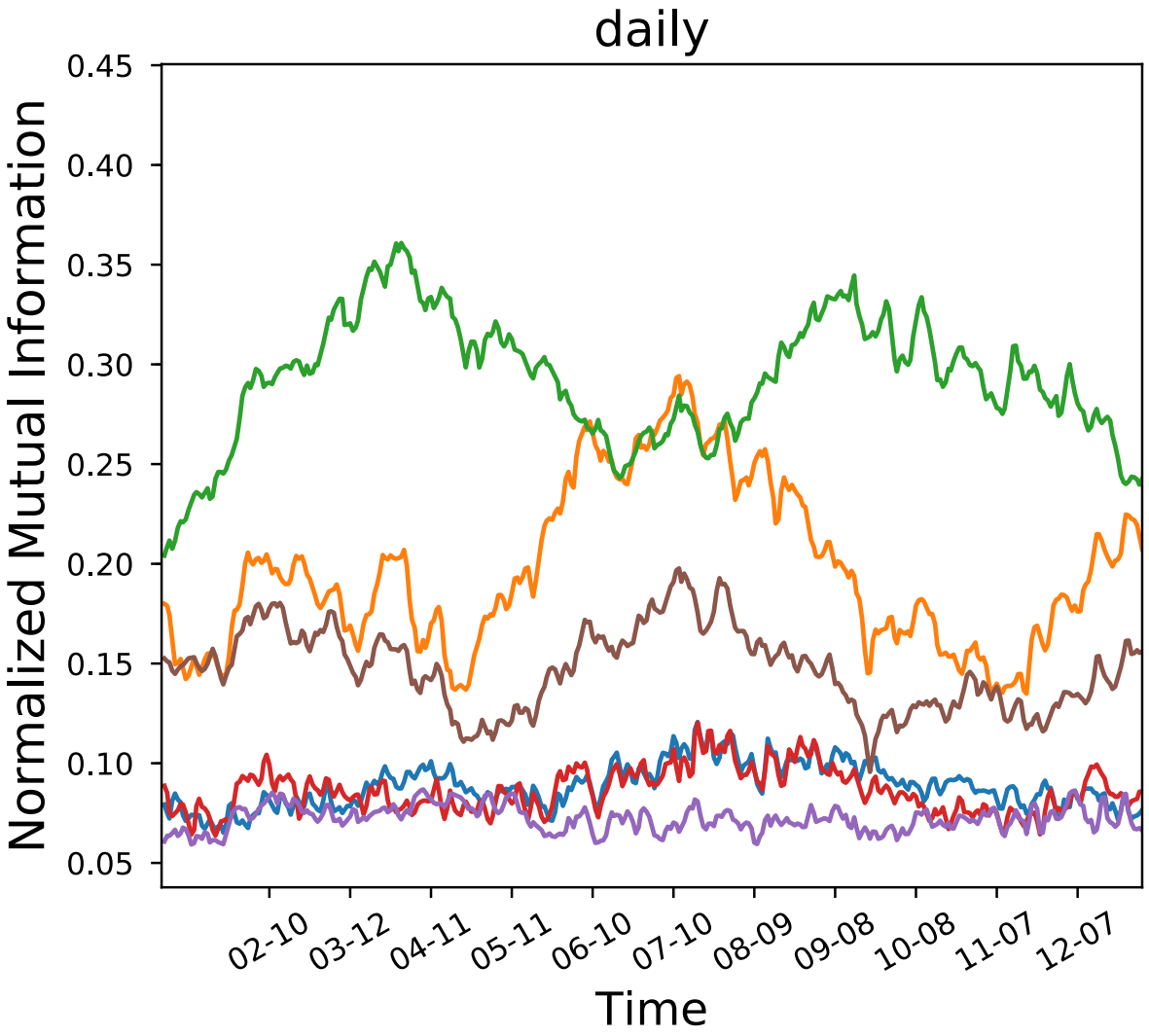
Sample Entropy



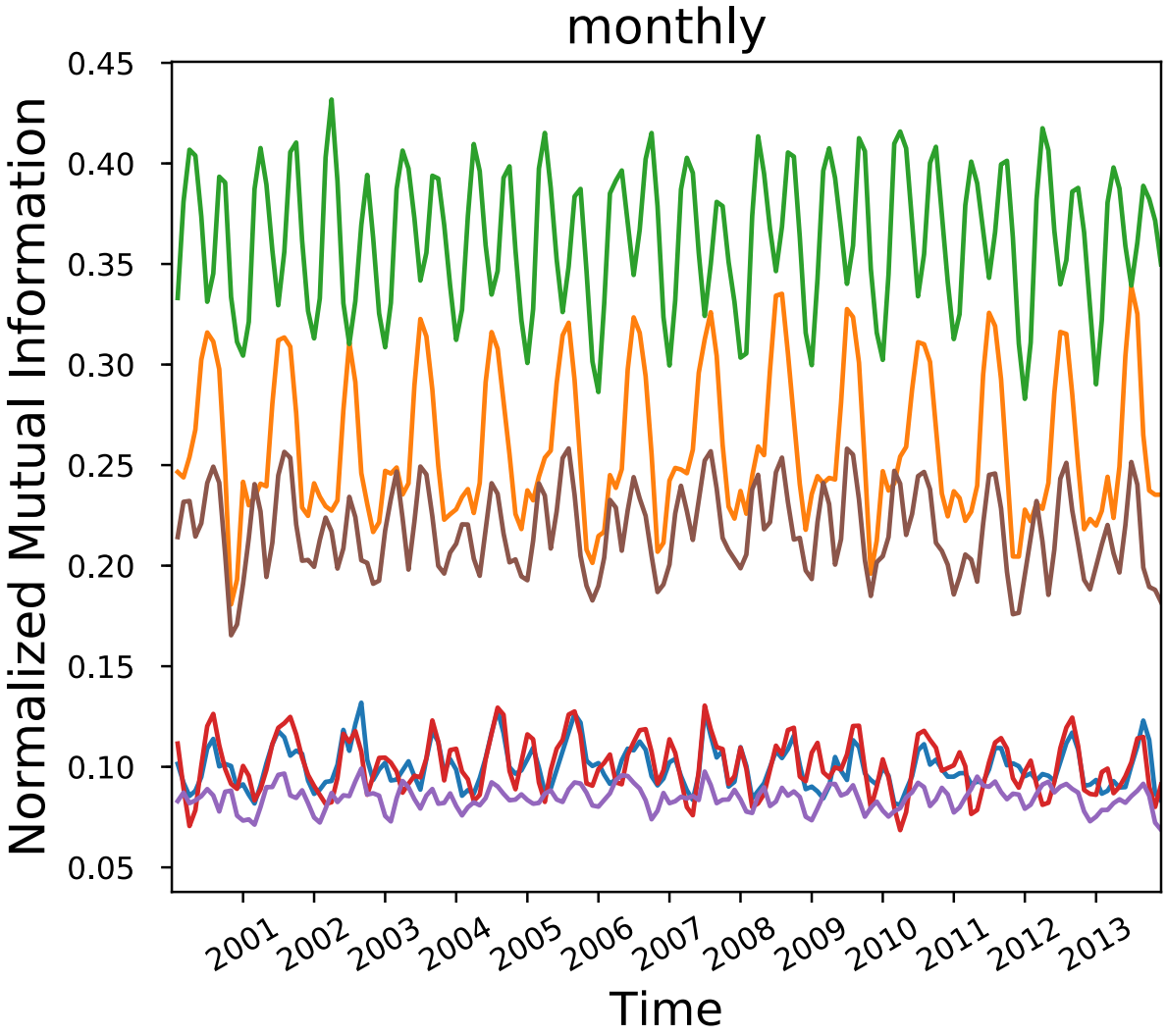
Mutual Information



Mutual Information



Mutual Information



Karlsruher Institut für Technologie

3. Integrating existing knowledge about the interactions of variables

Step 3

Take-home message

It is possible to improve lossy compression for **certain** time series data by only gradually increasing file size.

Application in climate research for compression of environmental indices.

Agenda

Karlsruher Institut für Technologie

Importance of compression for climate research

Introduction and description of proposed method

Application on environmental indices used in climate research

Climate data and importance of compression

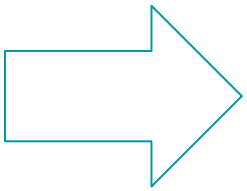
Karlsruher Institut für Technologie

- Environmental data is 4D (longitude, latitude, altitude, time)
- Current **E**uropean **Re**a**A**nalysis (ERA5) dataset needs 2.26 TiB p.a. and variable
  - Used by weather and climate simulations as ground truth

Climate data and importance of compression

Karlsruher Institut für Technologie

- Environmental data is 4D (longitude, latitude, altitude, time)
- Current **E**uropean **Re**-**A**nalysis (ERA5) dataset needs 2.26 TiB p.a. and variable
  - Used by weather and climate simulations as ground truth



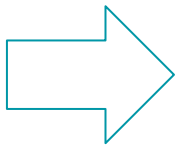
Generate a compression method specific for environmental data

But how?

Compression 101

$$\pi = 3 \ \vee \ \pi = \frac{C}{d}$$

- Lossy v lossless
- Distinguish between information and data
  - Understand the relationship within the data
  - Eliminate data without information
  - Compression works best with redundant data
- Temporal and spatial information can help to predict the behaviour variables.



Environmental indices

Compression 101

Karlsruher Institut für Technologie

$$\pi = 3 \ \vee \ \pi = \frac{C}{d}$$

- Lossy v lossless
- Distinguish between information and data
  - Understand the relationship within the data
  - Eliminate data without information
  - Compression works best with redundant data
- Temporal and spatial information can help to predict the behaviour variables.

Environmental indices

Karlsruher Institut für Technologie

- Temporal information of observations for forecasting weather phenomena like precipitation or monsoon season.
  - ENSO34
  - NAO
  - QBO30/50
  - …
- Idea: These indices can to be saved and used by the compression algorithm to gain information about the data

What are success metrics?

Karlsruher Institut für Technologie

- **Compression ratio**
  - Filesize (after) / Filesize (before)
- Memory usage
- Compression/Decompression time

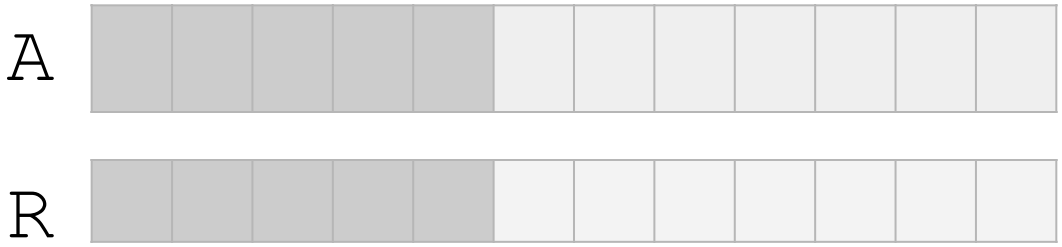
Additional criteria in case of lossy compression:

- **Quality of reconstructed data (community specific)**

Quality criteria for compressed indices

Karlsruher Institut für Technologie

- A lossy compression algorithm is considered successful, if the correlation between the **original time series A** and the **reconstructed time series R** is 1.



$$Corr_{s,e}(A,R) = 1.0$$

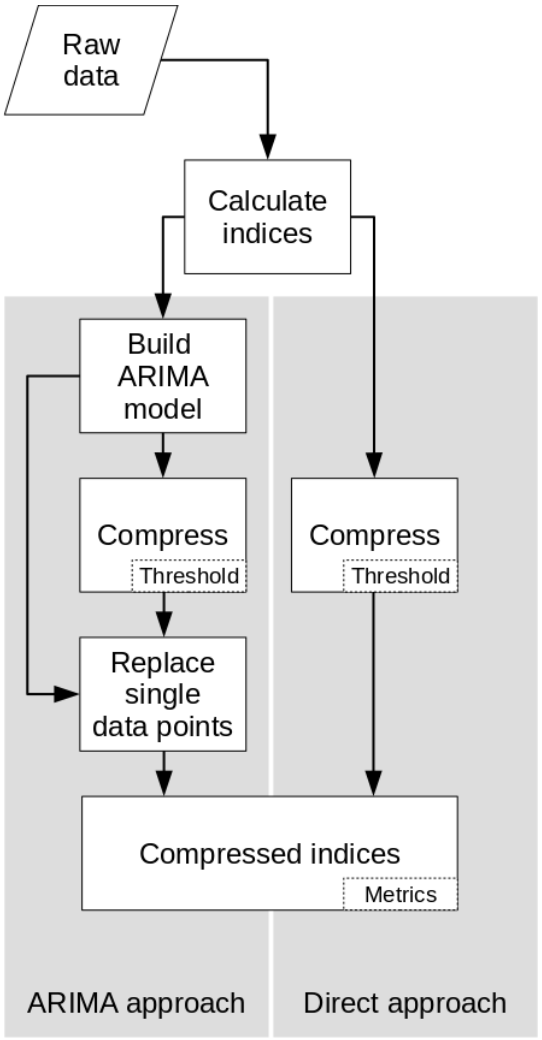
Direct & proposed approach

Direct approach

Compression using zfp
(which allows lossy compression by gradually lowering precision).

ARIMA approach

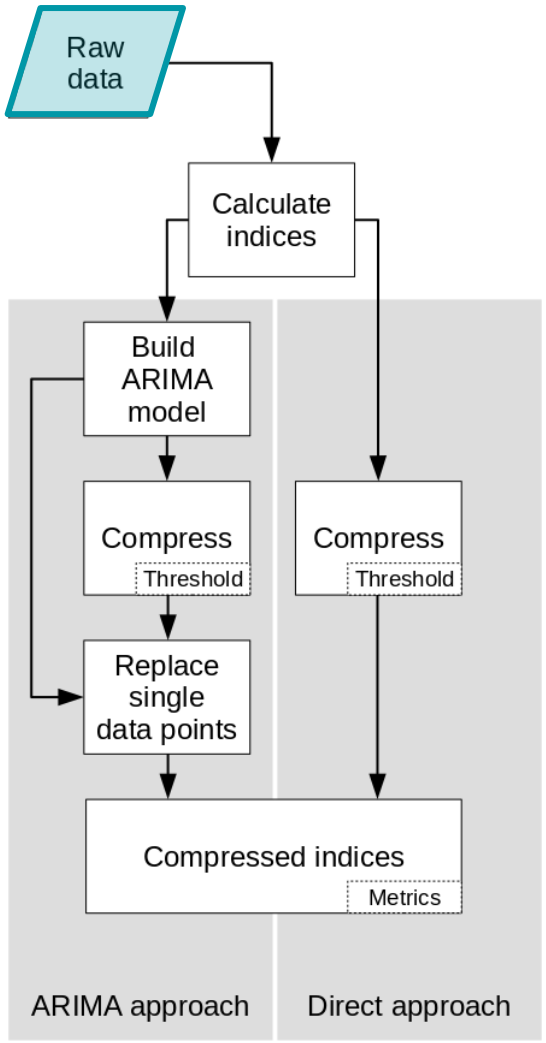
From us proposed approach.



Raw data

128 x 64	horizontal grid
6	vertical level
1979 - 2013	temporal (10h timesteps)

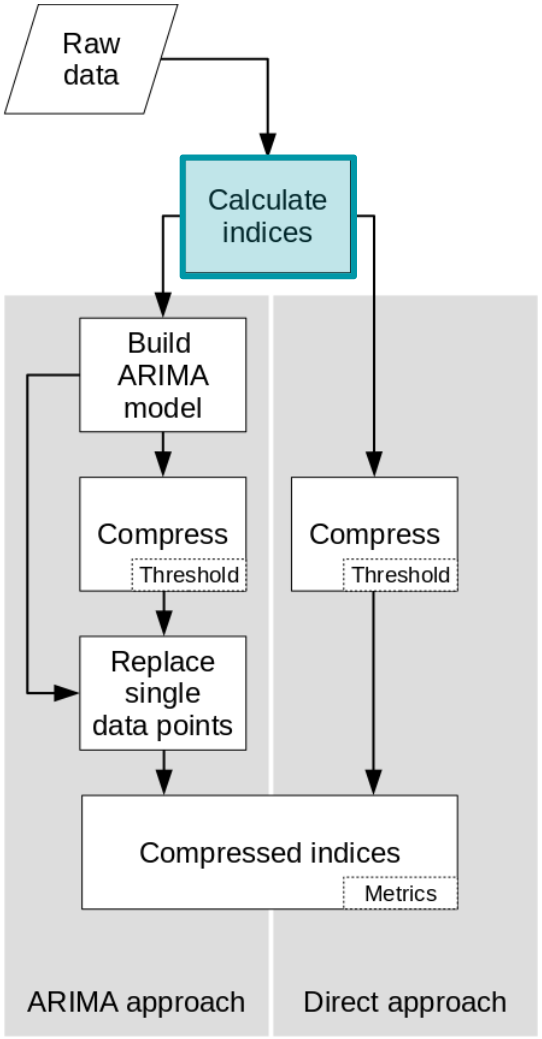
daily monthly



Calculate indices

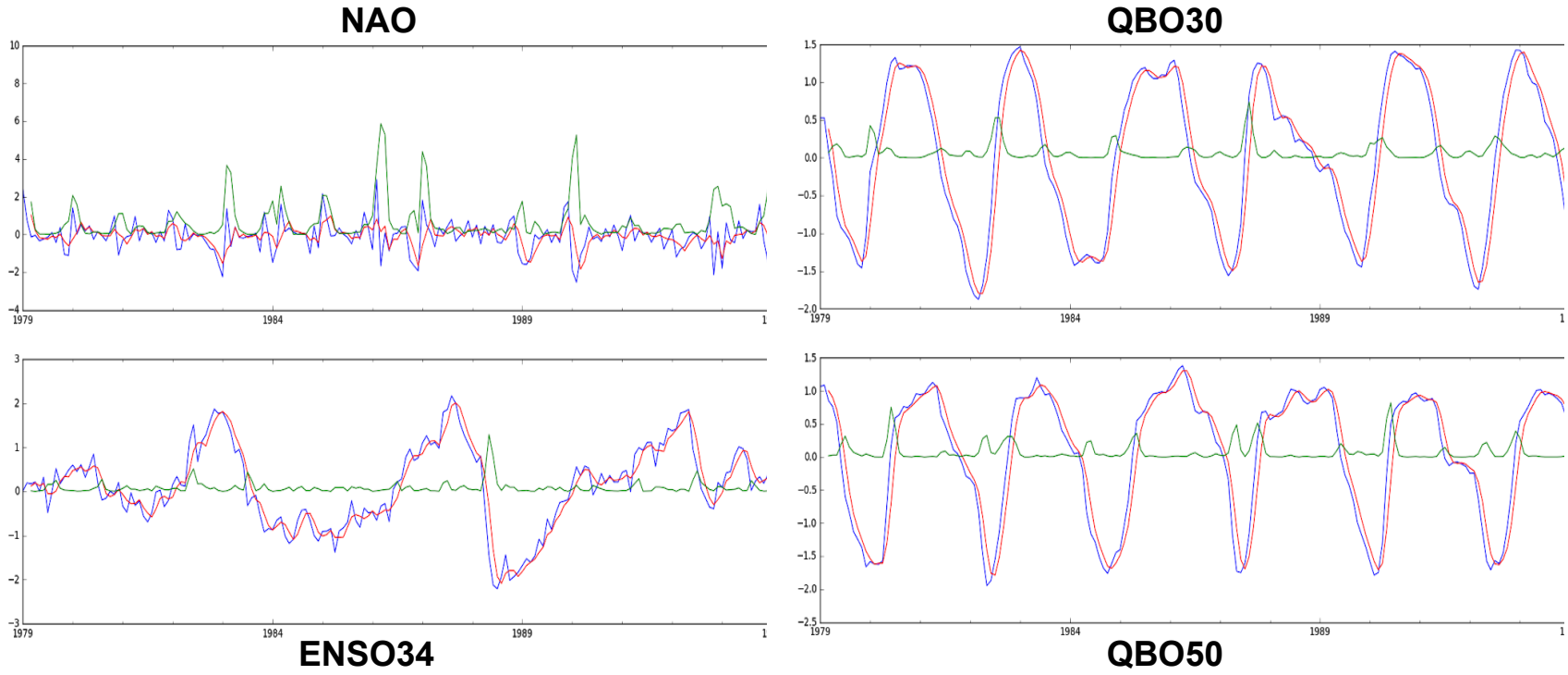
Index	Var	Lat [N]	Lon [E]	Alt [hPa]
ENSO34	T	[-5;5]	[190;240]	surface
QBO*x*	u	[-5;5]	[0;360]	*indicated by x*
NAO	p	Lisbon and Reykjavík		surface

T = temperature, u = westerly wind, p = pressure



Calculate indices

- Stationary time series
  - No trend
  - Variance is time independent

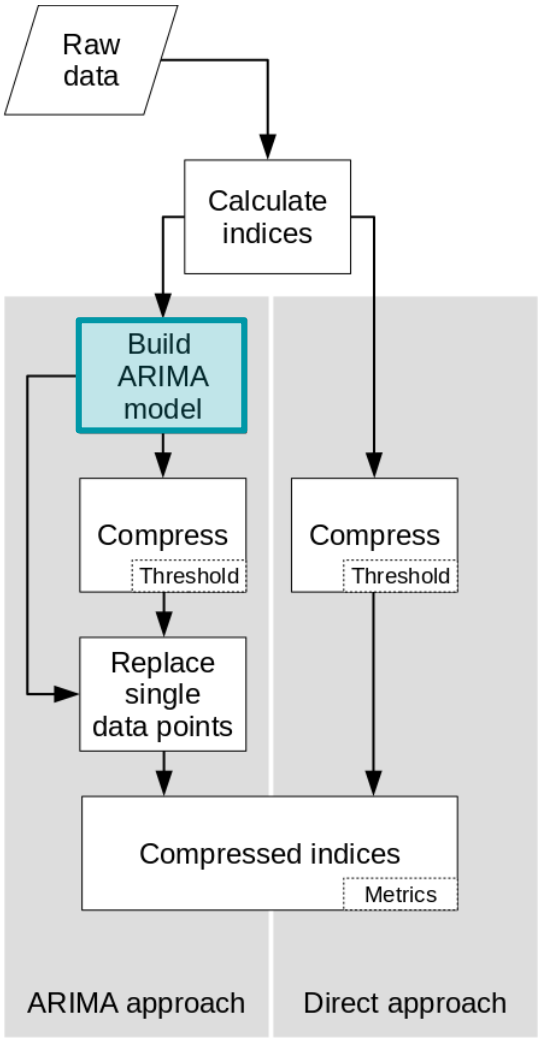


Build an Auto Regressive Integrated Moving Average model

Model the relationship of a data point x with its preceding values and predict future values.

Auto Regressive:
Regression on previous values.

Moving Average:
Regression on previous errors.



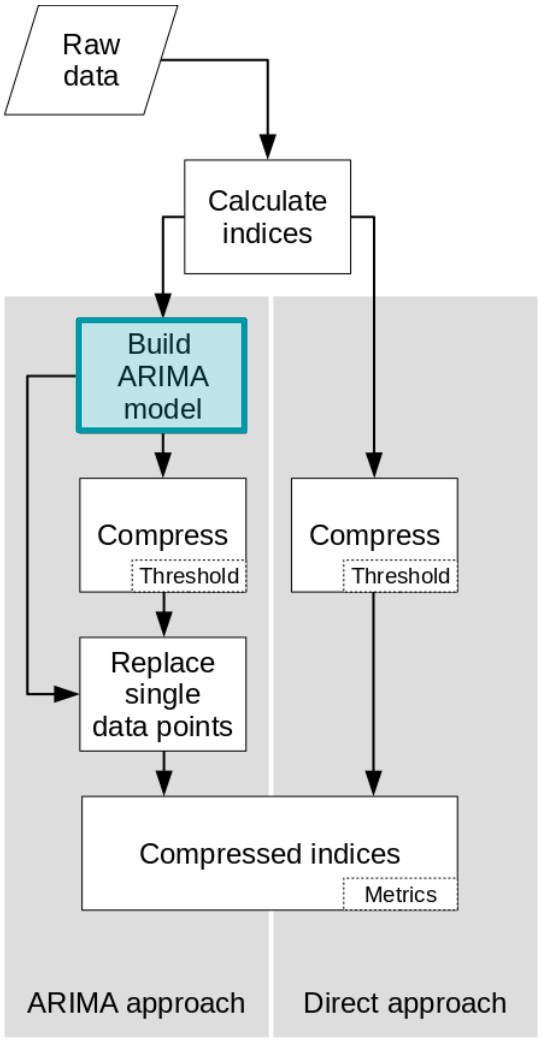
Build an Auto Regressive Integrated Moving Average model

Notation used for the ARIMA model:

$$ARIMA(p, d, q)(P, D, Q)_s$$

p = autoregressive order
d = differential order
q = moving average order
s = seasonal period

P, D, Q = appropriate seasonal order



Build an Auto Regressive Integrated Moving Average model

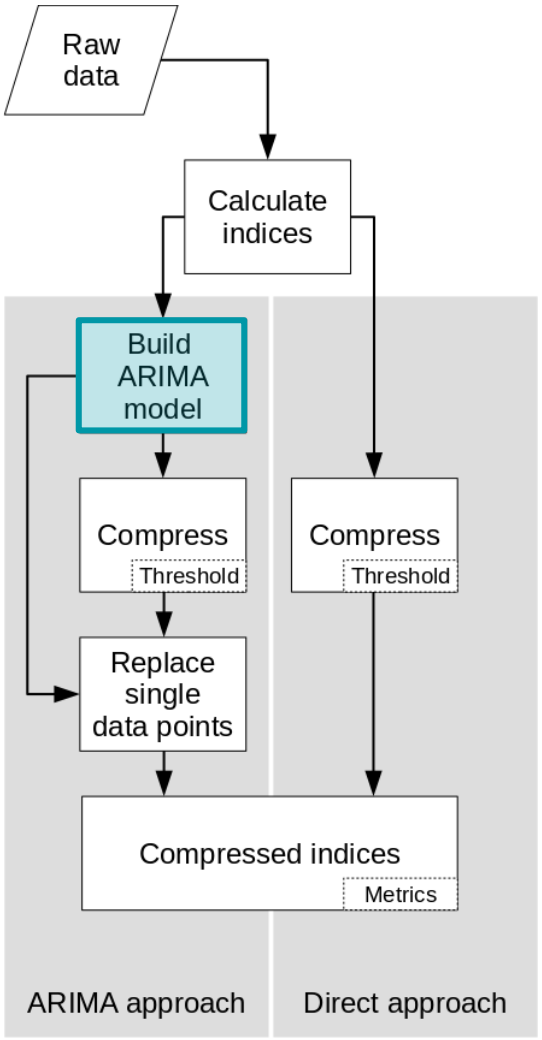
Notation used for the ARIMA model:

$$ARIMA(p, d, q)(P, D, Q)_s$$

The ARIMA model produces a prediction x' for the time series x which has an error of e.

$$x_i = x'_i + \epsilon_i$$

The time series x can be fully reproduced if the parameter of the ARIMA model and e are known.

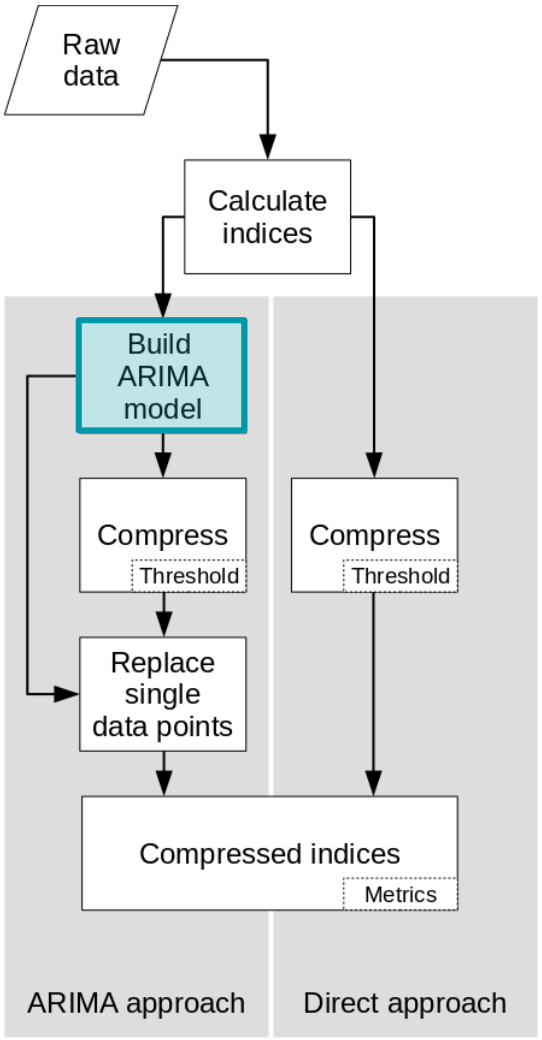


Build an Auto Regressive Integrated Moving Average model

ARIMA(p, 0, q)(0, 0, 0)0

$$x_i = x'_i + \epsilon_i$$

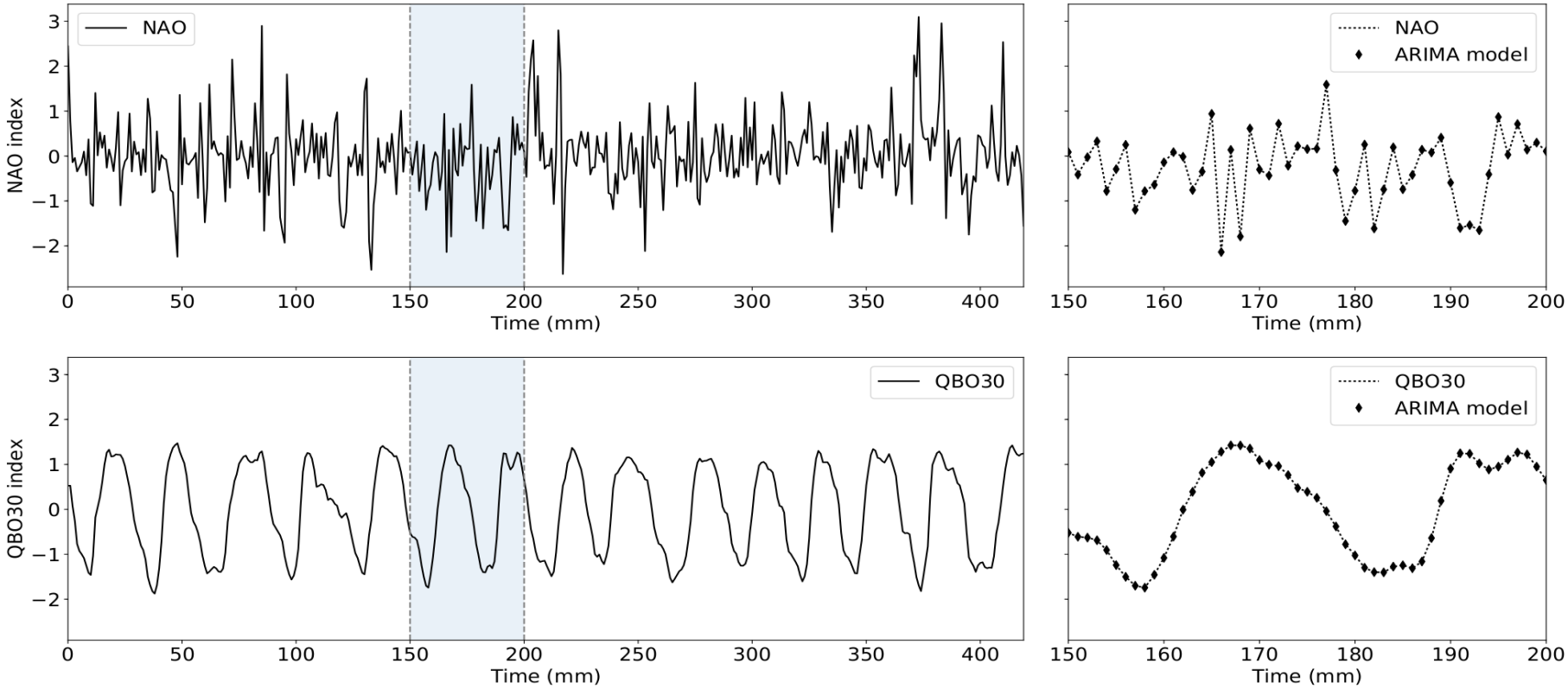
$$x'_i = \sum_{k=1}^{p} ar_k \cdot x_{i-k} + \sum_{j=1}^{q} ma_j \cdot x_{i-j}$$



Build an Auto Regressive Integrated Moving Average model

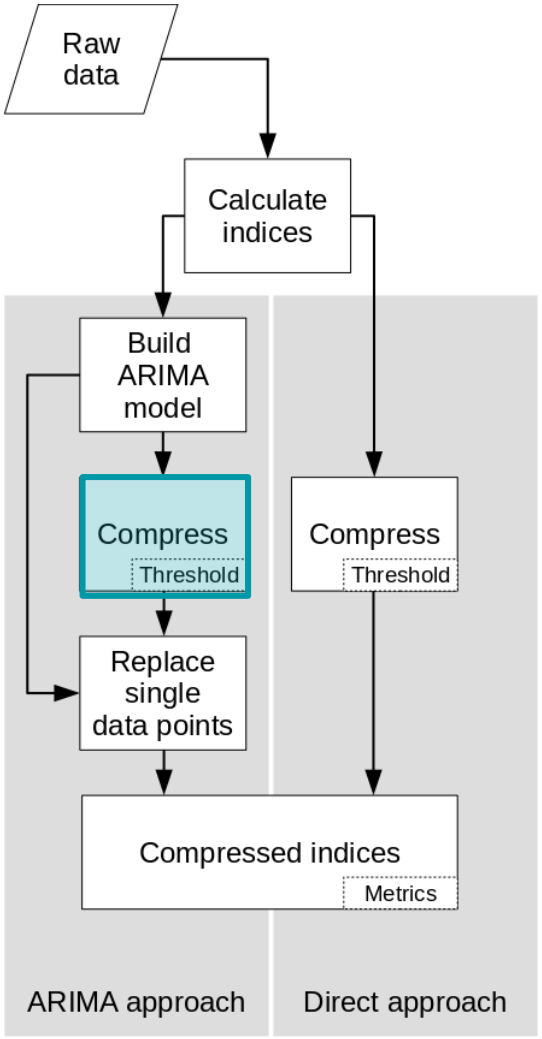
Timeline	ARIMA Model			
	Monthly		Daily	
	Model	RMSD	Model	RMSD
ENSO34	ARIMA(3,0,2)(1,0,0)$_{12}$	5.067e-8	ARIMA(5,2,4)(0,0,0)0	4.686e-4
NAO	ARIMA(1,0,0)(1,0,0)$_{12}$	8.195e-9	ARIMA(2,0,2)(0,0,0)0	1.440e-7
QBO30	ARIMA(2,0,3)(1,0,0)$_{12}$	1.0877e-7	ARIMA(5,0,4)(0,0,0)0	1.084e-7
QBO50	ARIMA(1,1,1)(1,0,1)$_{12}$	2.909e-6	ARIMA(5,0,4)(0,0,0)0	4.488e-8

How good is the reconstruction of the ARIMA model?



Different compression methods

- Experiment 1:
  Lossless compression
- Experiment 2:
  Lossy compression with threshold
- Experiment 3:
  Lossy compression with replacement



Exp. 1:
Lossless compression

	Compression ratio			
	Monthly		Daily	
	ARIMA	Direct	ARIMA	Direct
ENSO34	1.043	1.024	1.036	1.009
NAO	1.043	1.033	1.033	1.026
QBO30	1.038	1.005	1.032	0.961
QBO50	1.045	1.014	1.033	0.969

Exp. 2:
Lossy compression with threshold τ = 1e-5

	Compression ratio			
	Monthly		Daily	
	ARIMA	Direct	ARIMA	Direct
ENSO34	.386	.371	.658	.322
NAO	.386	.386	.377	.370
QBO30	.381	.357	.376	.273
QBO50	.668	.362	.377	.281

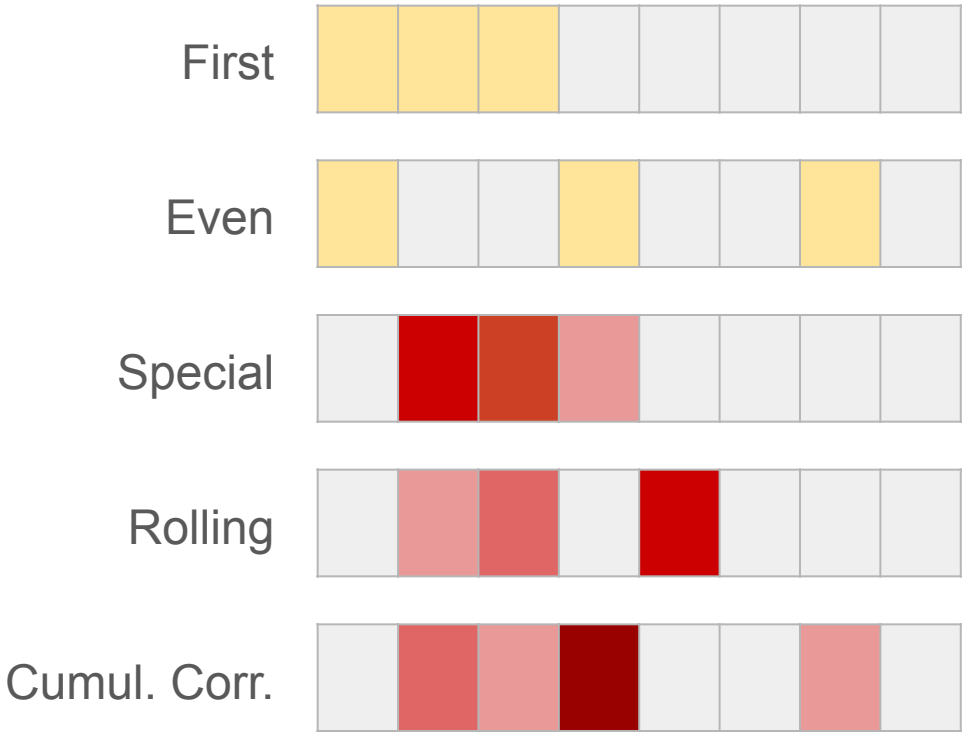
Exp. 2:
Lossy compression with threshold τ = 1e-5

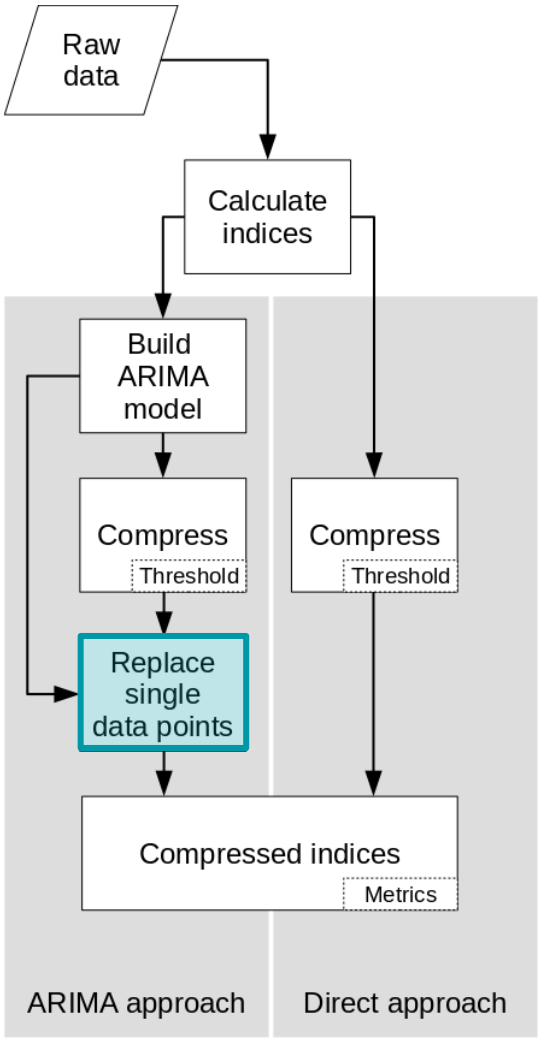
	Compression Ratio			
	Monthly		Daily	
	ARIMA	Direct	ARIMA	Direct
ENSO34	.386	.371	.658	.322
NAO	.386	.386	.377	.370
QBO30	.381	.357	.376	.273
QBO50	.668	.362	.377	.281

Exp. 3:
Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

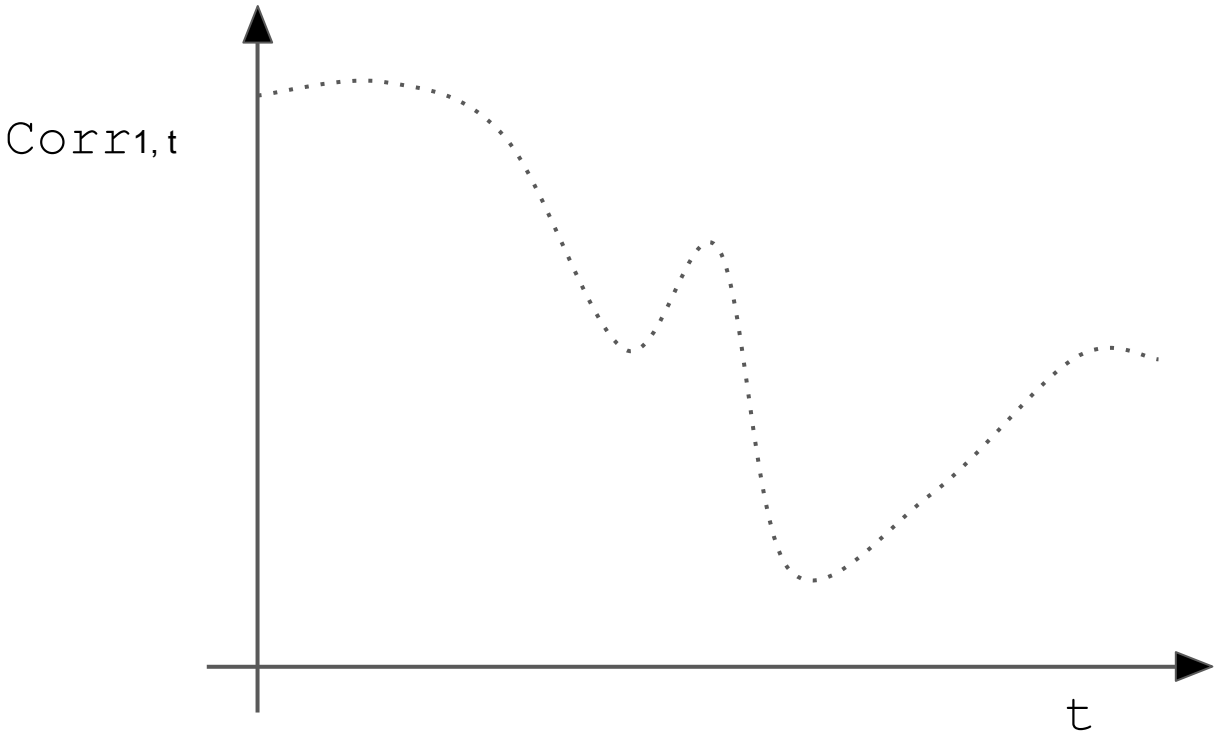
Methods for finding data points to be replaced:

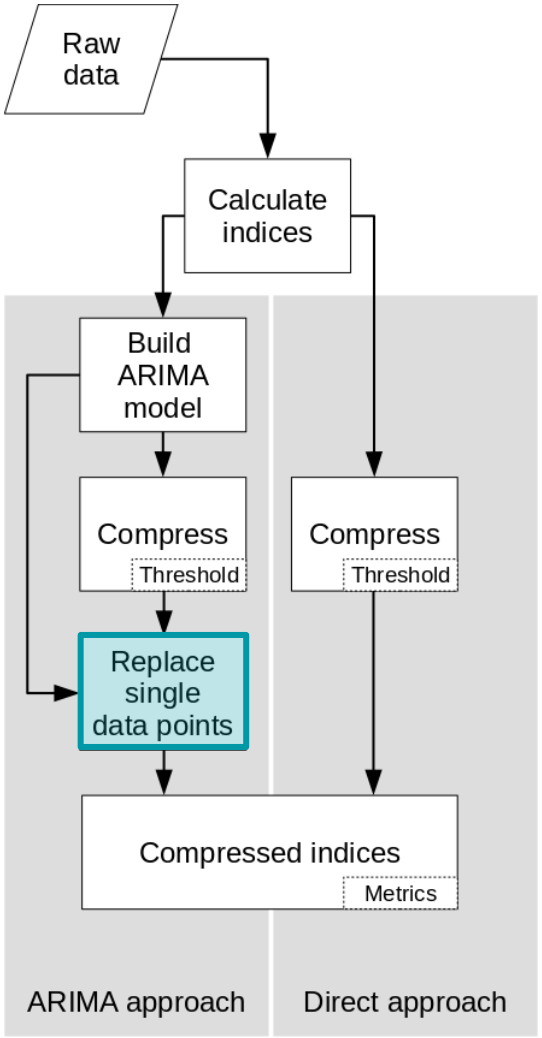




Exp. 3:
Lossy compression (w/ replacement)

Finding replacement methods

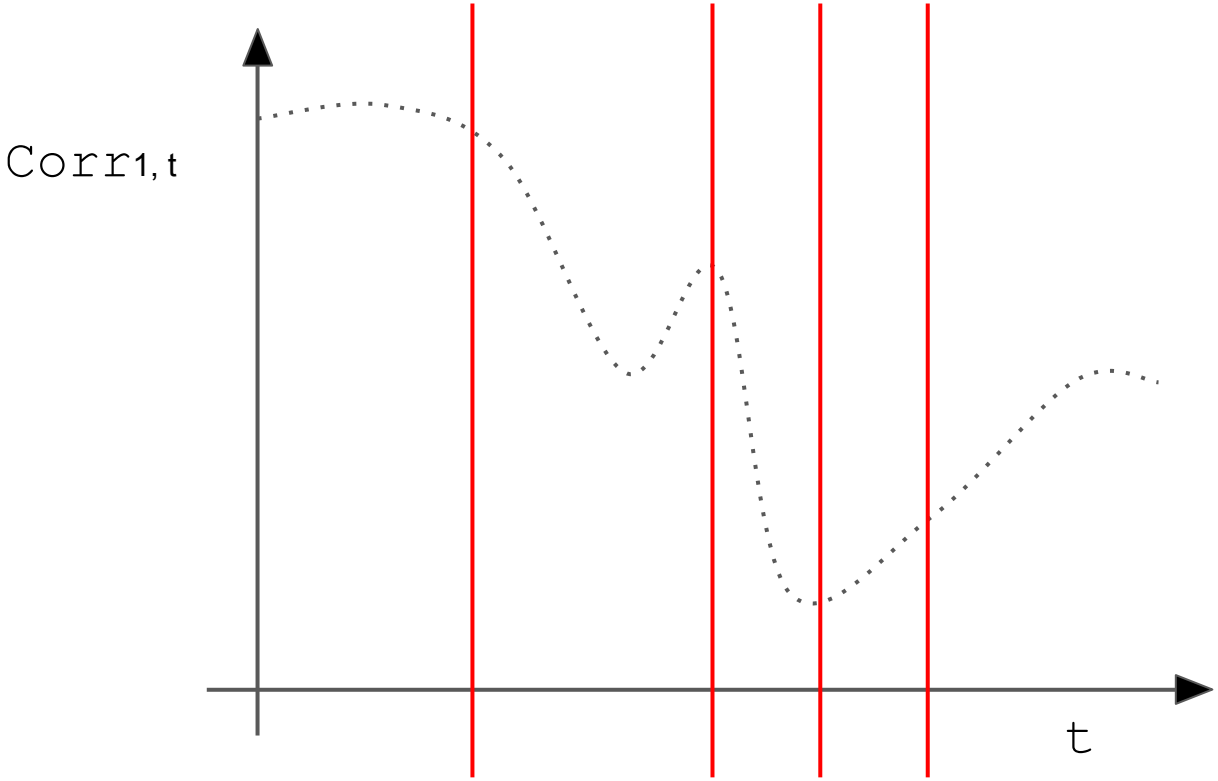


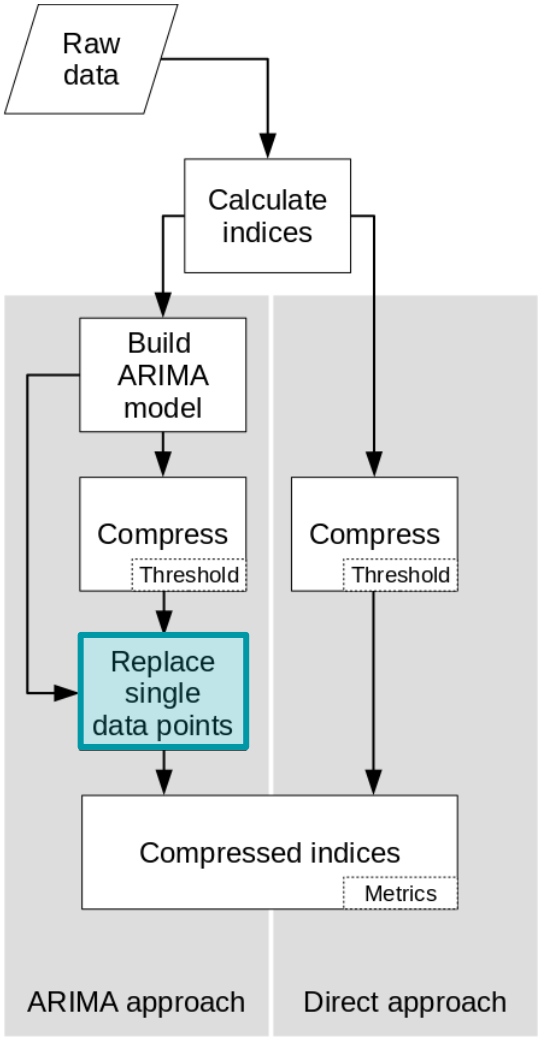


Exp. 3:
Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

Finding replacement methods

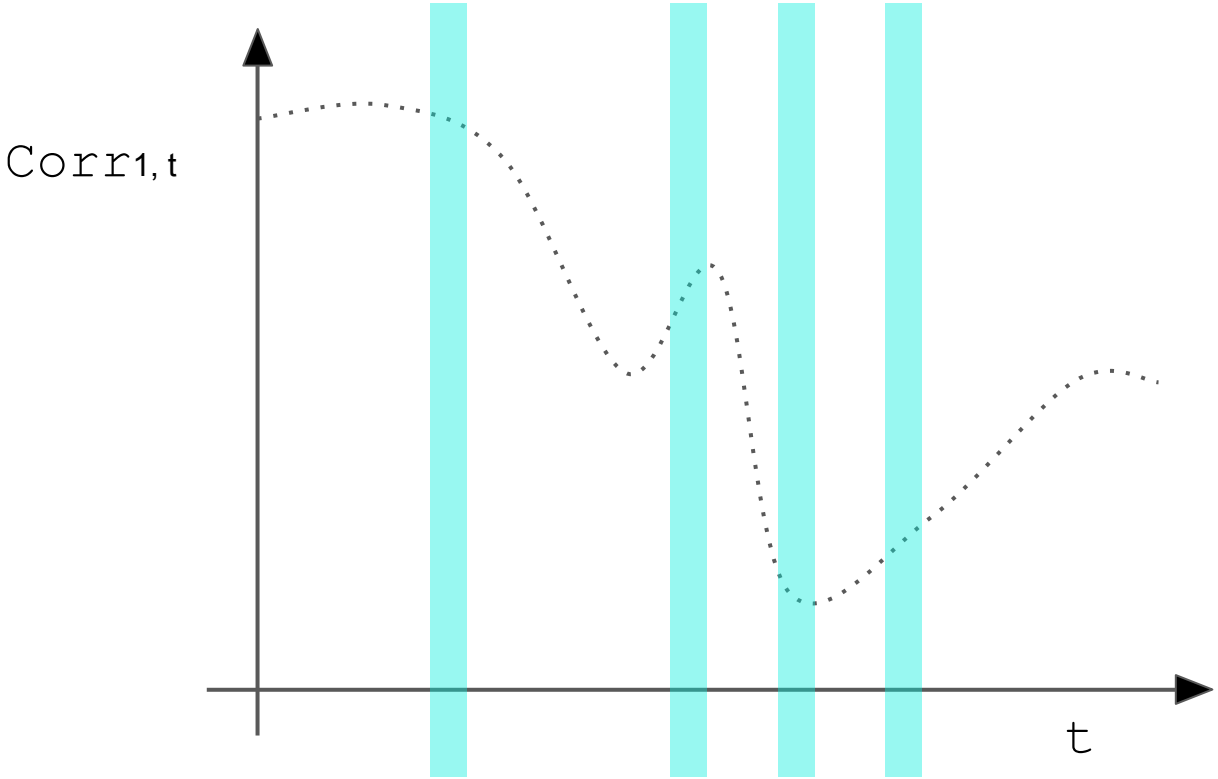


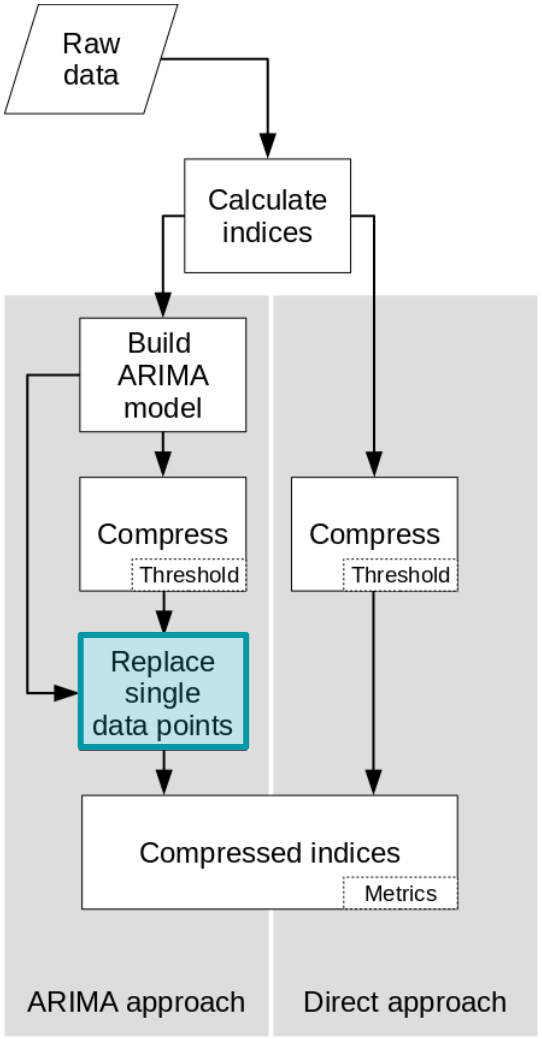


Exp. 3:
Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

Finding replacement methods

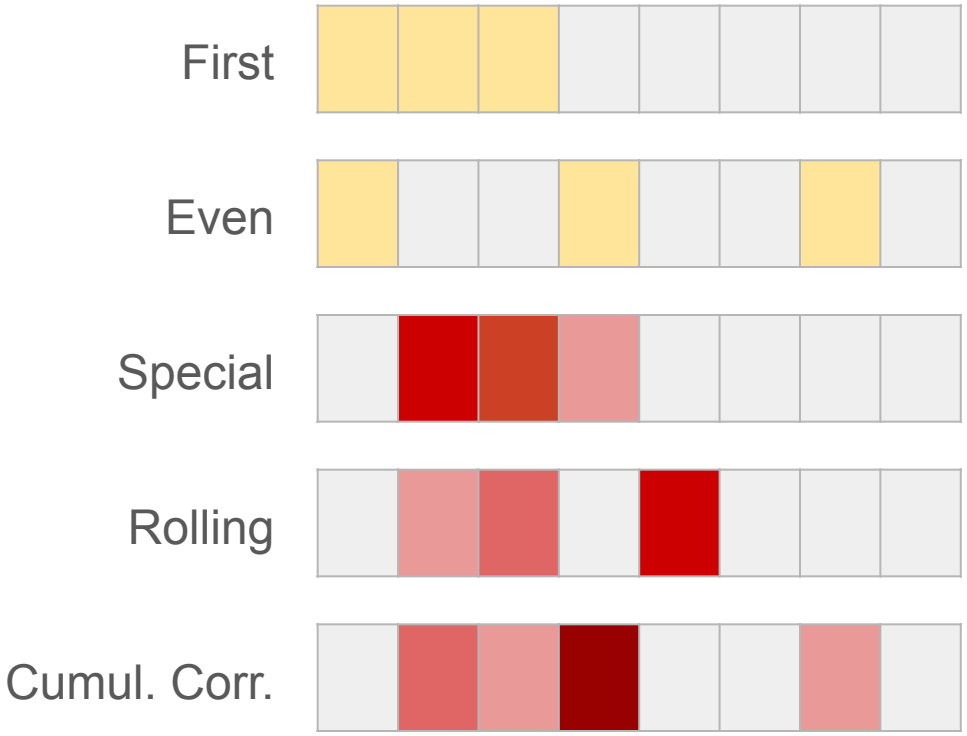


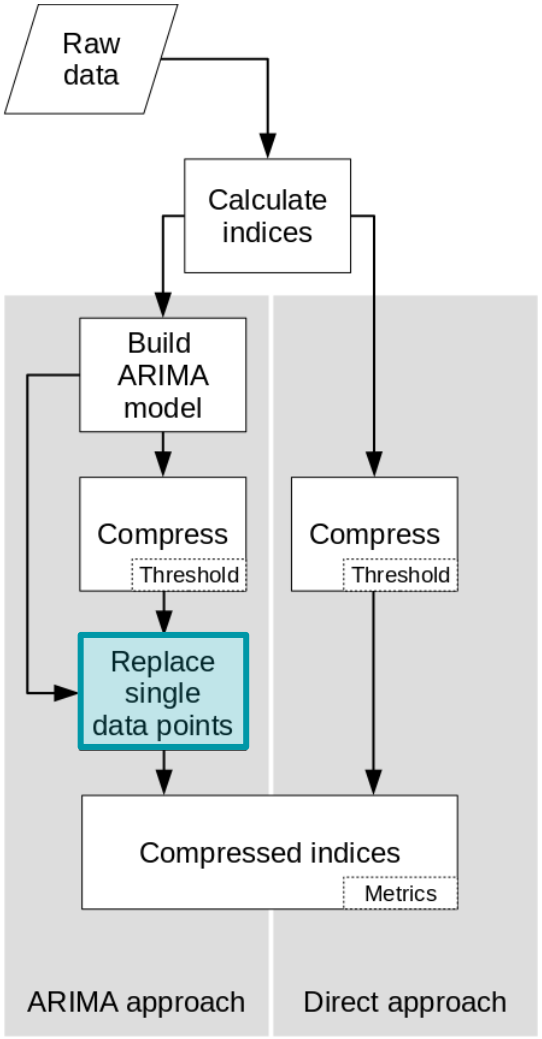


Exp. 3:
Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

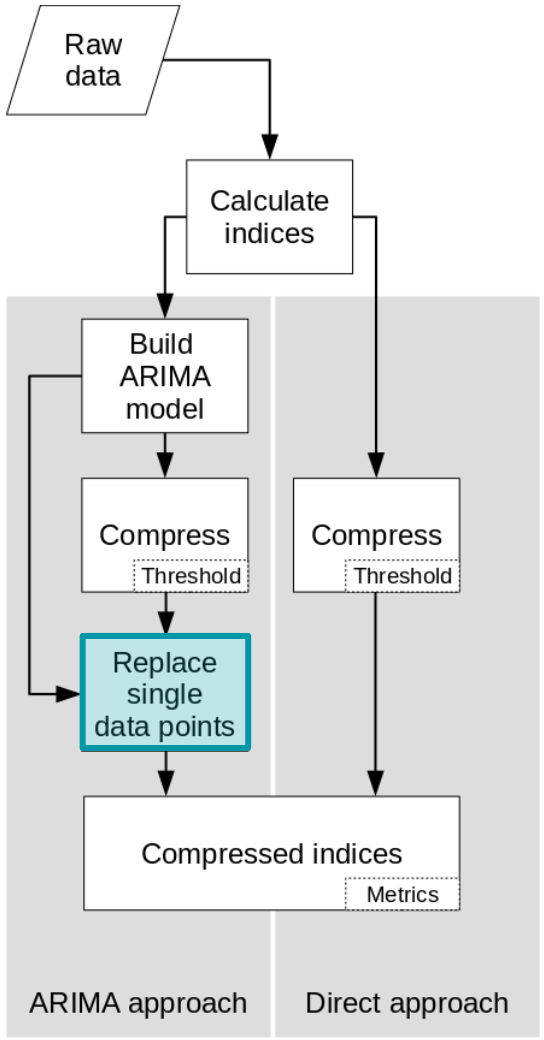
Methods for finding data points to be replaced:





Exp. 3:
Lossy compression (w/ replacement)

- Special
  - Calculate $Corr_{1,i}$
  - Sort from low to high
  - Replace values contributing to lowest $Corr_{1,n}$
- Rollling
  - Calculate windowed $Corr_{j\text{-}bs,j}$
  - Sort from low to high
  - Replace values contributing to lowest $Corr_{1,n}$
- Cumul.Corr.
  - Calculate $Corr_{1,i}$
  - Identify datum with biggest drop
  - Replace values contributing to identified datum



Exp. 3: Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

	Correlation Coefficient (Monthly, 10%, Method: Special)				
	zfp02	zfp03	zfp04	zfp05	zfp06
NAO	.354	.725	.924	.979	.994
+1 Bit					
+2 Bit					
+3 Bit					
QBO30	.139	.482	.635	.972	.986
+1 Bit					
+2 Bit					
+3 Bit					

Exp. 3: Lossy compression (w/ replacement)

Karlsruher Institut für Technologie

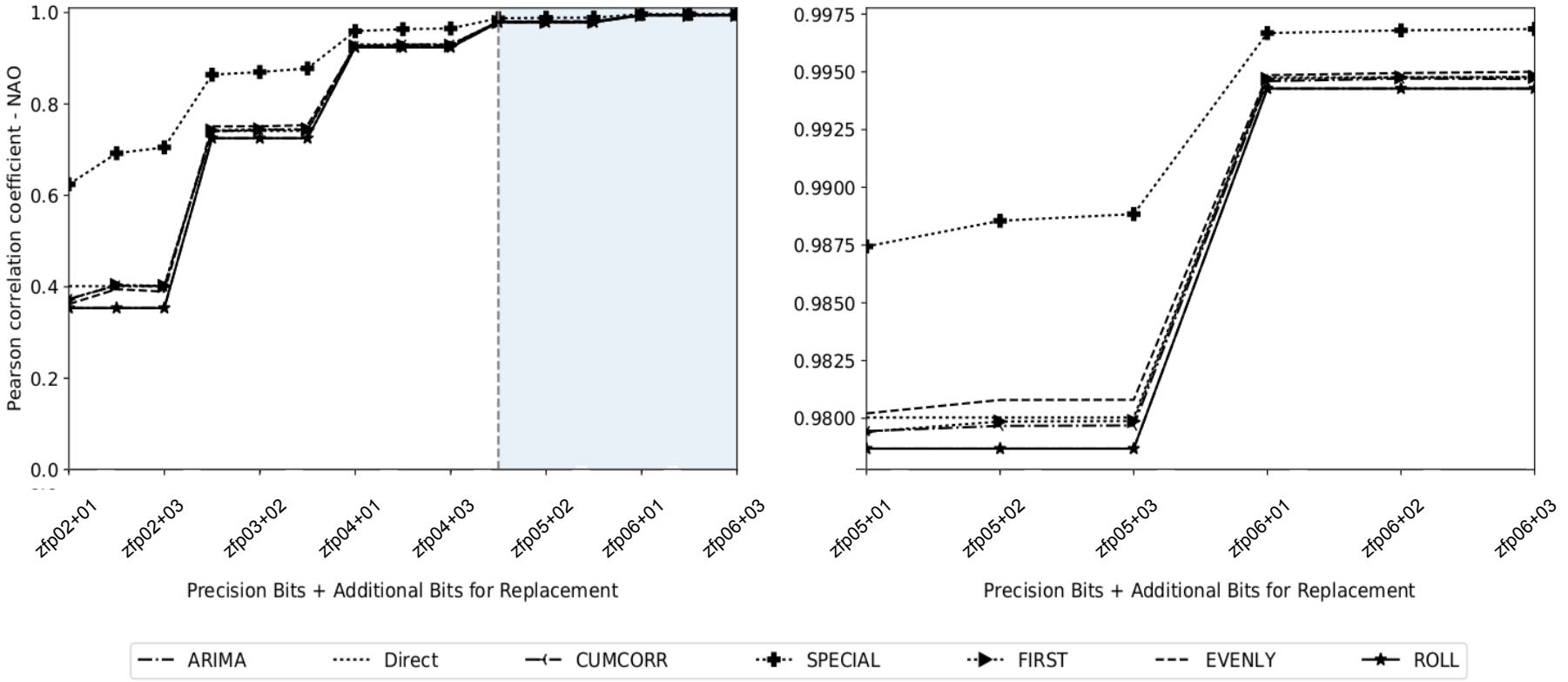
	Correlation Coefficient (Monthly, 10%, Method: Special)				
	zfp02	**zfp03**	**zfp04**	**zfp05**	**zfp06**
NAO	.354	.725	.924	.979	.994
+1 Bit	.624	.864	.959	.987	.997
+2 Bit	.692	.870	.964	.989	.997
+3 Bit	.705	.878	.965	.989	.997
QBO30	.139	.482	.635	.972	.986
+1 Bit	.050	.575	.935	.968	.996
+2 Bit	.082	.615	.940	.973	.993
+3 Bit	.084	.607	.944	.987	.996

Exp. 3: Lossy compression (w/ replacement)

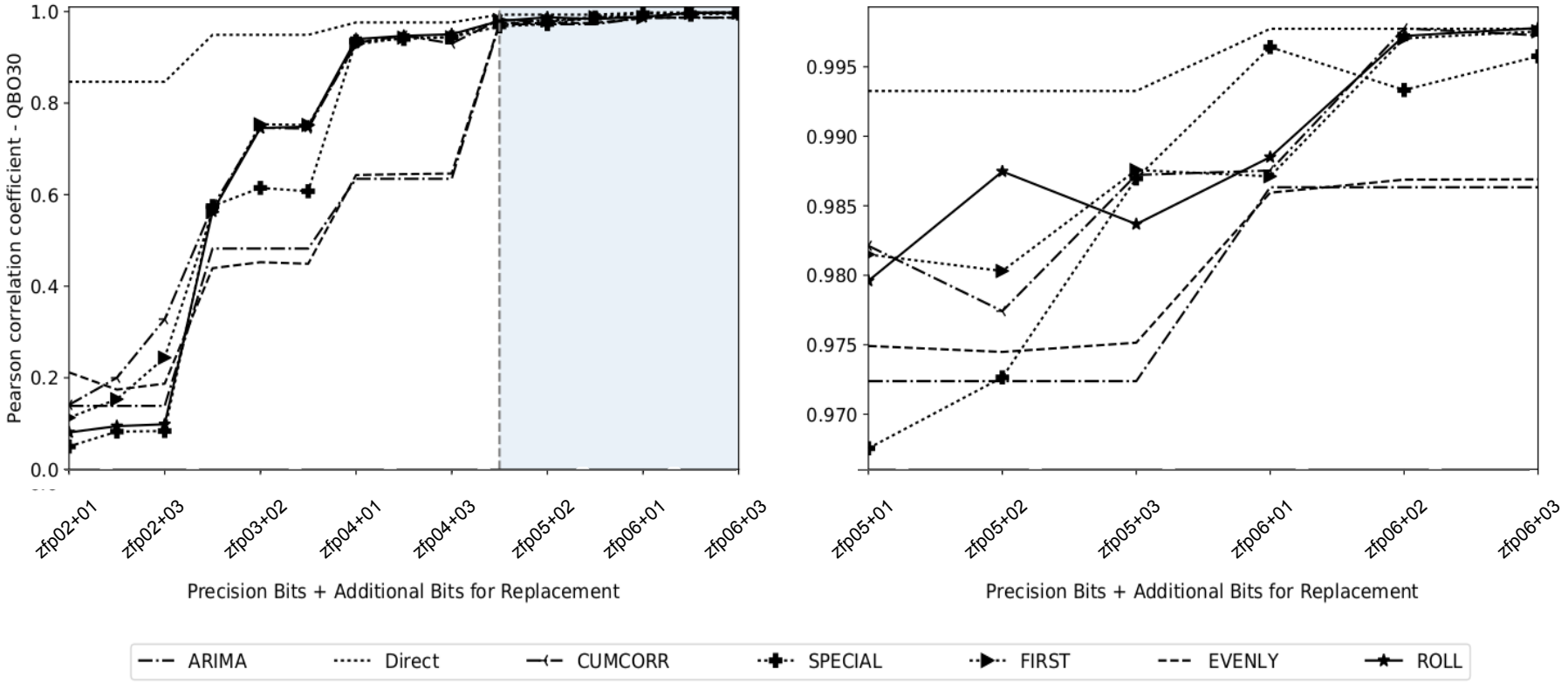
Karlsruher Institut für Technologie

	Correlation Coefficient (Monthly, 10%, Method: Special)				
	zfp02	**zfp03**	**zfp04**	**zfp05**	**zfp06**
NAO	.354	.725 ~15%	.924	.979	.994
+1 Bit	.624 ~25%	.864	.959	.987	.997
+2 Bit	.692	.870	.964	.989	.997
+3 Bit	.705	.878	.965	.989	.997
QBO30	.139 -10%	.482	.635 ~30%	.972 -0.4%	.986
+1 Bit	.050	.575 ~10%	.935	.968	.996 -0.3%
+2 Bit	.082	.615 -1%	.940	.973	.993
+3 Bit	.084	.607	.944	.987	.996

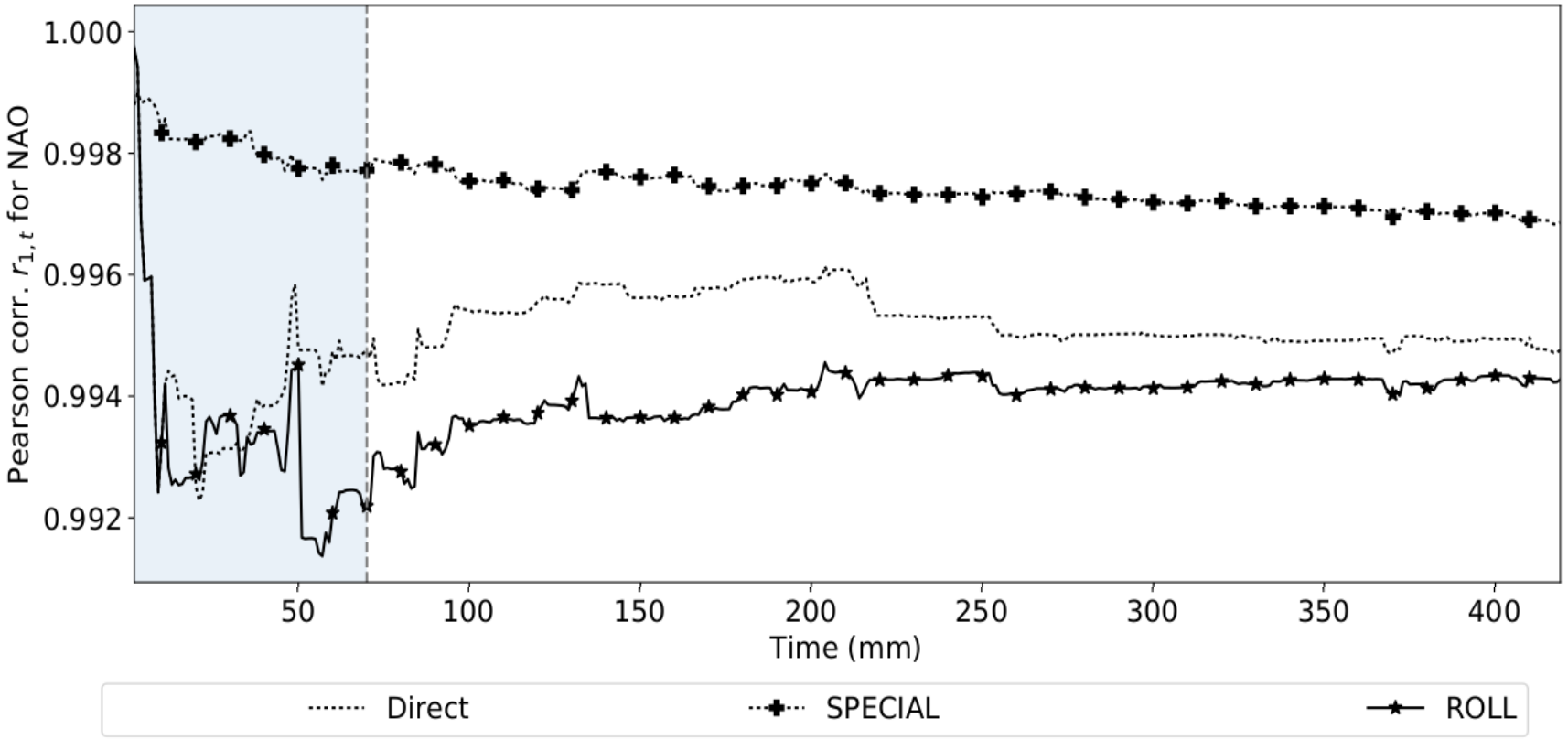
Exp. 3: Lossy compression (w/ replacement)
NAO



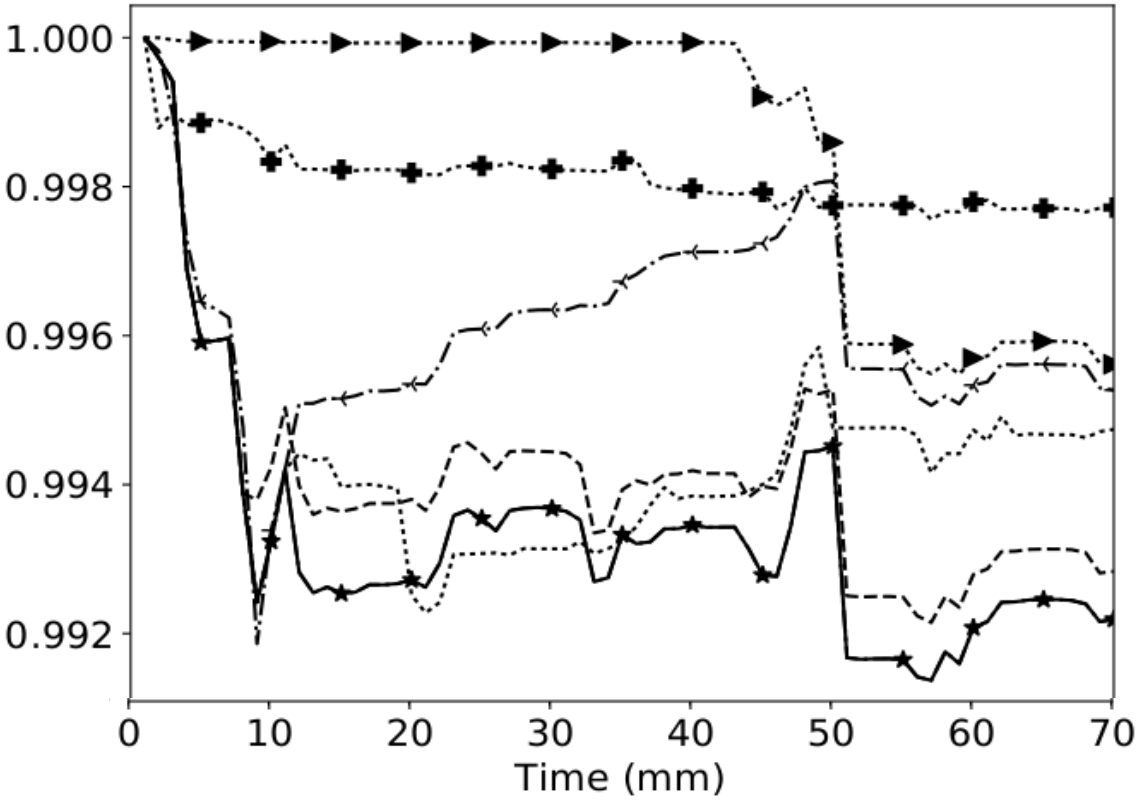
Exp. 3: Lossy compression (w/ replacement)
QBO30



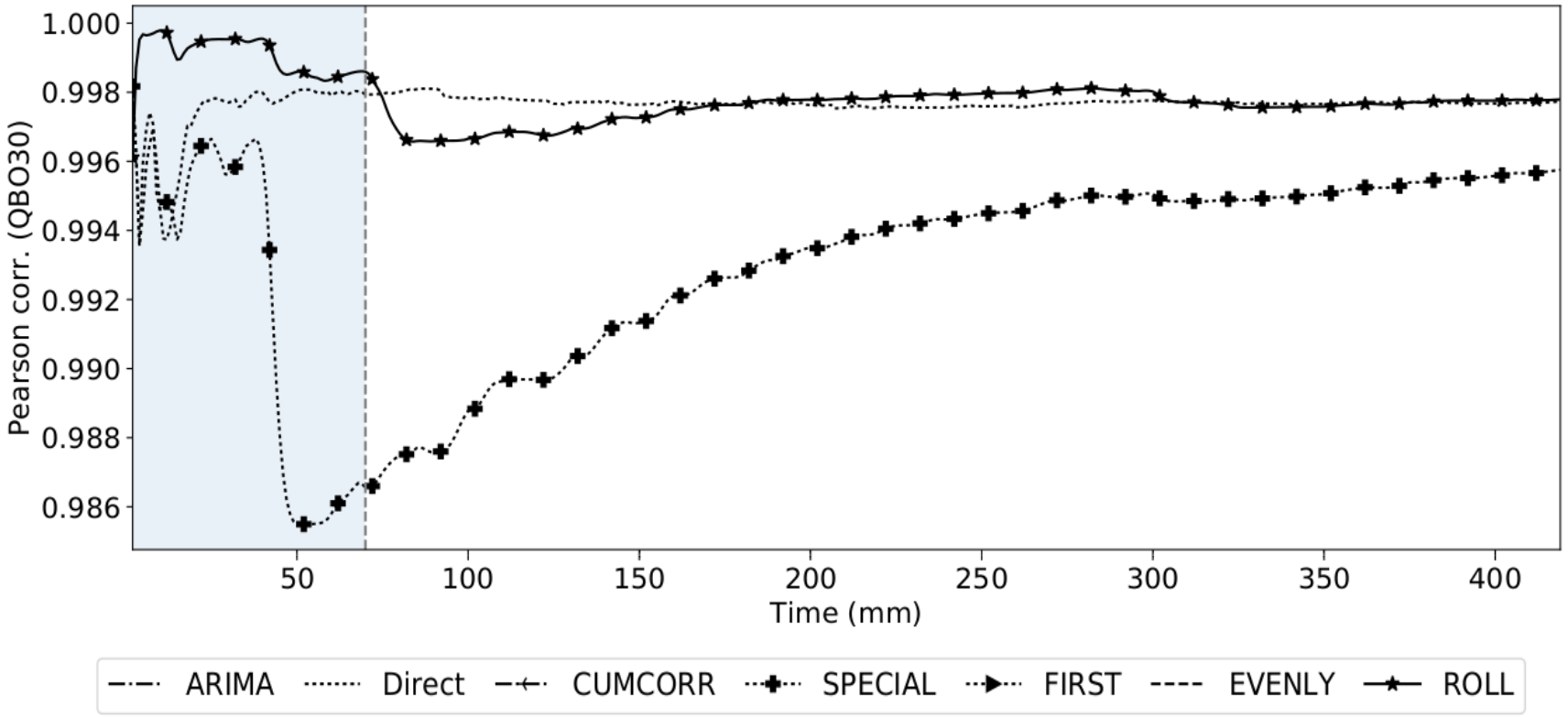
Exp. 3: Lossy compression (w/ replacement) over time series NAO



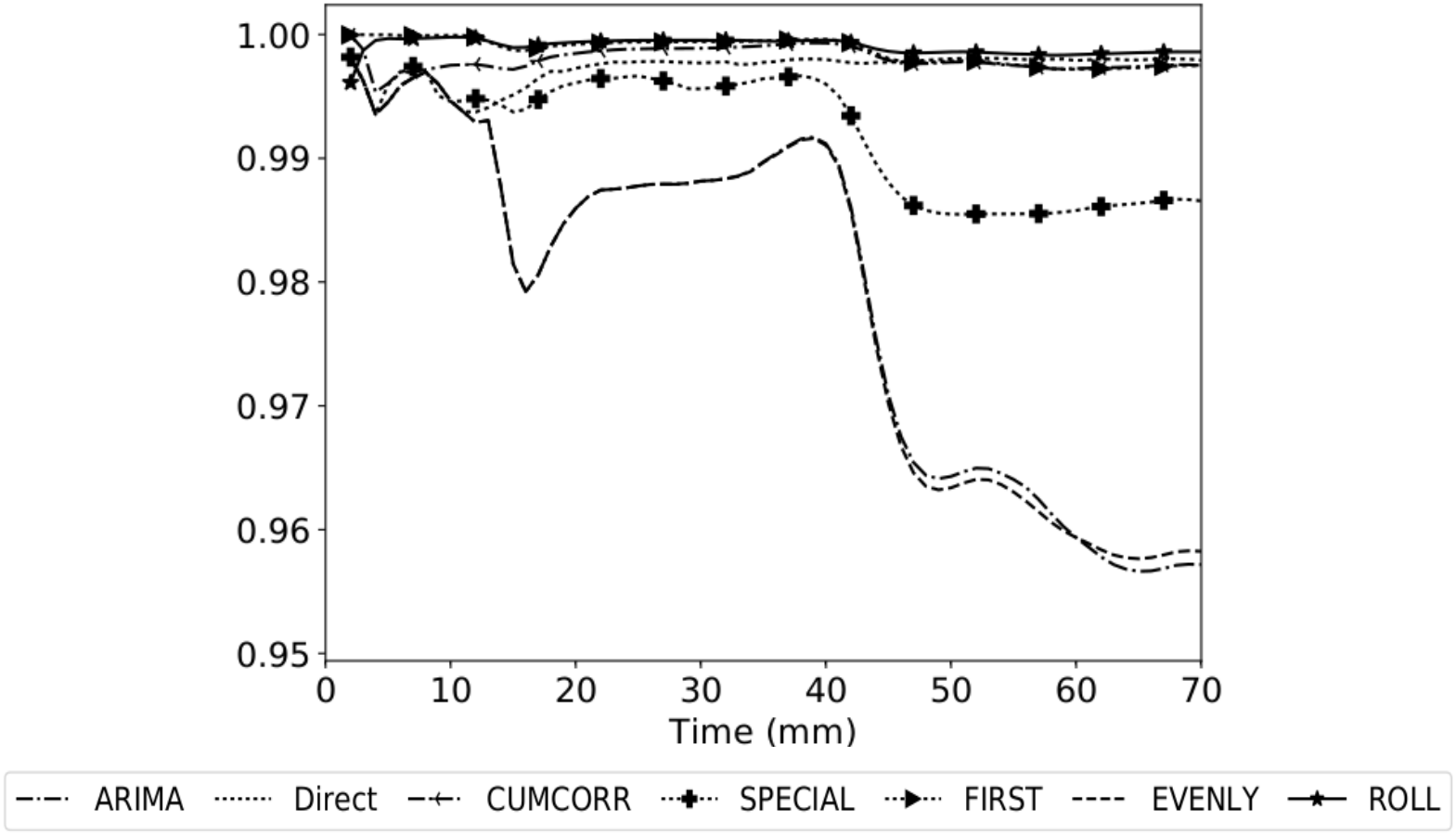
Exp. 3: Lossy compression (w/ replacement) over time series NAO



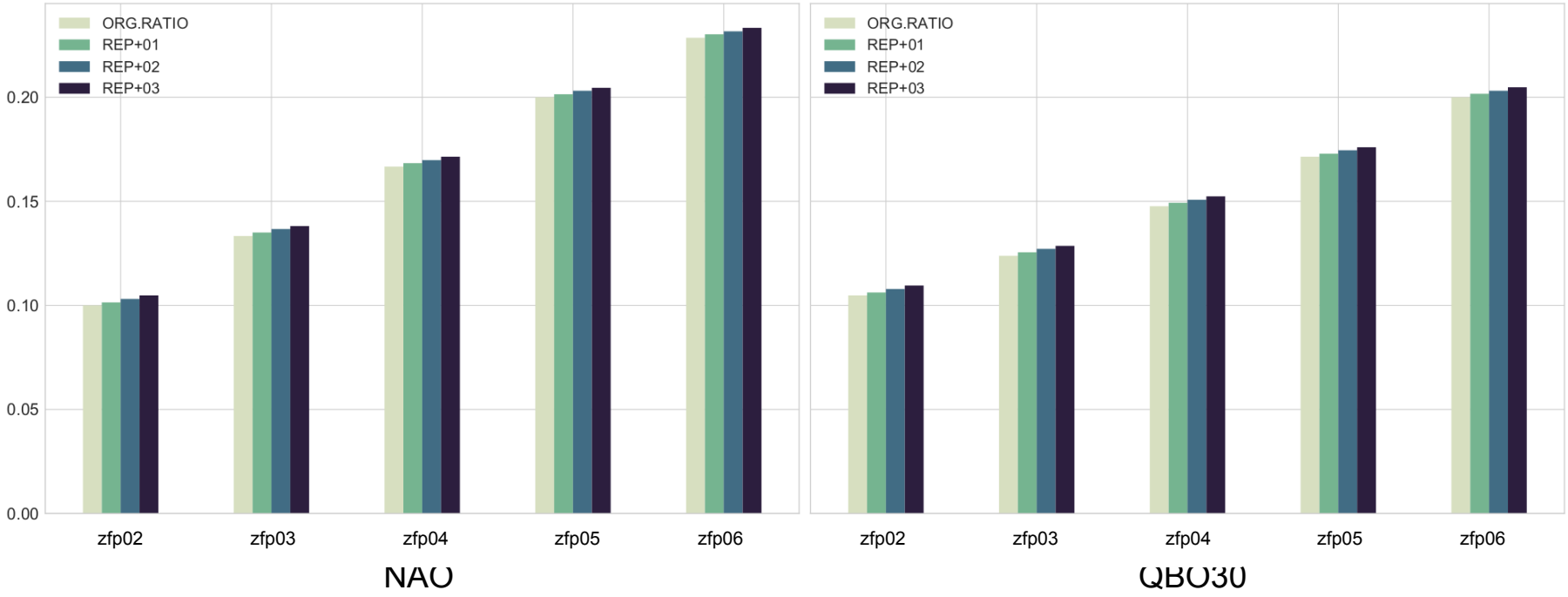
Exp. 3: Lossy compression (w/ replacement) over time series QBO30



Exp. 3: Lossy compression (w/ replacement) over time series QBO30



Exp. 3: Lossy compression (w/ replacement) compression ratio



Summary

Conclusion

- It is **possible to improve quality** of the reconstructed data by replacing several data points with slightly higher precision.
- ARIMA models using a **differentiation step** have difficulties and performed worse than other models.
- Time series expressed with **small auto-regressive and moving-average** order can be improved significantly.

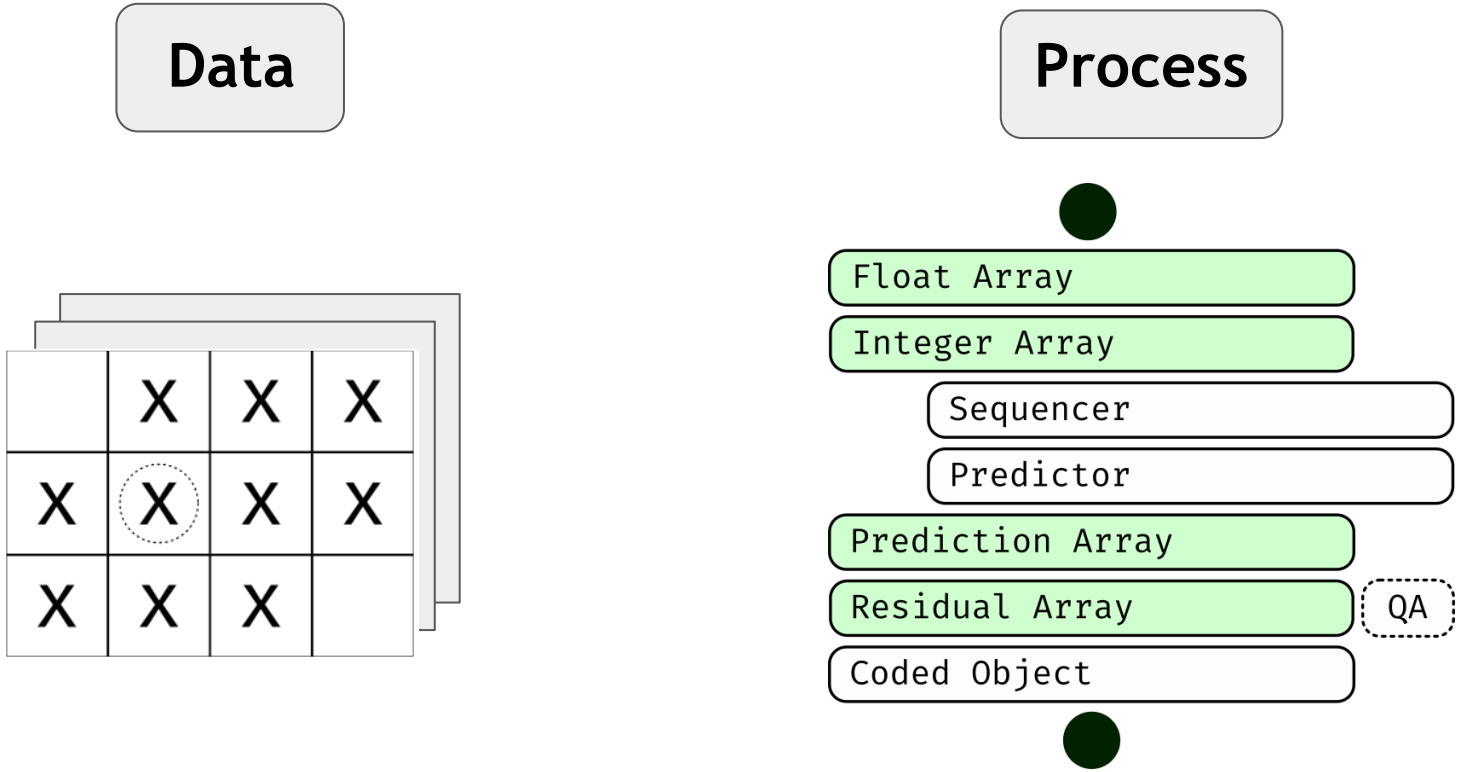
Further analysis

- Further analysis will focus on why certain time series (like QBO30) **do not show the same improvement** like NAO
- Analyse why there is sometimes **loss in quality** using higher precision data.
- It is a **complementary method** to the direct approach and will not replace it *yet*.

4. Identifying new patterns and relationships within and between variables for the current data

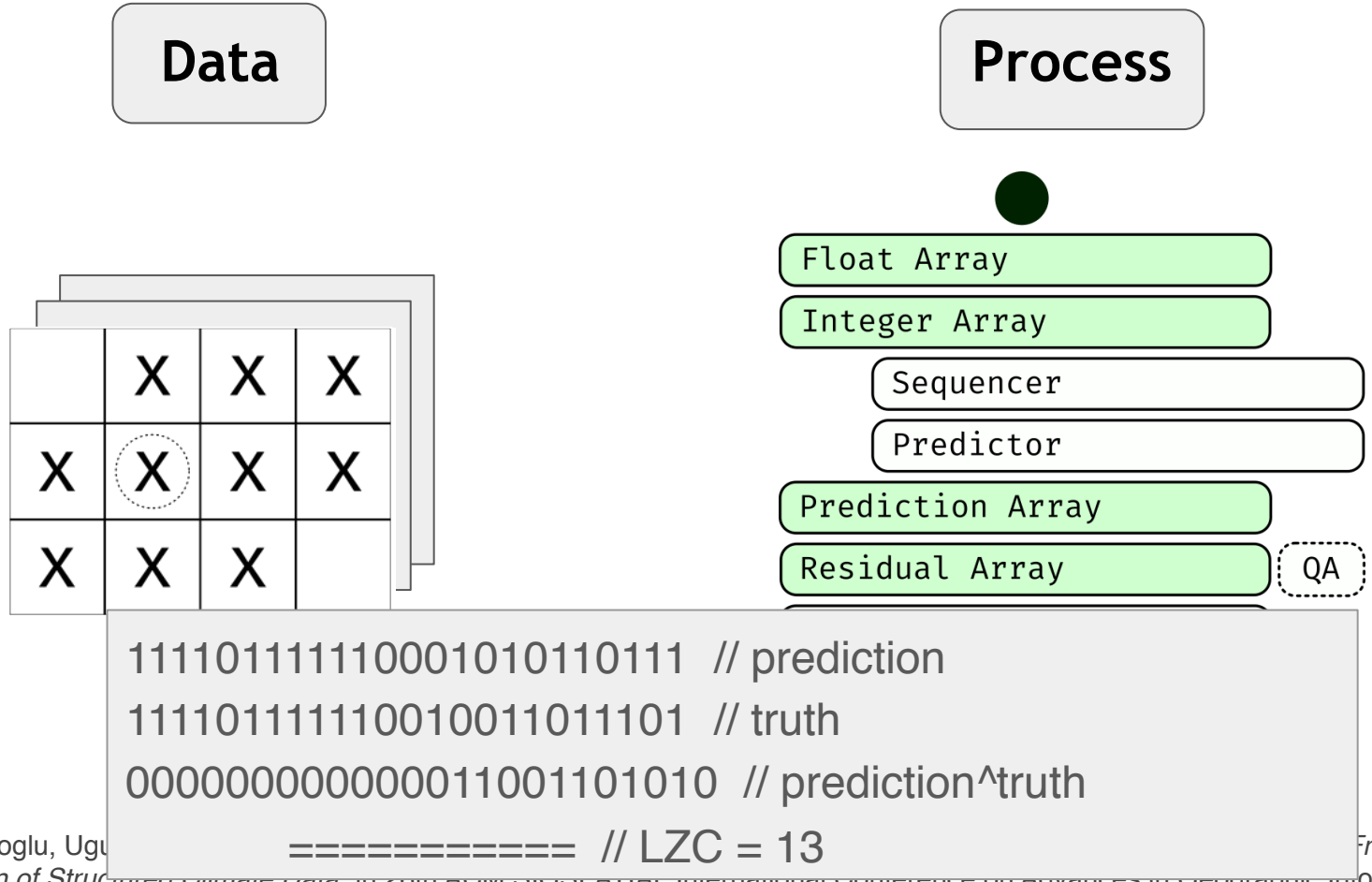
Step 4

Prediction-based compression process



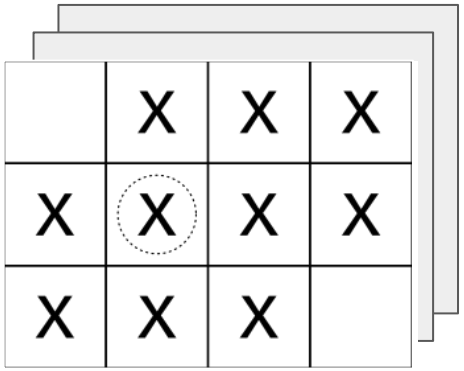
Source: Cayoglu, Ugur and Schröter, Jennifer and Meyer, Jörg and Streit, Achim and Braesicke, Peter "*A Modular Software Framework for Compression of Structured Climate Data*" in 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2018.

Prediction-based compression process



Source: Cayoglu, Ugu ... *Framework for Compression of Structured Climate Data* ... Systems 2018.

Prediction-based compression process



- Decisions to be made:
  - How do we want to traverse the data?
  - Where do we want to start?
  - Which neighbouring points we want to use for prediction?
  - How do we want to calculate the predictions?

Contributions

Karlsruher Institut für Technologie

- Analysis of different **starting points** and **traversal sequences** and their influence on compression ratio.
- Introduction of new **traversal sequences** and analysis of their influence on compression ratio.
- Definition of **Information Spaces** and usage during the prediction process.
- Introduction of **consolidation techniques** for further improvement of the compression ratio.

Contributions (and the take home messages!)

Karlsruher Institut für Technologie

- Analysis of different **starting points** and **traversal sequences** and their influence on compression ratio.
- Introduction of new **traversal sequences** and and analysis of their influence on compression ratio.
- Definition of **Information Spaces** and usage during the prediction process.
- Introduction of **consolidation techniques** for further improvement of the compression ratio.

little influence on compression rate

major influence on compression rate

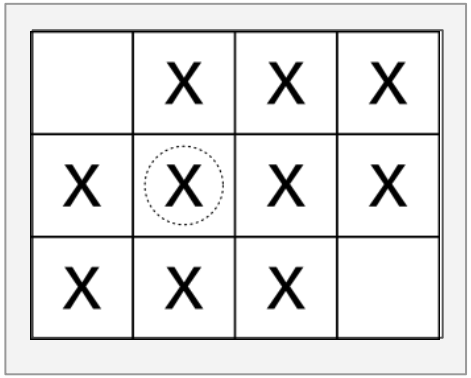
stabilize the compression rate by reducing structure dependency

further improve prediction by understanding biases

What are Information Spaces?

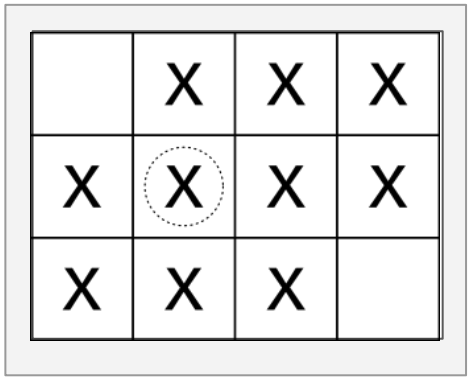
Karlsruher Institut für Technologie

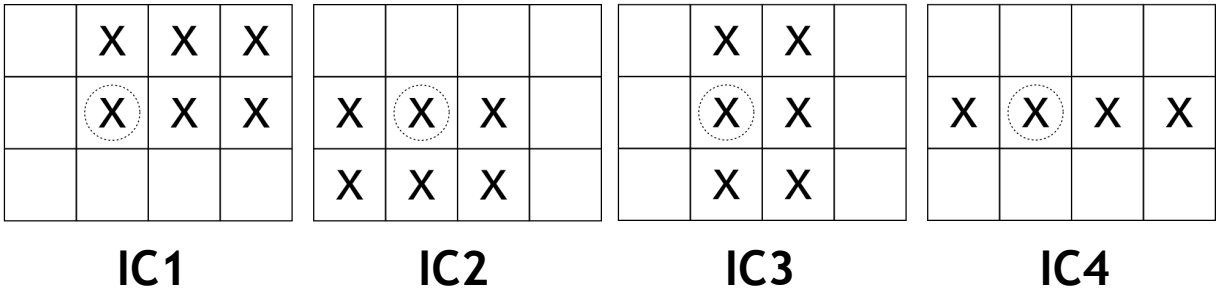
- Information Space:
  - All **available information** for the prediction of a data point



What are Information Spaces?

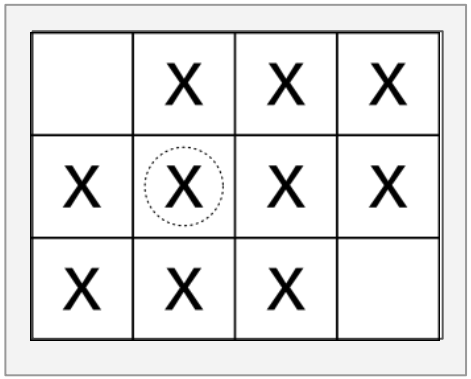
- Information Space:
  - All **available information** for the prediction of a data point
  - **Composed of** smaller groups of information called **Information Context (IC)**

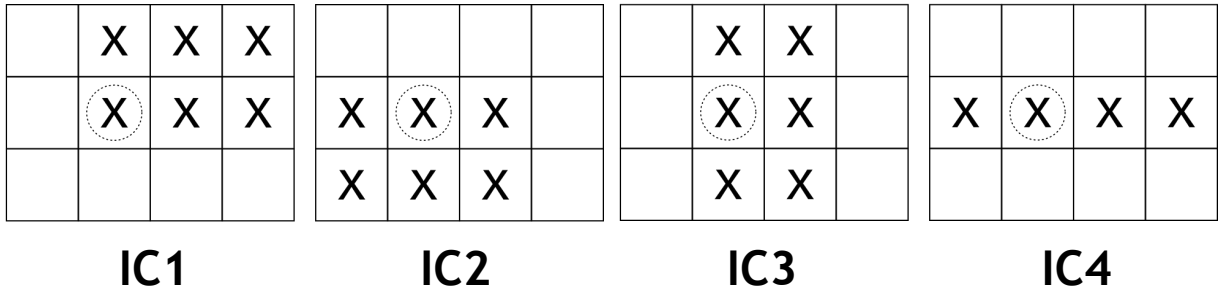




What are Information Spaces?

- Information Space:
  - All **available information** for the prediction of a data point
  - **Composed of** smaller groups of information called **Information Context (IC)**
- Information Context:
  - Splitting of the Information Space in **blocks of subsets**
  - Lagrange interpolation problem

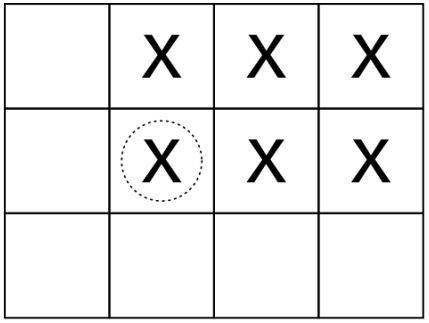




Predictions using neighbouring values in multi-dimensional space.

Karlsruher Institut für Technologie

$$f(x,y) = \sum_i \sum_j w_{ij} \cdot f(x+i, y+j) \qquad i,j \in \mathbb{Z}$$



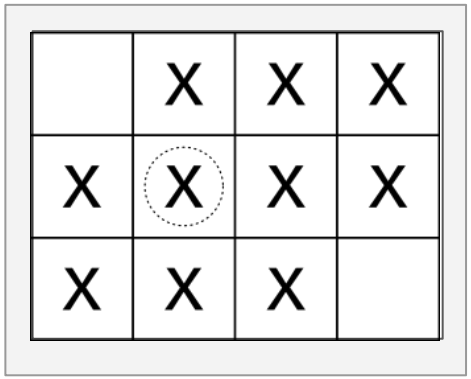
IC1

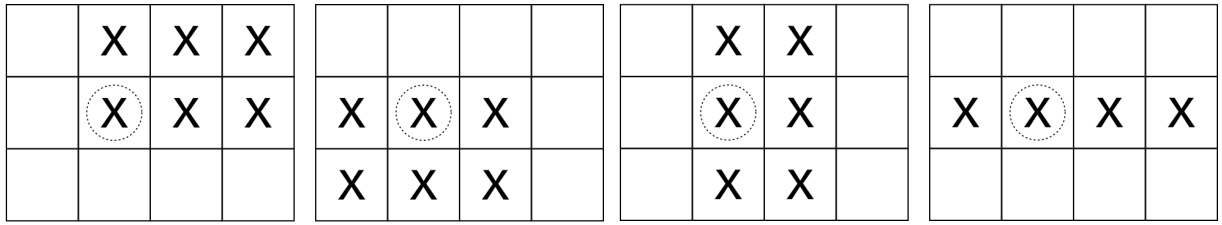
- Known interpolation problem in mathematics:
  - **Lagrange** (considering only value)
  - **Hermite** (value + derivative interpolation) and
  - **Birkhoff** (value + derivative incl. missing values interpolation)
- No general solution: It is **computationally difficult** and solutions rely on **floating-point arithmetics.**
  - **Integer arithmetic** solutions exists for the Lagrange problem iff neighbouring values are **arranged in blocks**.

Consolidation methods for final prediction from different information contexts.

Karlsruher Institut für Technologie

But we still have the problem of choosing which IC to use for the best prediction.



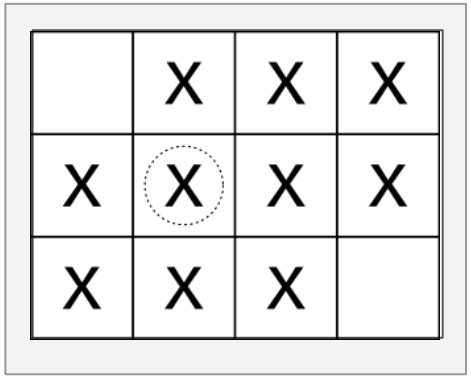


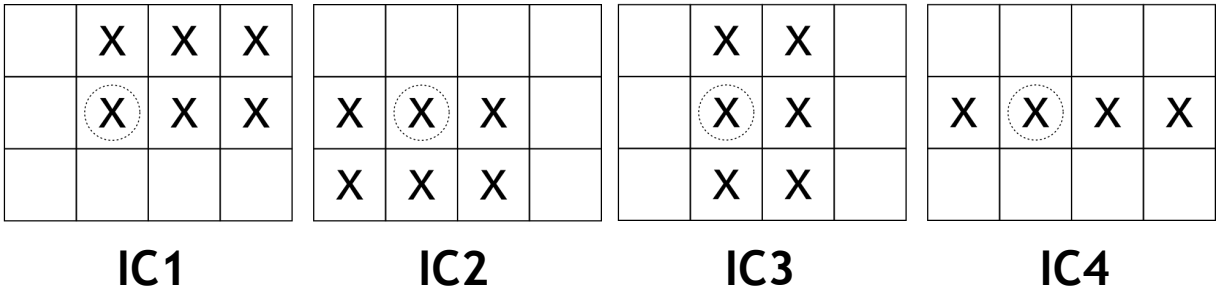
Consolidation methods for final prediction from different information contexts.

Karlsruher Institut für Technologie

But we still have the problem of choosing which IC to use for the best prediction. This is where the consolidation methods come into play:

- Average of all prediction candidates
- Prediction of the best predictor for the previous element
- Maximum or minimum
- User-defined enforced prediction





Experimental setup

Karlsruher Institut für Technologie

Data

- Climate simulation output by ECHAM/MESSy model with 128 x 64 x 47 x *time* resolution (32 bit floating-point data)
  - January, 2013 with 74 timesteps (every 10 hours)
  - 2013 with 365 timesteps (daily mean)
  - 2000-2013 with 168 timesteps (monthly mean)
- Intel Xeon (E5-2640 v2) with 16 cores and 128 GB memory
- Experiments were conducted on randomly selected chunks

Metrics

- Leading Zero Count (LZC)
- Compression Ratio (CR)

Experiments

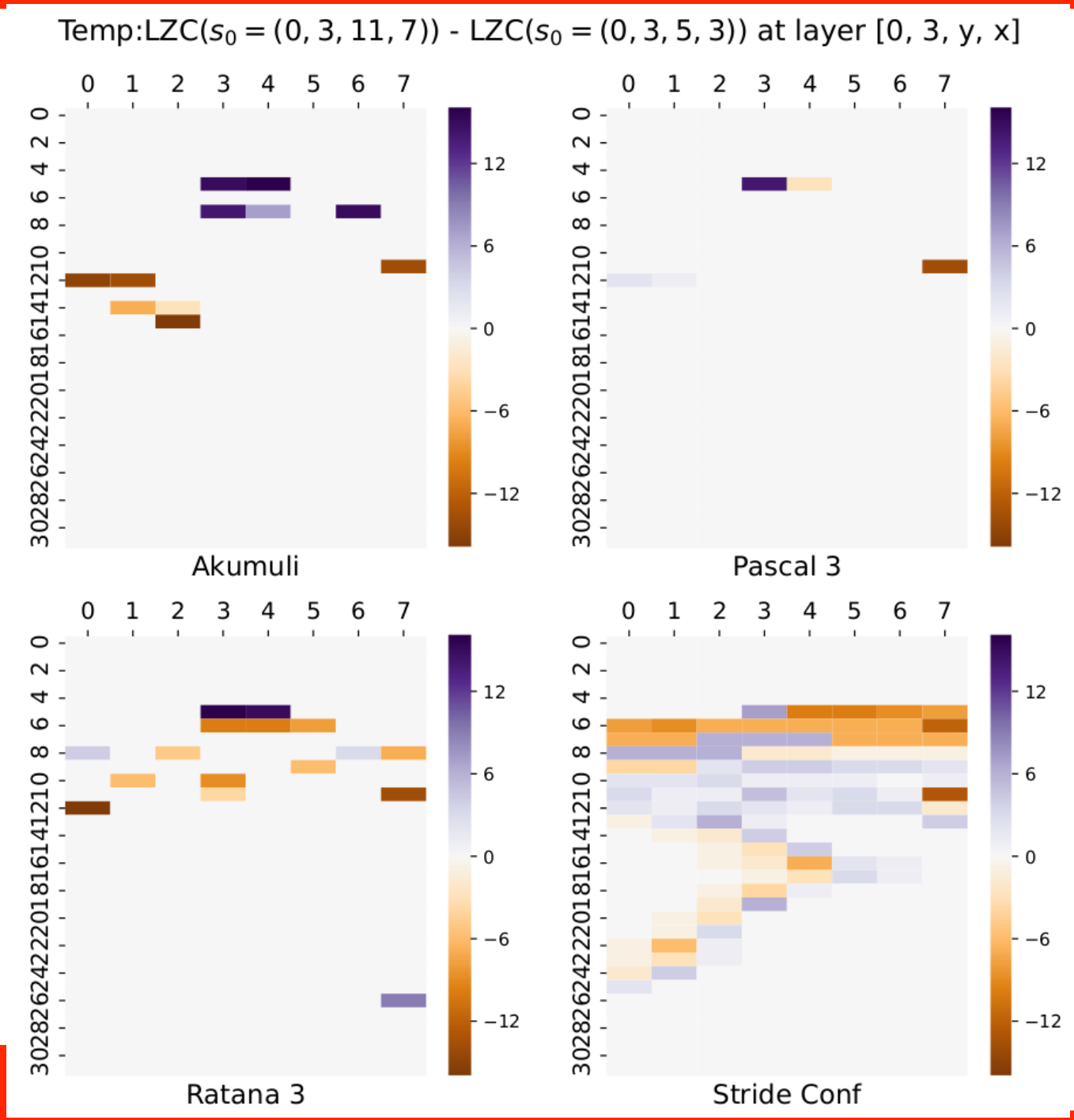
Karlsruher Institut für Technologie

- Influence of starting point:
  - Randomly select ten different starting points per chunk
- Traversal order of dimensions:
  - Compare linear traversal with rotated chunks
  - Compare linear traversal with block, chequerboard and blossom traversal
- Traversal order with the use of Information Spaces:
  - Experiment 2 including the use of Information Spaces
- Consolidation techniques to discover bias of the predictor:
  - Average, Minimum, Maximum, Last Best, Reinforced (ordering defined by user)

Influence of starting point

Karlsruher Institut für Technologie

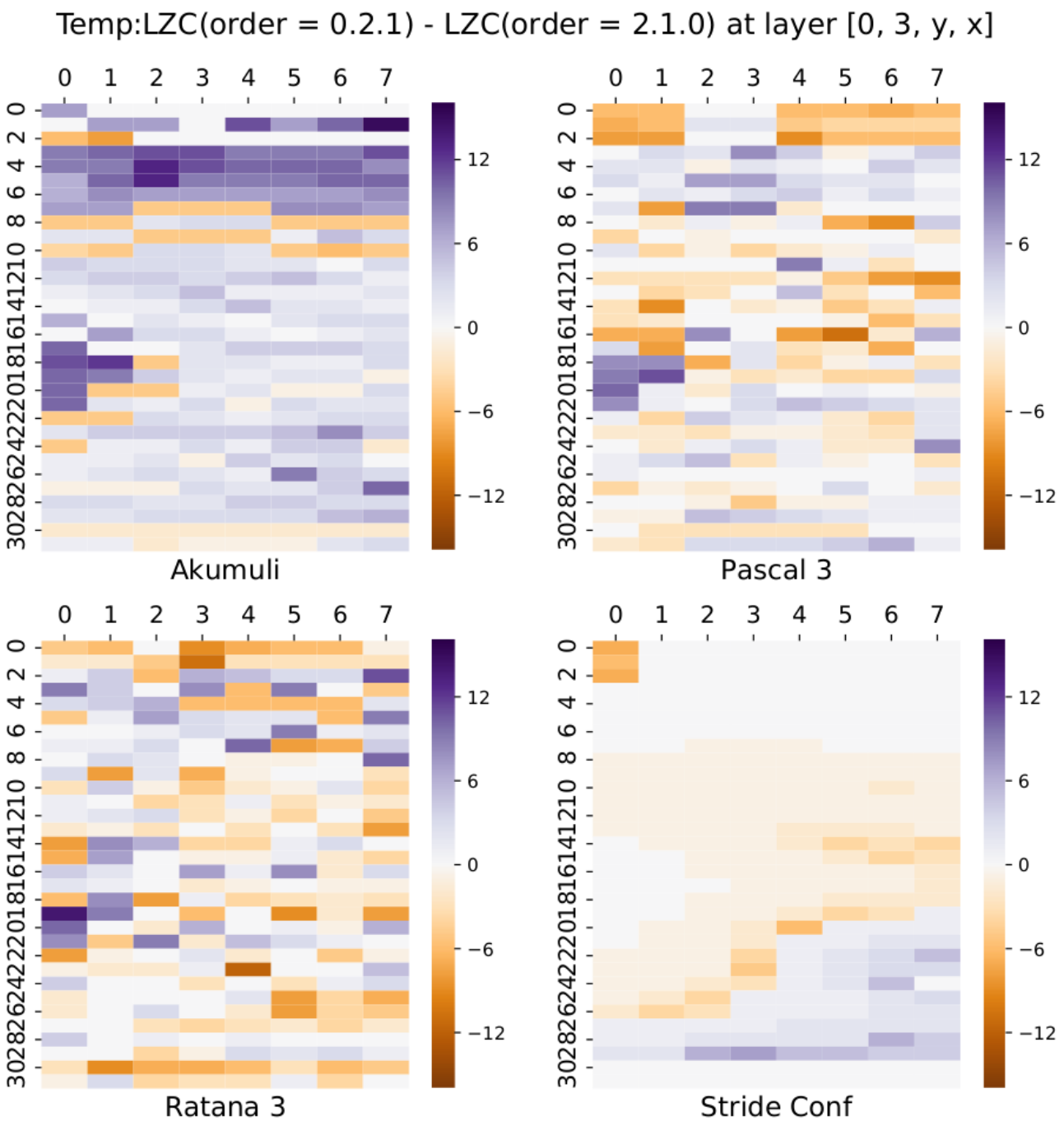
1. Randomly select ten different starting points per chunk
2. Compress each chunk using linear traversal
3. Calculate predictions using different predictors:
   - Akumuli, Stride, Pascal, Ratana, …
4. Difference comparison of the LZC for all data points



Traversal order of dimensions

Karlsruher Institut für Technologie

1. Set starting value to (0,0,0,0) of each chunk
2. Traverse each chunk using linear traversal
3. Calculate predictions using different predictors
4. Reshape each chunk e.g.
   (5,4,7) > {(7,5,4),(4,7,5),(7,4,5),...}
5. Traverse each chunk using linear traversal
6. Calculate predictions using different predictors for reshaped chunks
7. Difference comparison of the LZC for all data points


Temp:LZC(order = 0.2.1) - LZC(order = 2.1.0) at layer [0, 3, y, x]
Akumuli
Pascal 3
Ratana 3
Stride Conf

Traversal order of dimensions

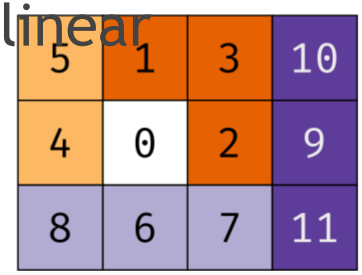
Karlsruher Institut für Technologie

Influence of traversal direction

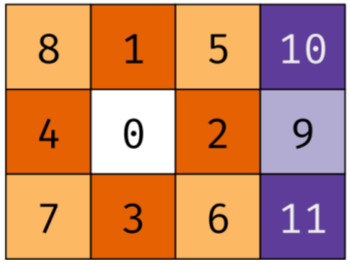
1. Set starting value to (0,0,0,0) of each chunk
2. Traverse each chunk using linear along the dimensions
3. Reshape each chunk e.g.
   (5,4,7) > {(7,5,4),(4,7,5),(7,4,5),...}
4. Traverse each chunk **using new traversal order**
5. Compare standard deviation of CR for all reshaped chunks

block

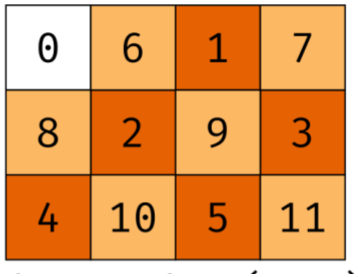
linear

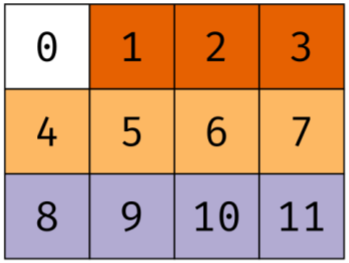


blossom

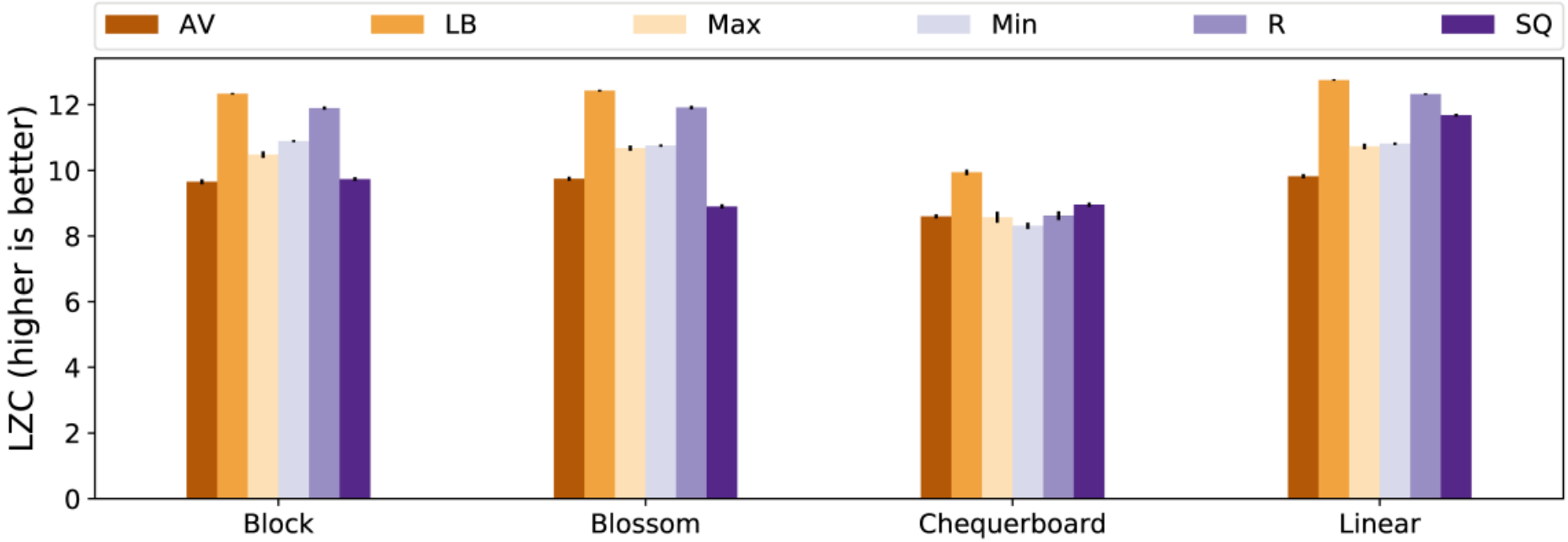


chequerboard





LZC has highest for last best consolidation method.



block

5	1	3	10
4	0	2	9
8	6	7	11

blossom

8	1	5	10
4	0	2	9
7	3	6	11

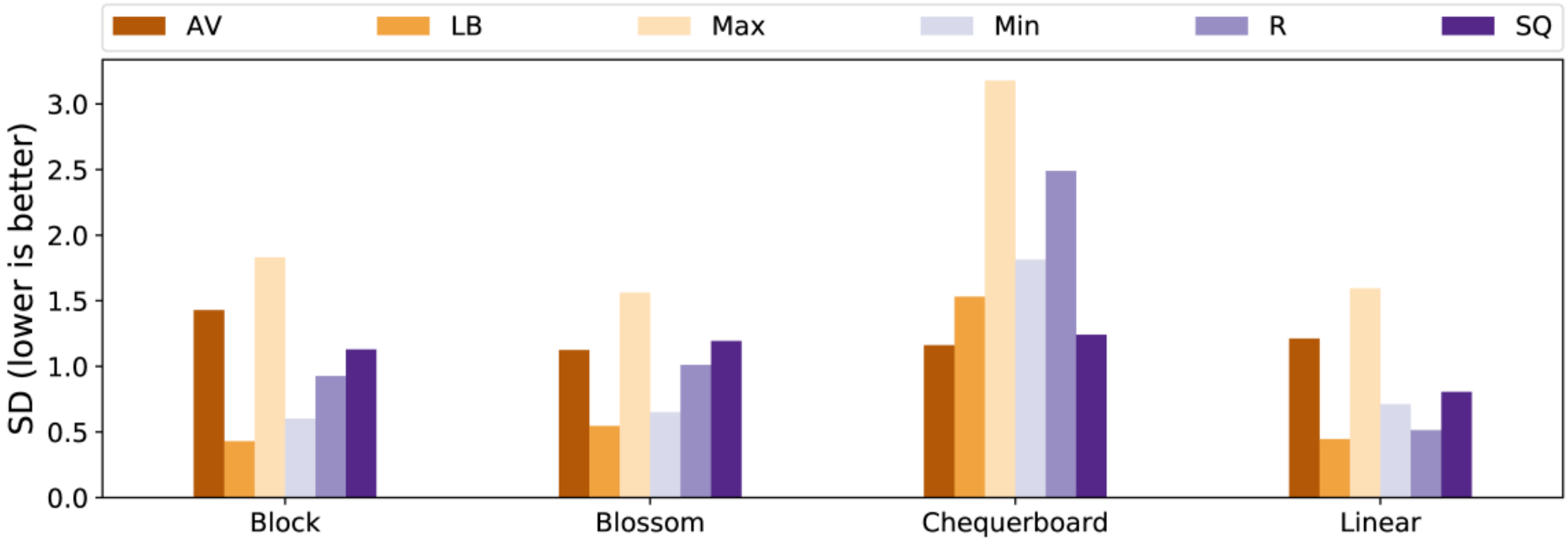
chequerboard

0	6	1	7
8	2	9	3
4	10	5	11

linear

0	1	2	3
4	5	6	7
8	9	10	11

Standard deviation of LZC to express vulnerability of CR for reshaped chunks.

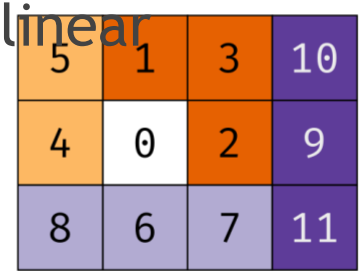


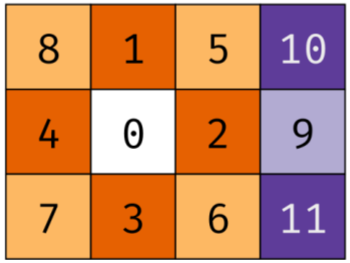
block

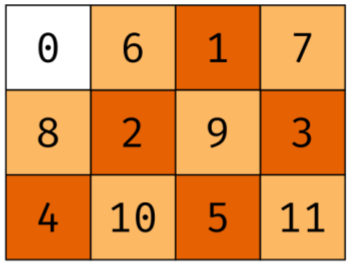
blossom

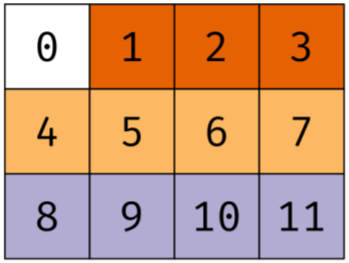
chequerboard

linear

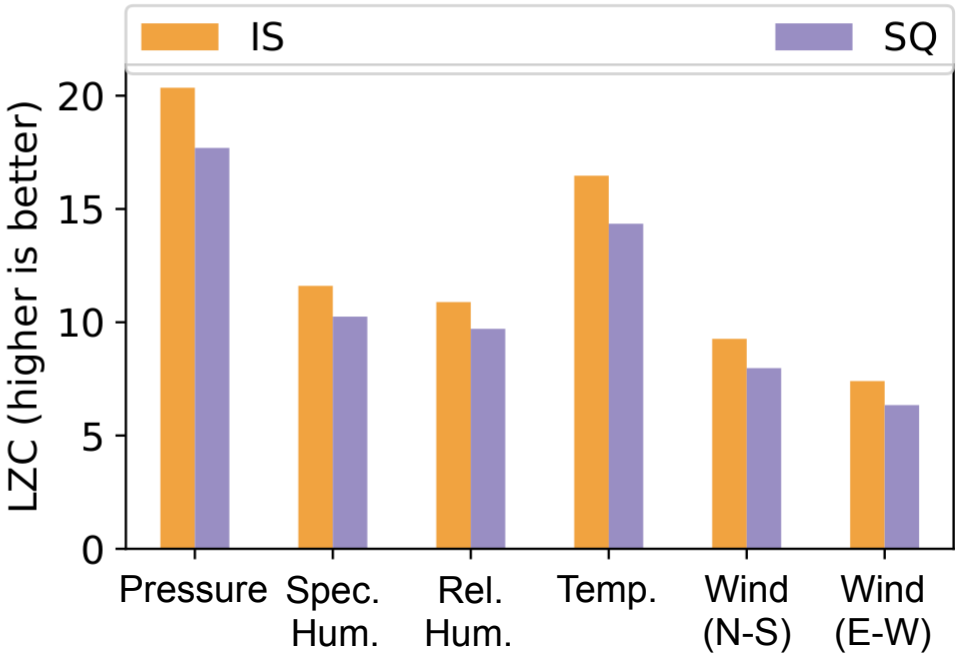








Experimental results show that using information spaces improves compression rates.



Performance of predictors using information spaces (IS) and classic sequential prediction (SQ).

Summary of the results/further research areas

Karlsruher Institut für Technologie

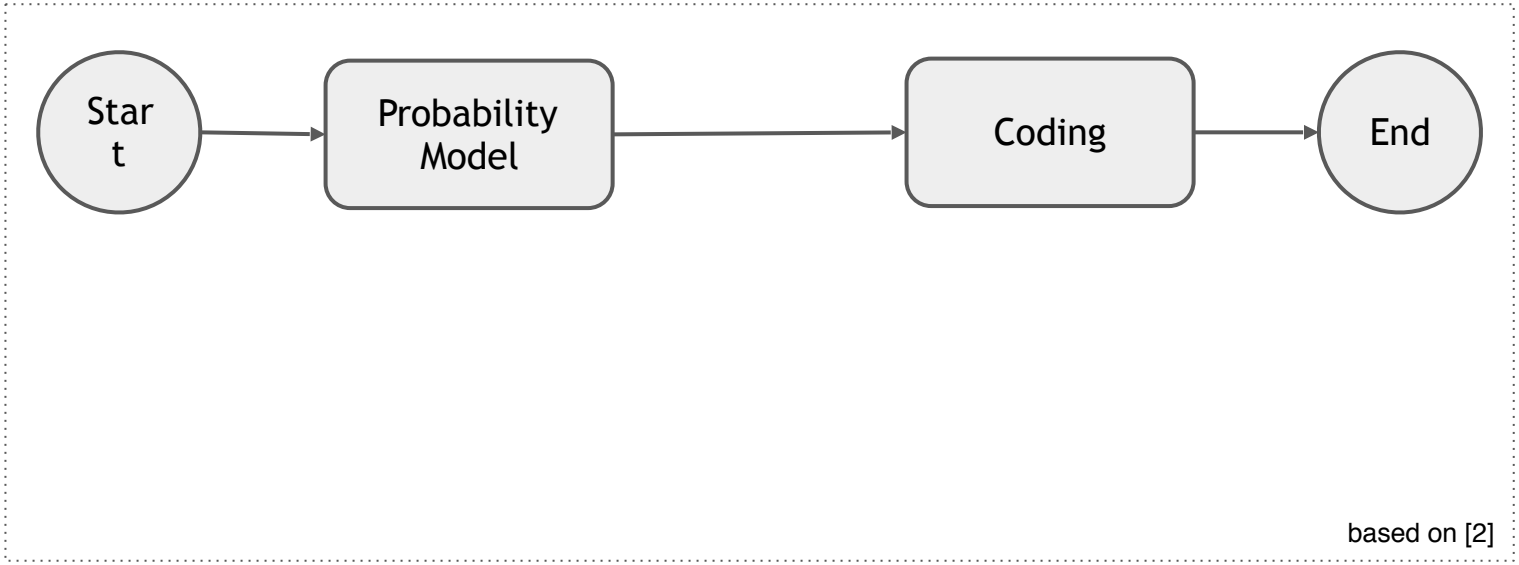
- **Stability** Compression results show independence of data rotation/ positioning
- **Better compression rates** Even in early stages of development the compression rates are higher than before
- **New possibilities** Using new traversal sequences might improve compression rate
- **Not fully utilised (*yet*)**
  - Weights for the different Information Contexts yet to be defined
  - Optimal subgrids for Information Contexts yet to be defined
  - Results for 1D predictors and ND predictors are possible
- **Complexity/Slow** Added complexity for calculation of IS
- **Memory usage** Higher memory usage due to adoption of several concurrent predictors

Karlsruher Institut für Technologie

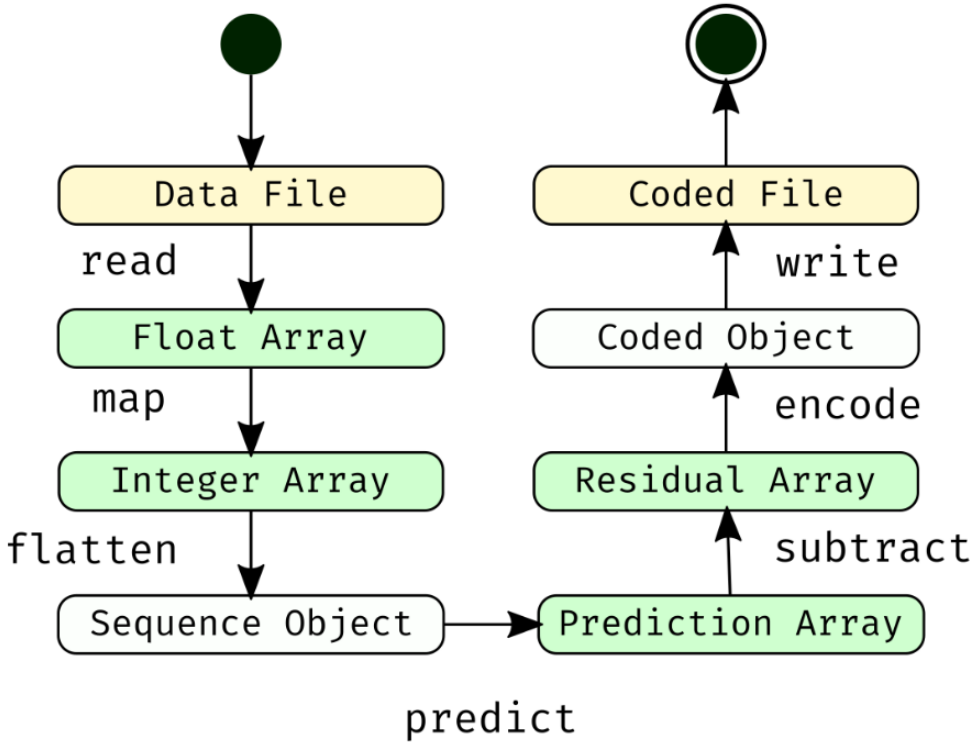
5. Optimizing coding method to write on disk

Step 5

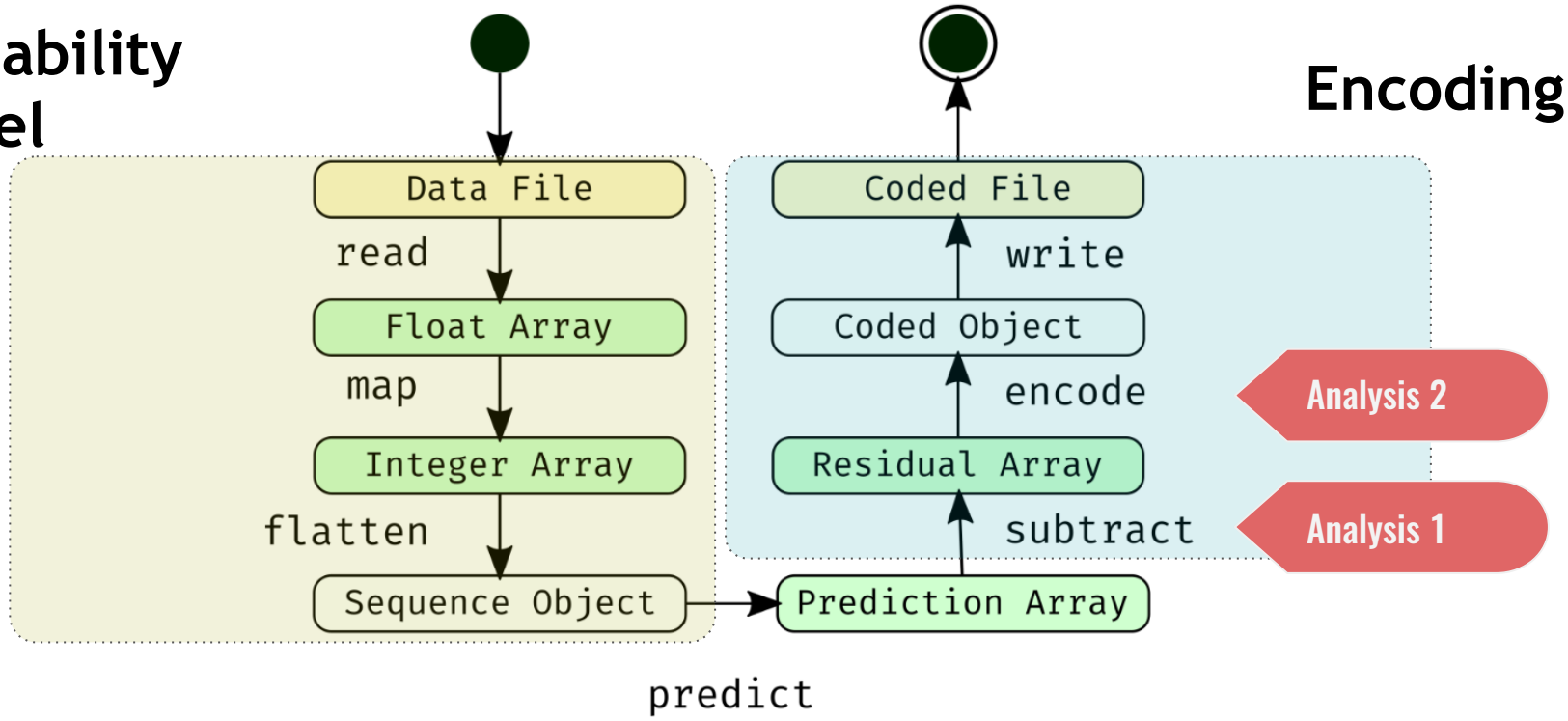
Currently our focus is on the last step of the chain: encoding of the data.

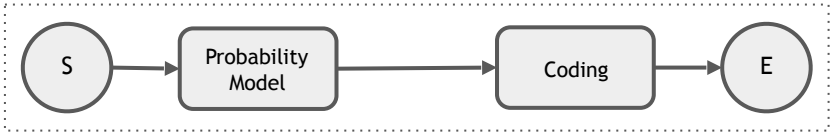


The last three steps are influenced by the encoding method.



The last three steps are influenced by the encoding method.





How to calculate the difference between two floating-point values

Karlsruher Institut für Technologie

Analysis 1

$$\text{diff}_{xor}(p,t) = p \oplus t$$
$$\text{diff}_{abs}(p,t) = \mid p - t \mid$$

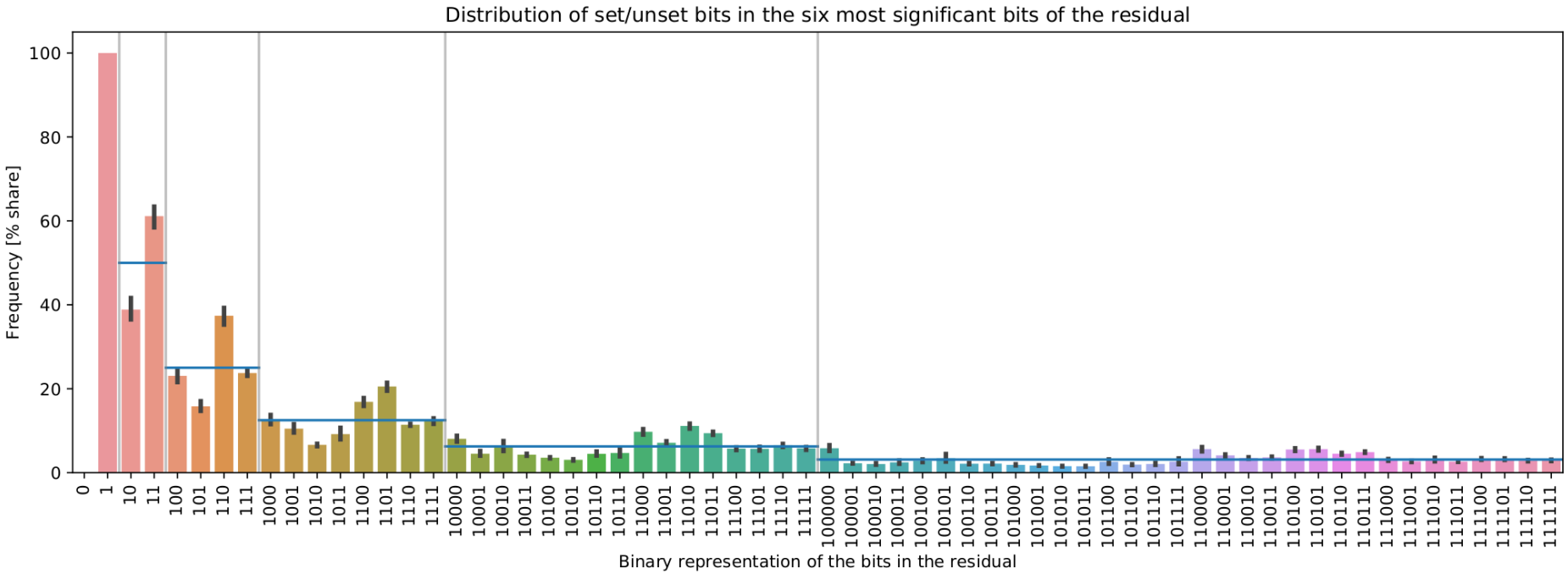
pred: 256.321
true: 255.931
diff: 0.390
diff (xor): 16,762,689

pred: 847,390.837
true: 847,794.417
diff: 403.580
diff (xor): 6,458

```
bin(p=256.321) = 01000011100000000010100100010111
bin(t=255.931) = 01000011011111111110111001010110
         p ⊕ t = 00000000111111111000111010000001
index 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
```

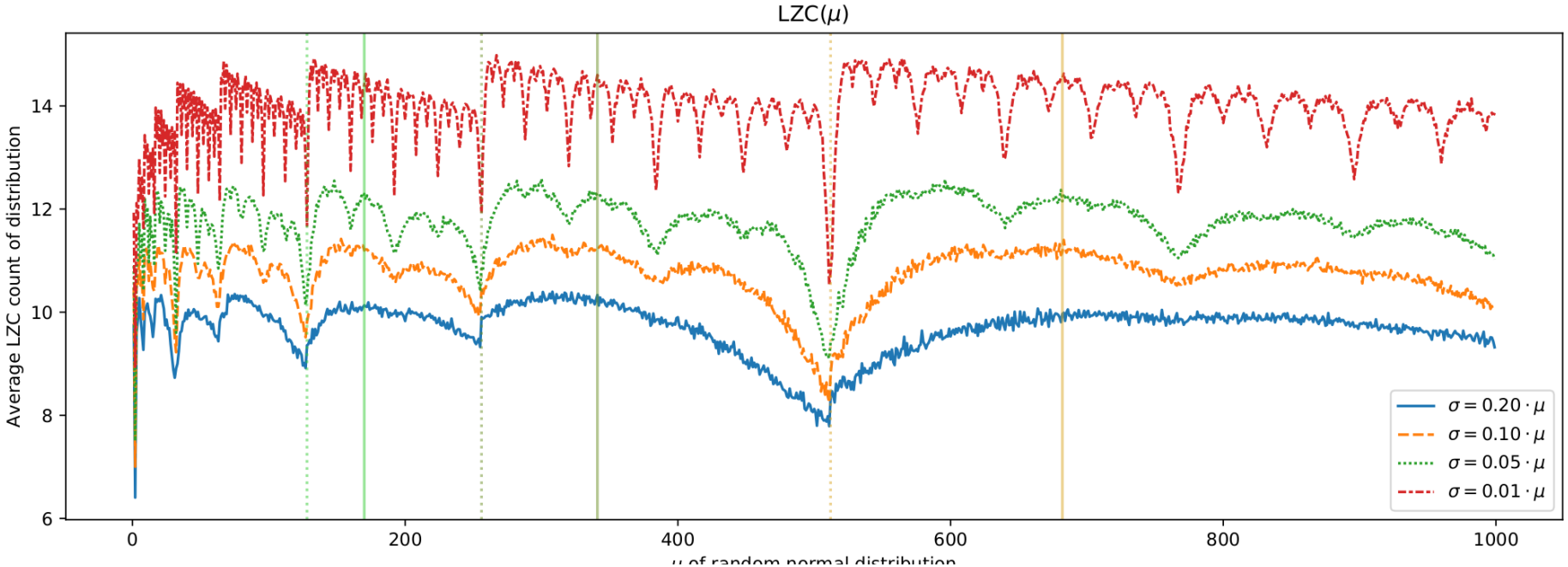
LZC **FOC**

Currently our focus is on the last step of the chain: encoding of the data.

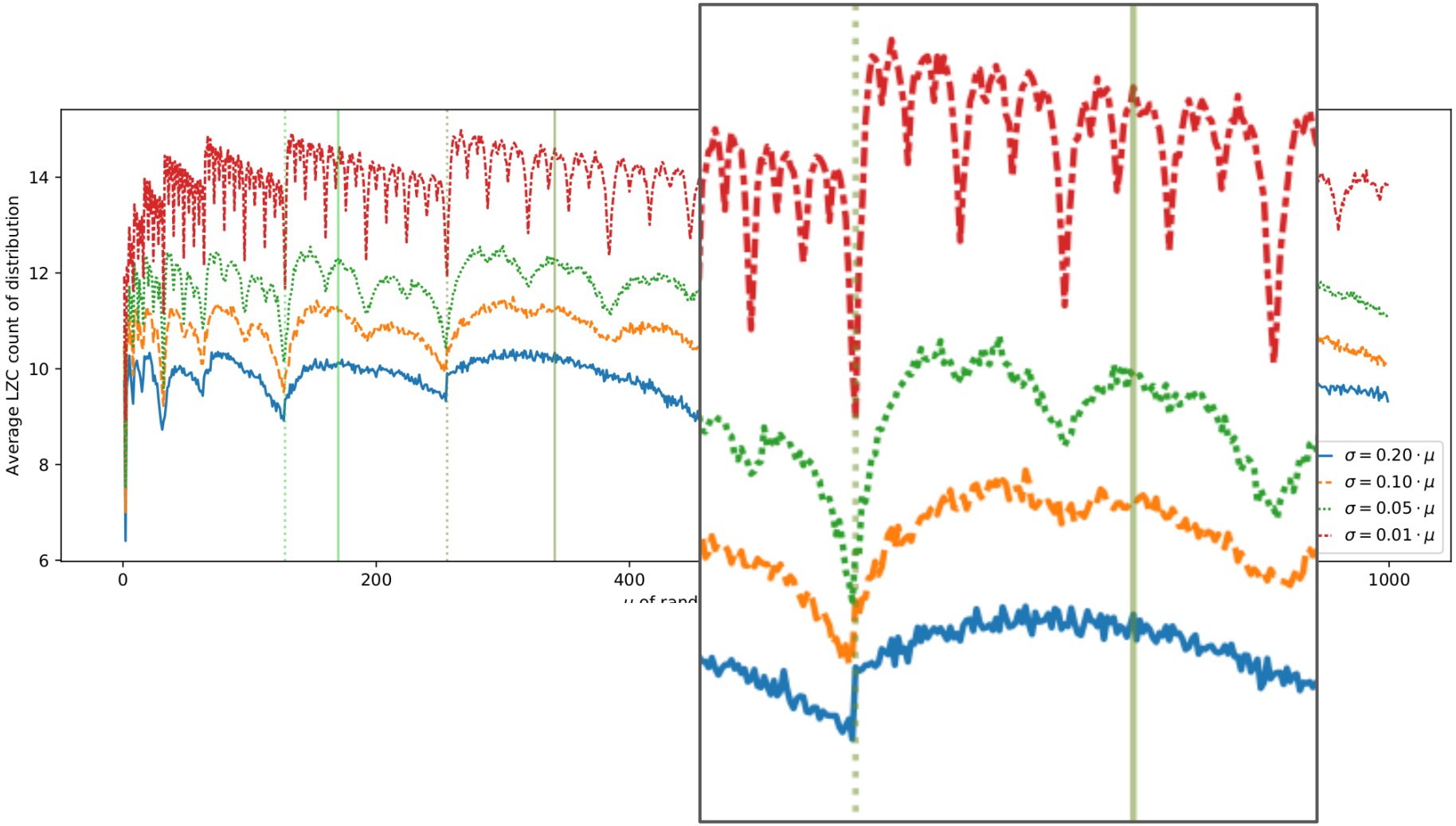



```
bin(p=256.321) = 01000011100000000010100100010111
bin(t=255.931) = 01000011011111111111011001010110
         p ⊕ t = 00000000111111111110001110100001
index 1  3  5  7  9  11 13 15 17 19 21 23 25 27 29 31
```

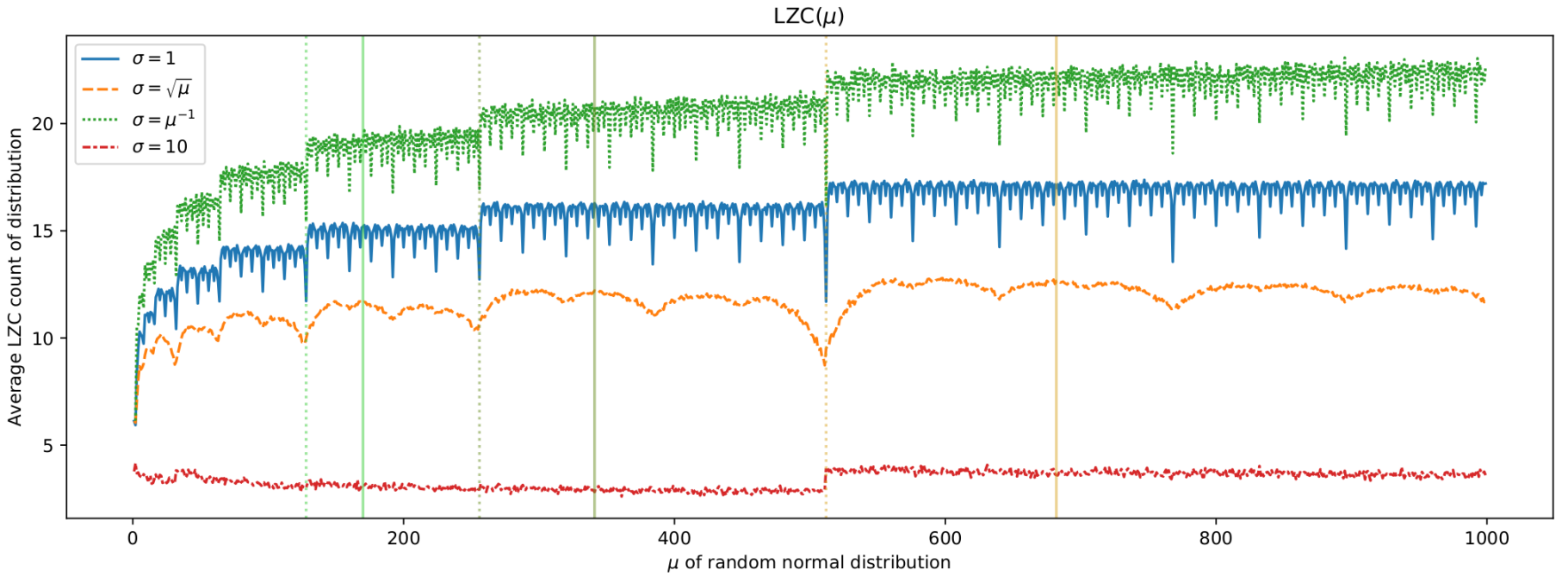
Average LZC for different normal distributed random datasets



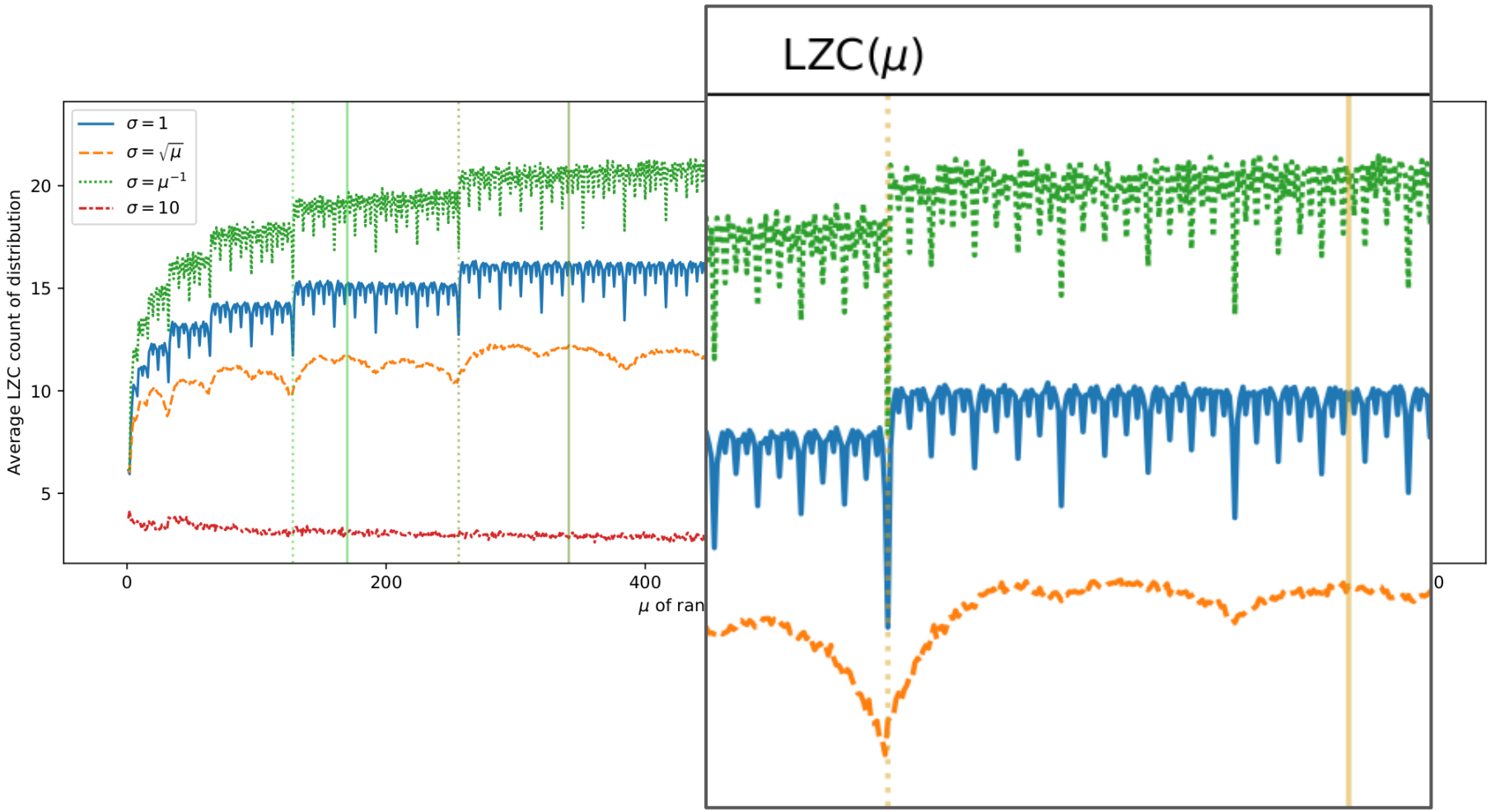
Average LZC for different normal distributed random datasets



Average LZC for different normal distributed random datasets (2/2)



Average LZC for different normal distributed random datasets (2/2)



Shifting the values and performing the residual calculation afterwards a results in better compression factor

Karlsruher Institut für Technologie

- If the values are fluctuating around powers of 2 it is advantageous **to shift** the values before calculating the residual

$$s = g(p) - p$$

$$g(p) = \begin{cases} \sum_{k=1}^{15} 2^{2k-1} & \text{if } p < 2^{30} \\ \sum_{k=0}^{14} 2^{2k} & \text{if } 2^{30} \leq p < 2^{31} \\ \sum_{k=1}^{16} 2^{2k-1} & \text{if } 2^{31} \leq p < 2^{32} \\ \sum_{k=0}^{15} 2^{2k} & \text{if } 2^{32} \leq p \end{cases}$$

$$g_1 = \sum_{k=1}^{15} 2^{2k-1} \text{ and } g_2 = \sum_{k=0}^{15} 2^{2k}$$

$$\text{bin}(g_1) = 00101010101010101010101010101010$$

$$\text{bin}(g_2) = 01010101010101010101010101010101$$

What effects has the application and chaining of encoding methods on the compression factor?

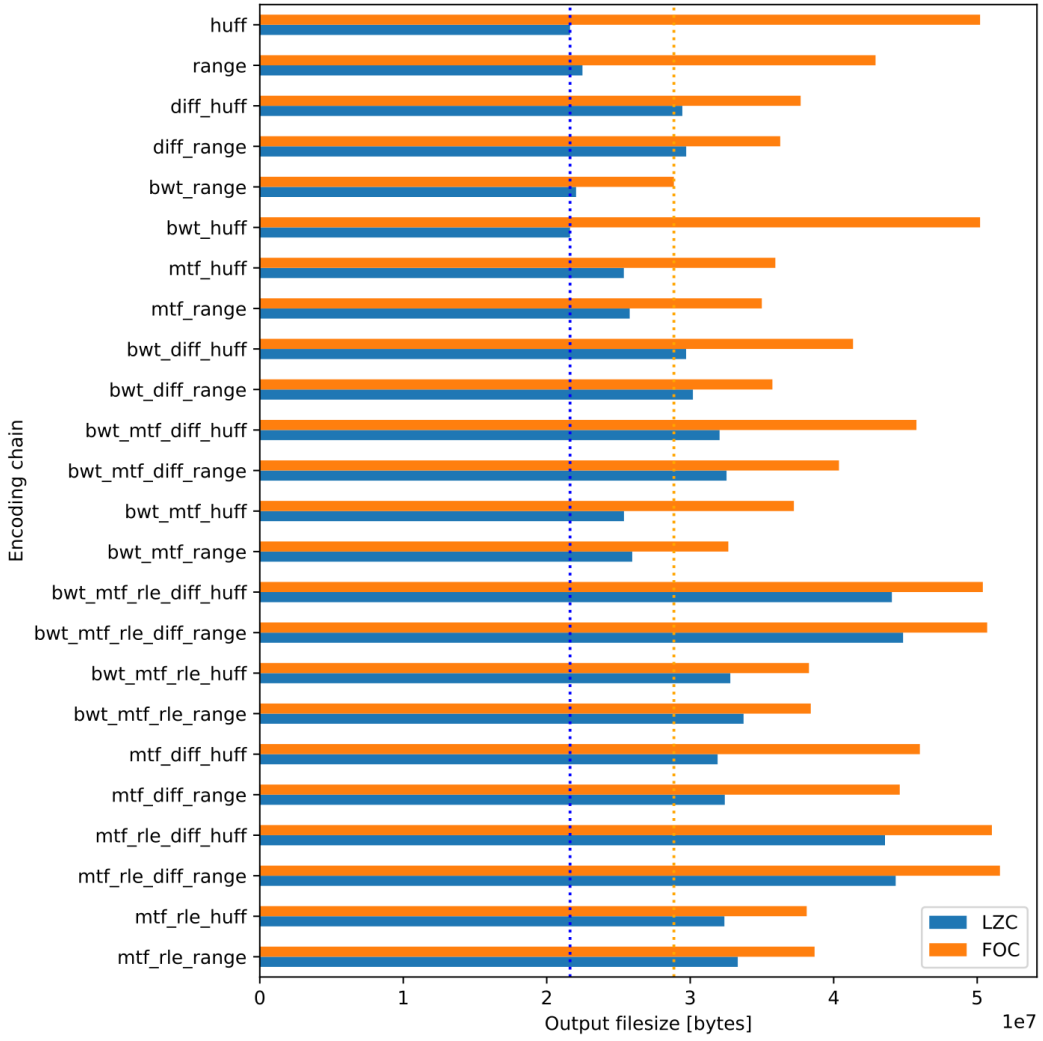
Karlsruher Institut für Technologie

Analysis 2

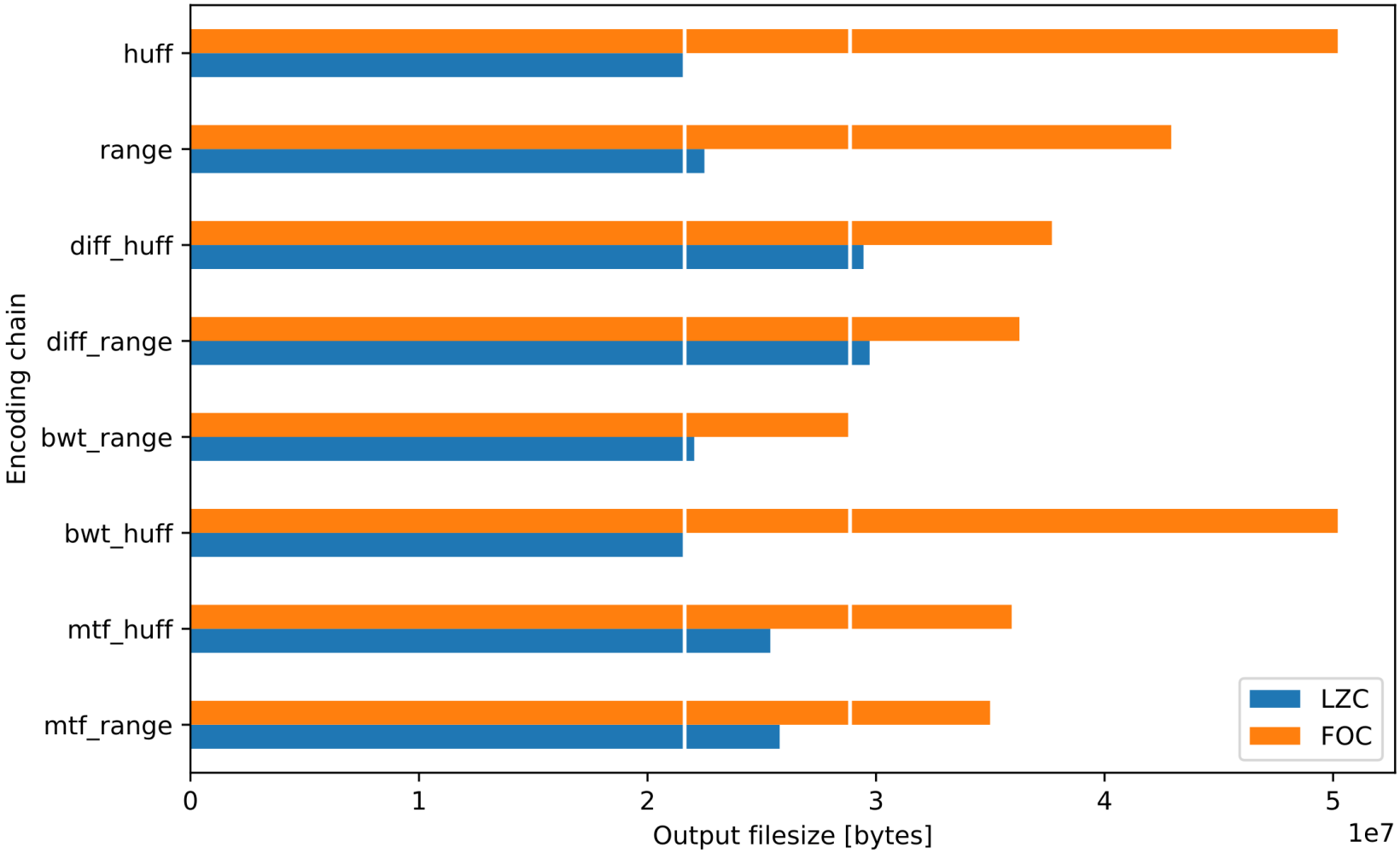
- Burrow-Wheeler-Transform (BWT)
- Delta Difference (Delta)
- Huffman Encoding (Huff)
- Range Encoding (Range)
- Move-to-Front (MTF)
- Run-length Encoding (RLE)

Transformation and Encoding Methods	FOC	LZC
bwt_diff_huff	29715223	41347759
bwt_diff_range	30187482	35727381
bwt_huff	21625804	50204194
bwt_mtf_diff_huff	32046132	45767747
bwt_mtf_diff_range	32531946	40367196
bwt_mtf_huff	25382523	37223861
bwt_mtf_range	25958407	32657611
bwt_mtf_rle_diff_huff	44054552	50393489
bwt_mtf_rle_diff_range	44836681	50698553
bwt_mtf_rle_huff	32794335	38277196
bwt_mtf_rle_range	33716265	38402906
bwt_range	22043891	**28862090**
diff_huff	29451797	37695831
diff_range	29720133	36274895
huff	**21625389**	50203689
mtf_diff_huff	31908557	46008528
mtf_diff_range	32408277	44602439
mtf_huff	25374374	35935034
mtf_range	25782694	34990704
mtf_rle_diff_huff	43577596	51026248
mtf_rle_diff_range	44320799	51588757
mtf_rle_huff	32380683	38119328
mtf_rle_range	33313109	38667389
range	22489931	42918321

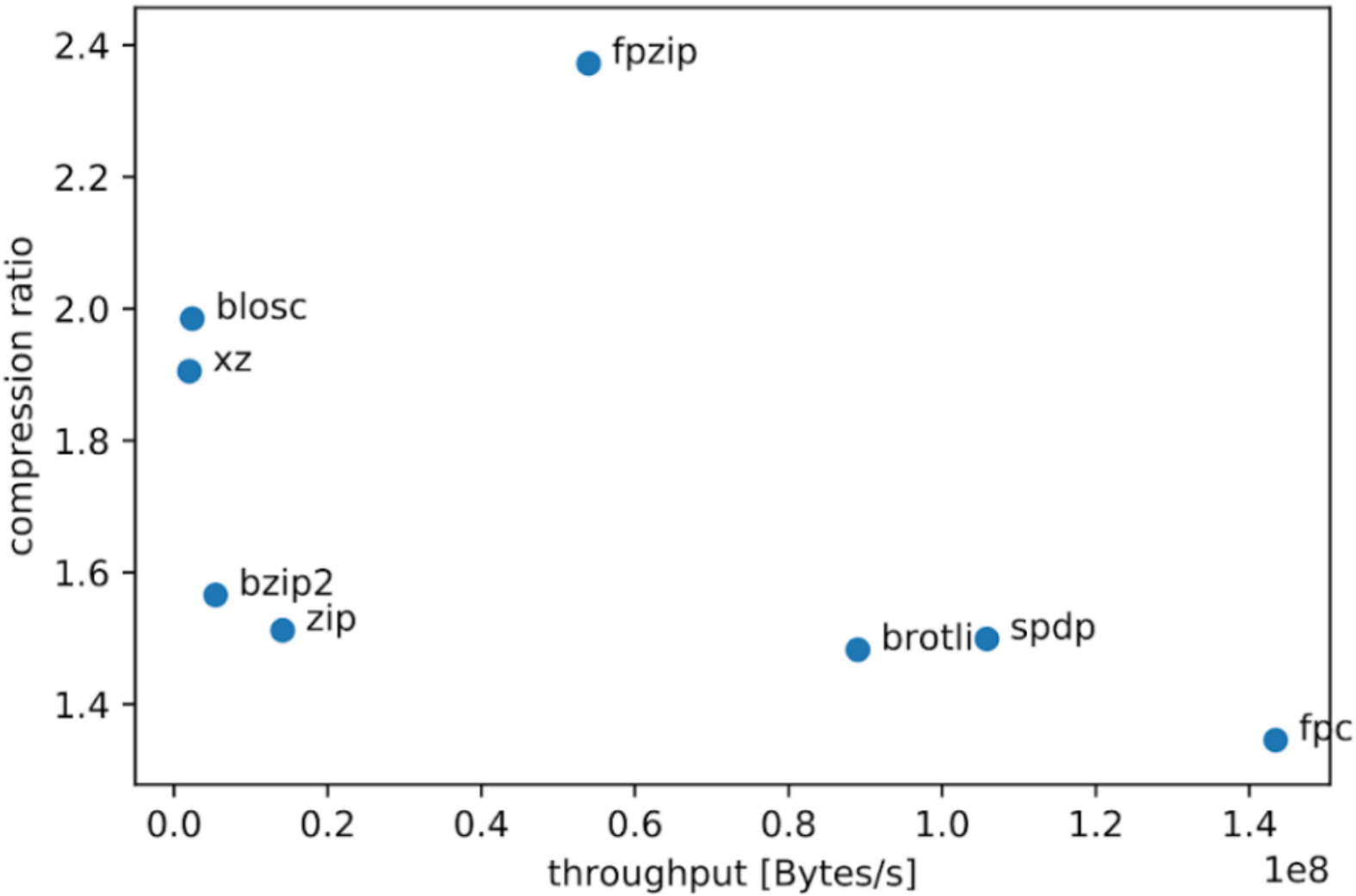
BWT + Range Encoder perform (close to) best for (LZC) FOC



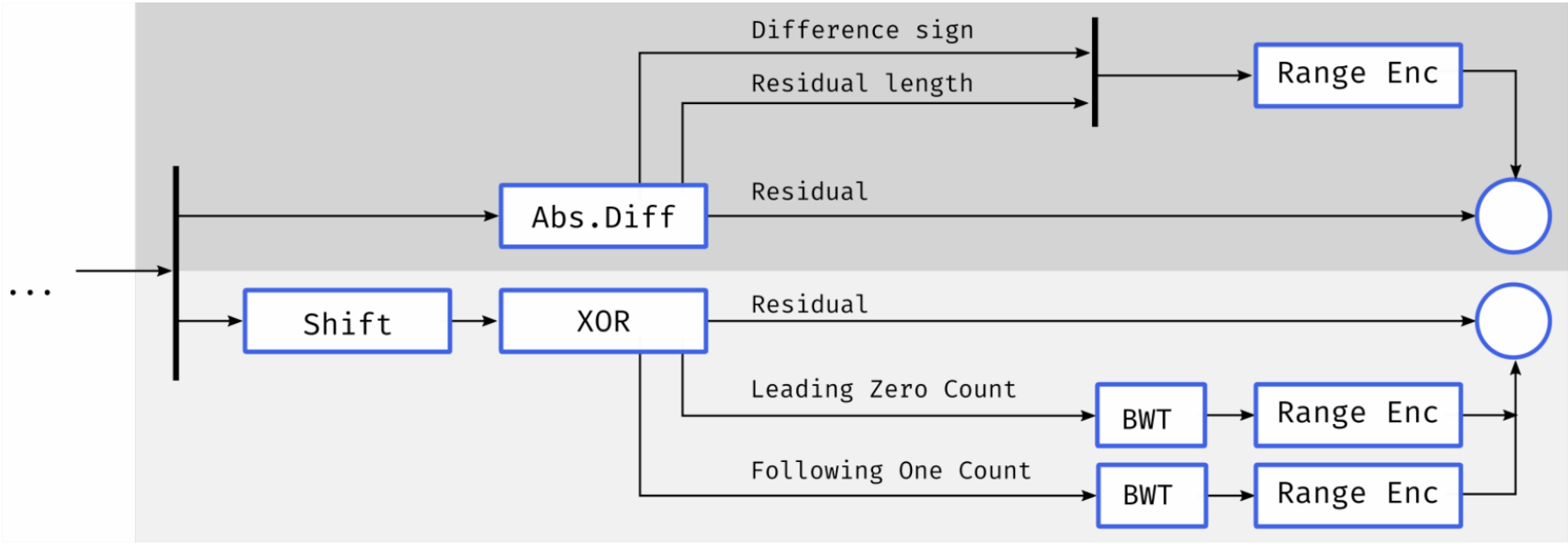
BWT + Range Encoder perform (close to) best for (LZC) FOC



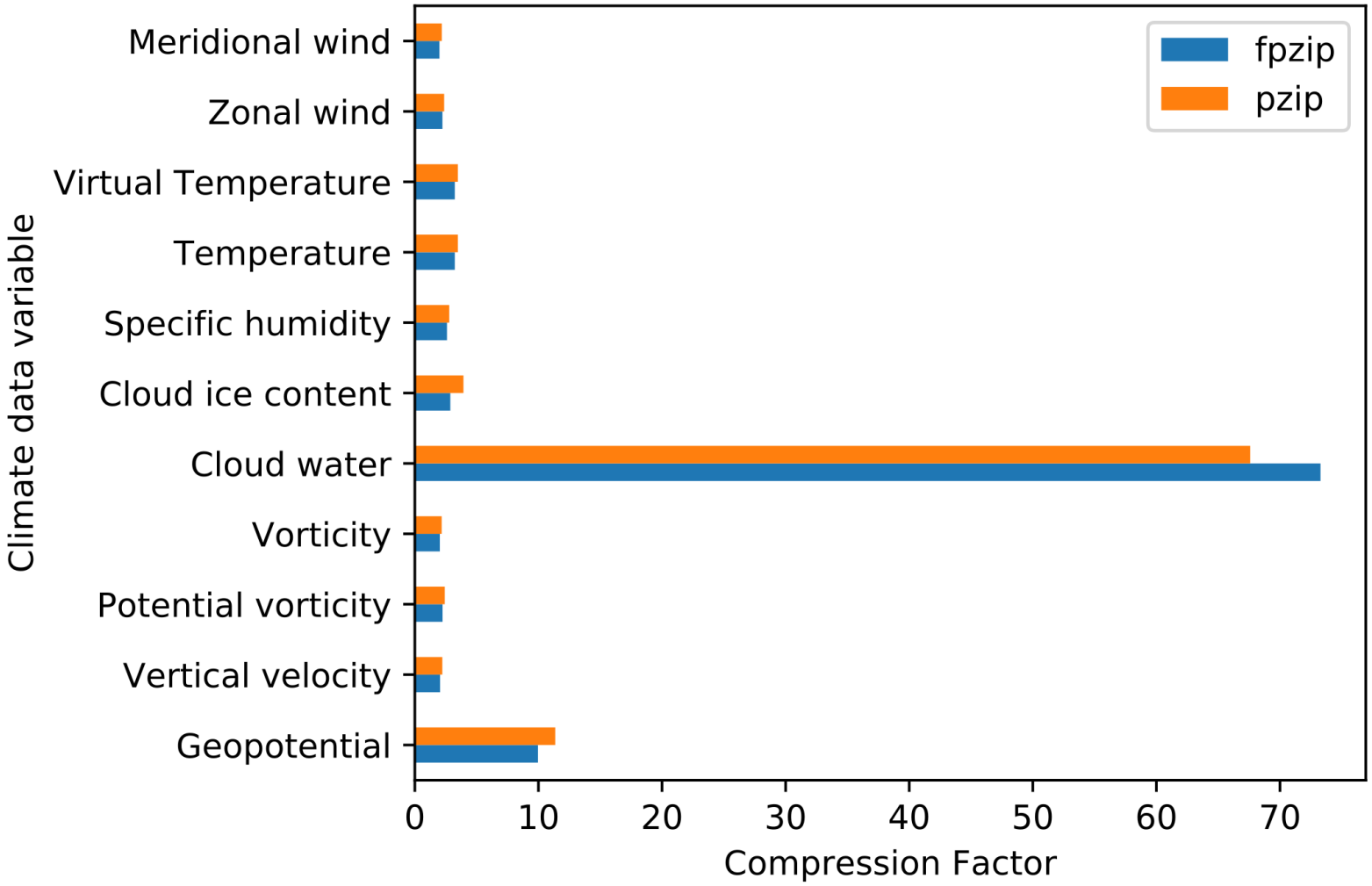
Performance of current state-of-the-art compression algorithms.



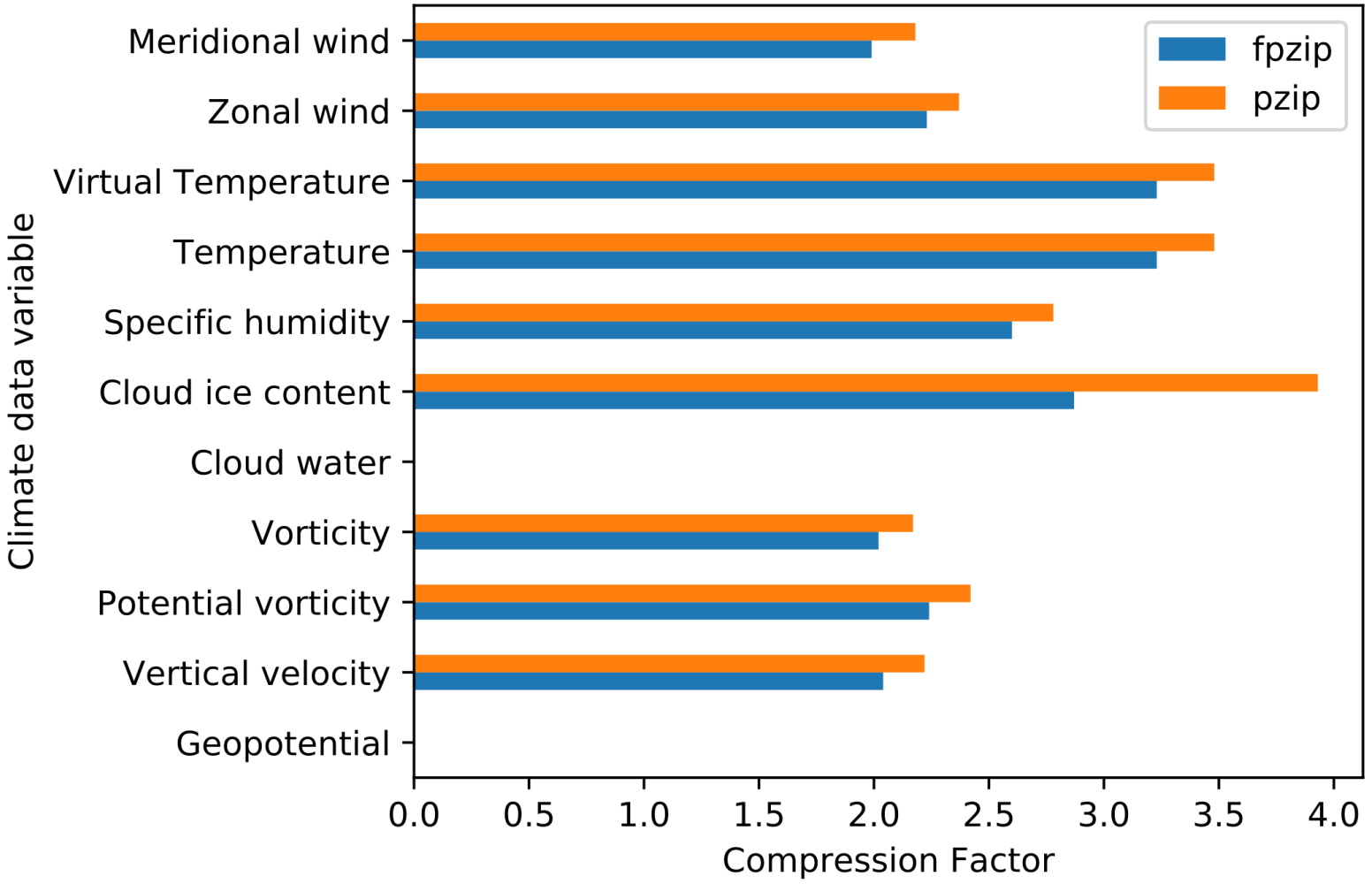
Final scheme of compression method



Final results of compression factor using our proposed approach (1)



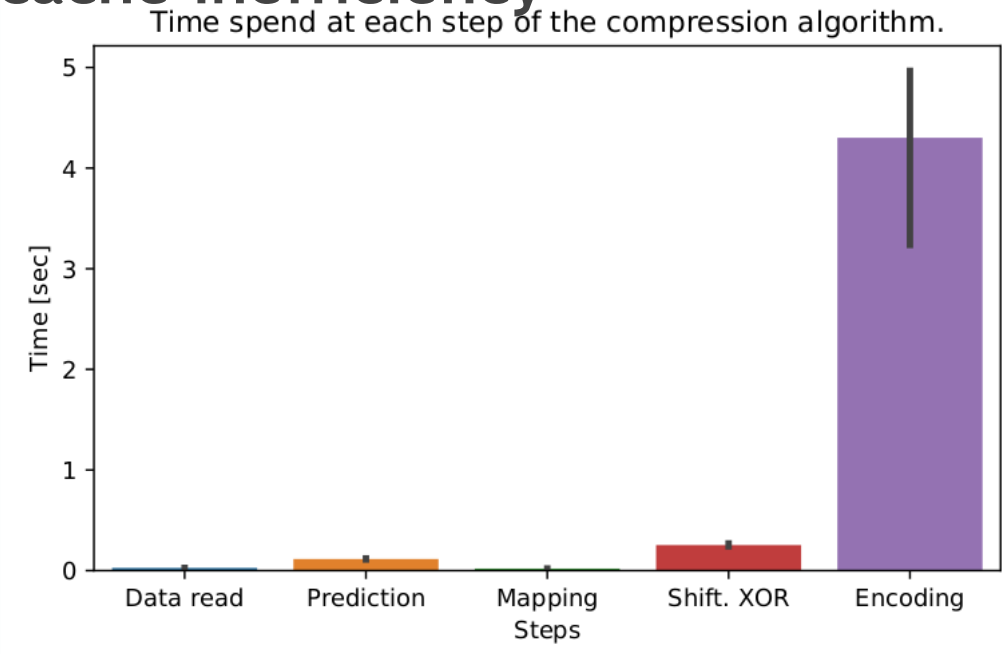
Final results of compression factor using our proposed approach (2)



Pitfalls and future work regarding our proposed approach

Karlsruher Institut für Technologie

- 5x slower than fpzip
- Memory space needed by BWT
- Big-O complexity of fpzip and pzip is same > O(n)
- Might be due to L1-L3 cache inefficiency



Karlsruher Institut für Technologie

6. Building a framework to perform rapid testing of new compression algorithms

Step 6

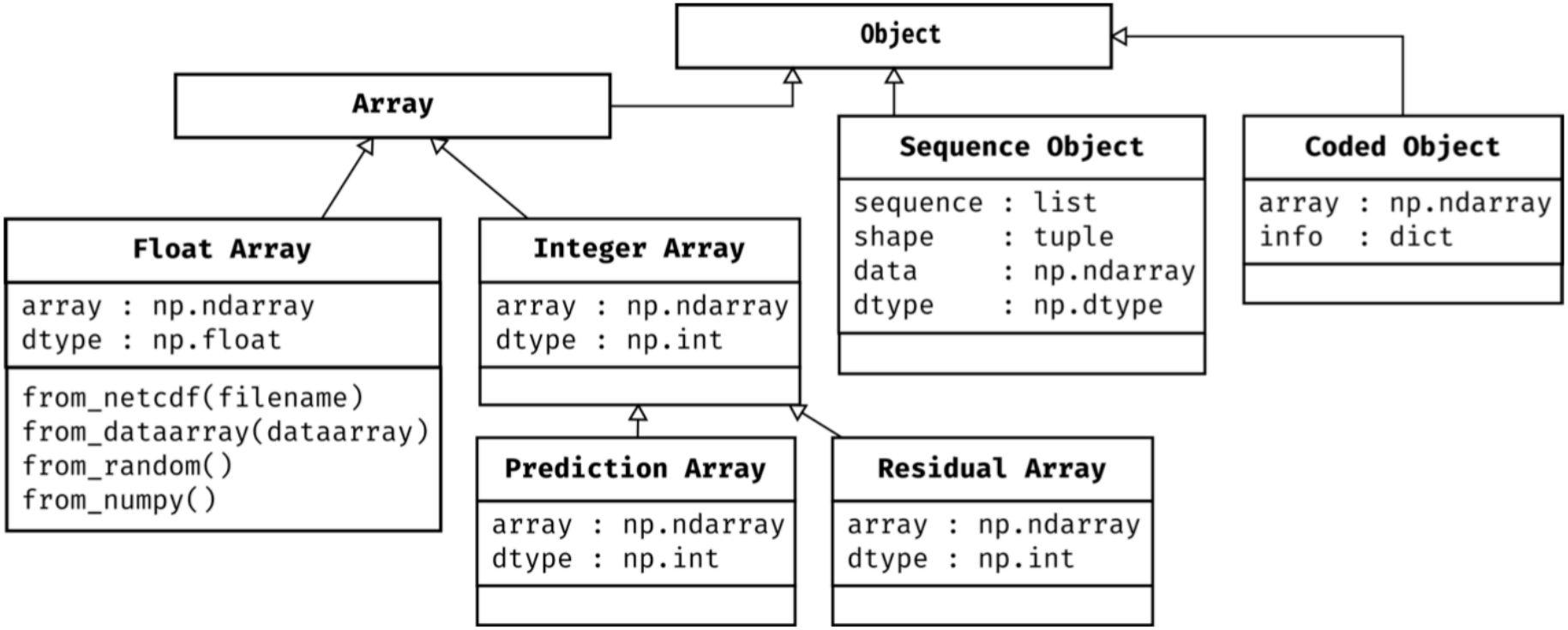
Building a framework to perform rapid testing of new compression algorithms

Karlsruher Institut für Technologie

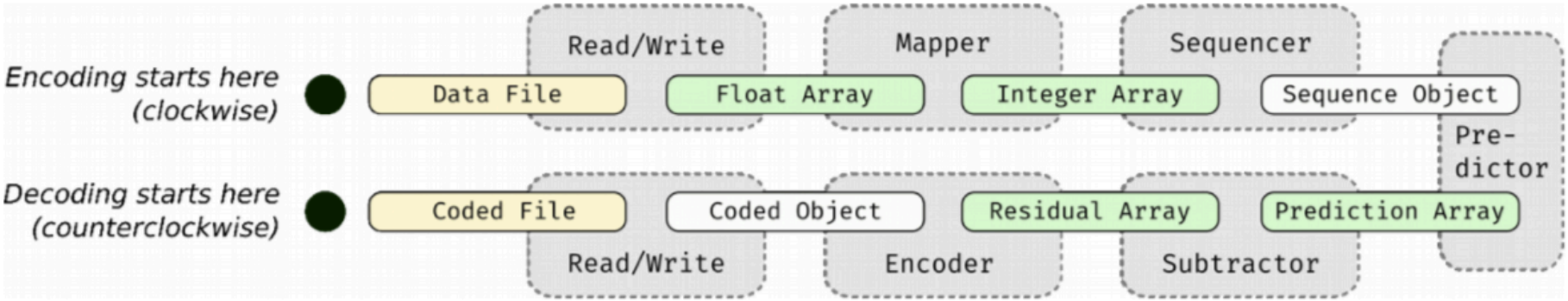
Development of a custom compression algorithm is an iterative process with a lot of parameters to think about. Rapid testing of new settings and environments is crucial for the development.
Requirements:

- Extensible architecture using strict interfaces
- Testing of different predictors
- Support for quality assessment

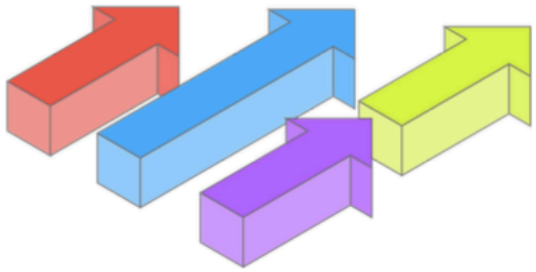
Framework


Object
Array
Sequence Object
sequence : list
shape : tuple
data : np.ndarray
dtype : np.dtype
Coded Object
array : np.ndarray
info : dict
Float Array
array : np.ndarray
dtype : np.float
from_netcdf(filename)
from_dataarray(dataarray)
from_random()
from_numpy()
Integer Array
array : np.ndarray
dtype : np.int
Prediction Array
array : np.ndarray
dtype : np.int
Residual Array
array : np.ndarray
dtype : np.int

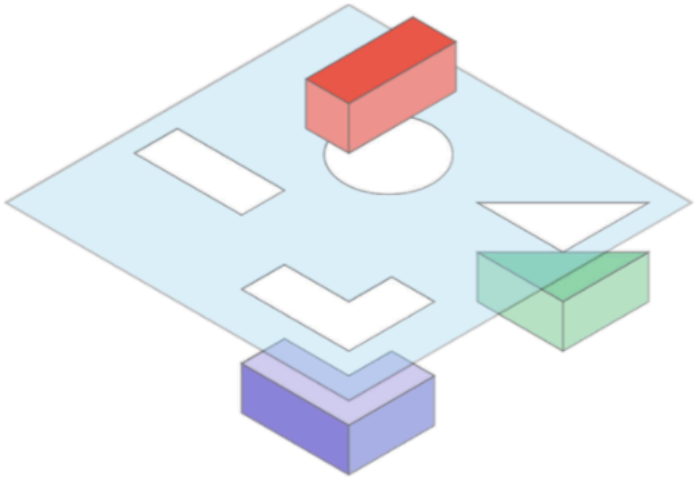
Framework


Read/Write
Mapper
Sequencer
Encoding starts here
(clockwise)
Data File
Float Array
Integer Array
Sequence Object
Pre-
dictor
Decoding starts here
(counterclockwise)
Coded File
Coded Object
Residual Array
Prediction Array
Read/Write
Encoder
Subtractor

Framework

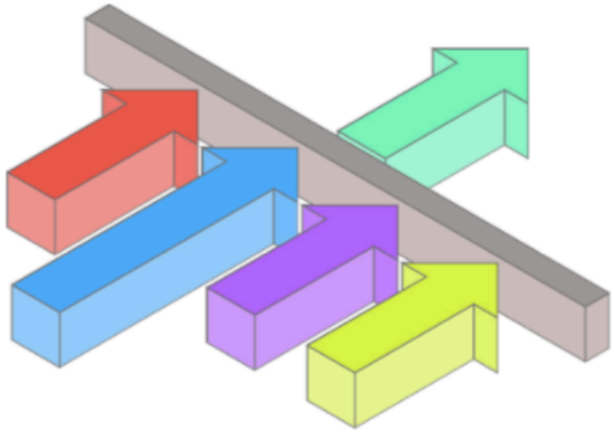


Concurrent compression of a single dataset using several predictors or using a single predictor with several datasets

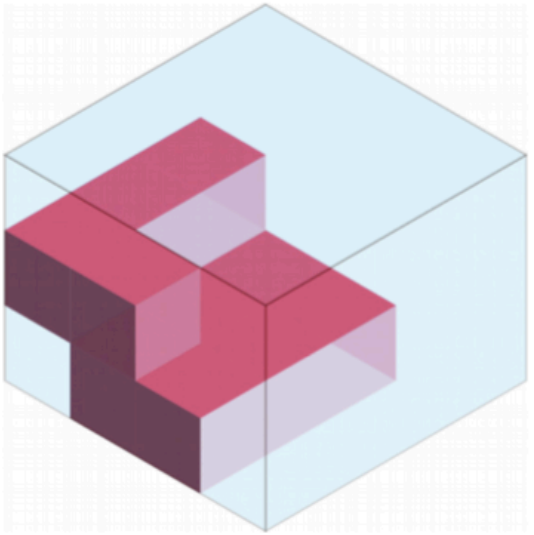


Interaction of modifiers and objects follow strictly defined interfaces to guarantee interoperability

Framework



Several predictors can be merged to a single entity to improve prediction accuracy



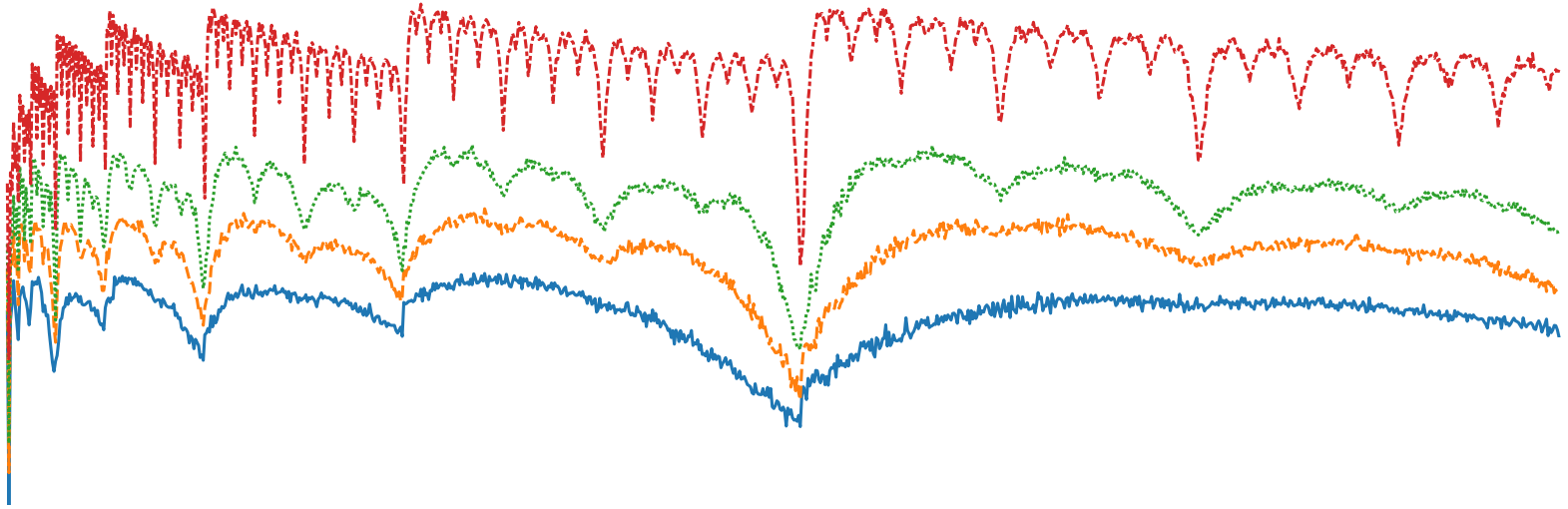
Support for random subsetting of multi-dim. data for fast iteration

Karlsruher Institut für Technologie

Kompressionsmethoden für strukturierte Gleitkommazahlen und ihre Anwendung in den Klimawissenschaften

von Uğur Çayoğlu

STEINBUCH CENTRE FOR COMPUTING (SCC) und
INSTITUT FÜR METEOROLOGIE UND KLIMAFORSCHUNG (IMK-ASF)



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem

Hohes Datenaufkommen durch Klimasimulationen

ERA5

Datensatz für die Initialisierung und Validierung von Simulationsläufen umfasst 10.89 PiB

IMK-ASF

Einer der größten Speicherplatzbenutzer am SCC mit >770 TiB (steigend)

Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem

Hohes Datenaufkommen durch Klimasimulationen (ERA5, 10.89 PiB)

Aktuelle Lösung

Reduzierung der zeitlichen Auflösung und gespeicherten Variablen

Folgen

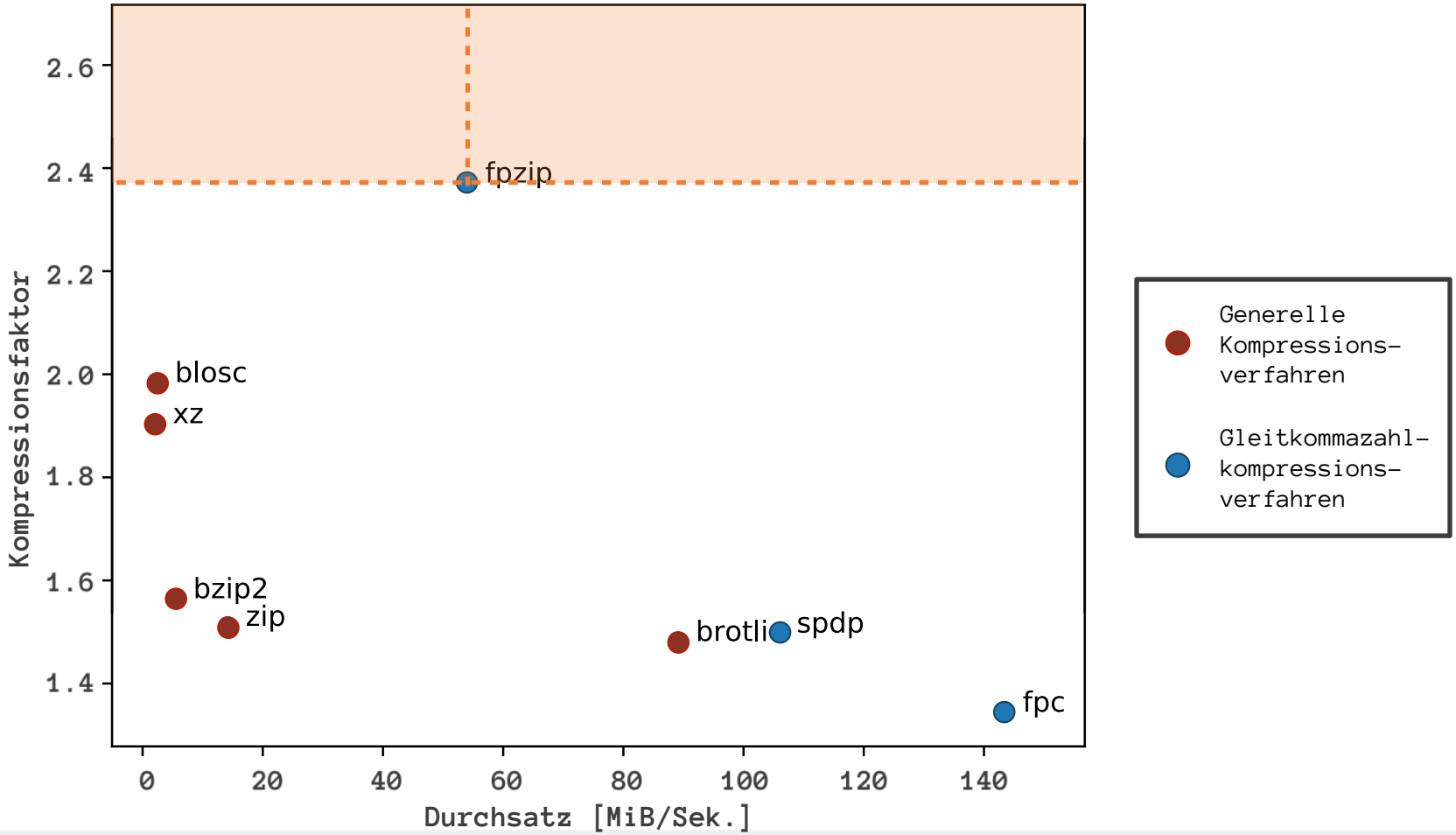
- Benutzung von Interpolationen
- Klimaereignisse (z.B. Entstehung von Stürmen) möglicherweise nicht abgebildet
- Neuberechnung von Simulationen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor

Kompressionsfaktor und Durchsatz

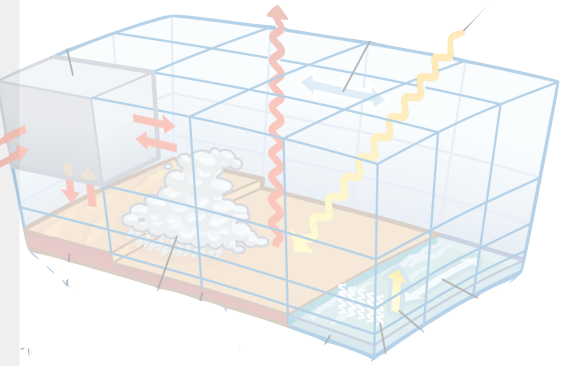
$$\text{Kompressionsfaktor} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Komprimierte Dateigröße [Bytes]}} \qquad \text{Durchsatz} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Kompressionszeit [Sek.]}}$$

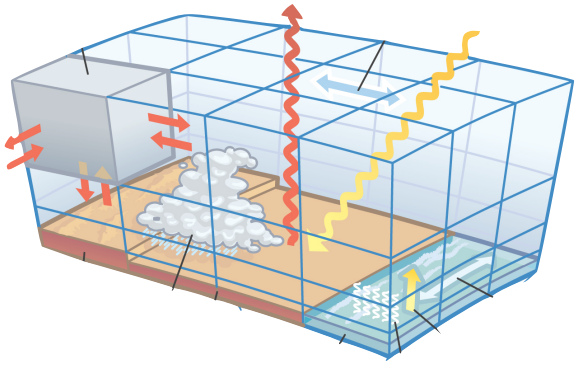


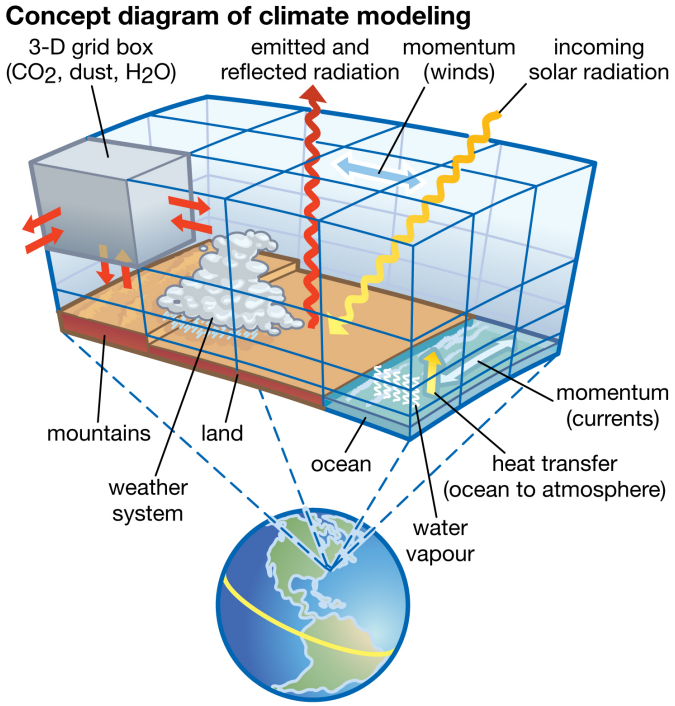
Verlustfreie Kompression von strukturierten Gleitkommazahlen

Karlsruher Institut für Technologie

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

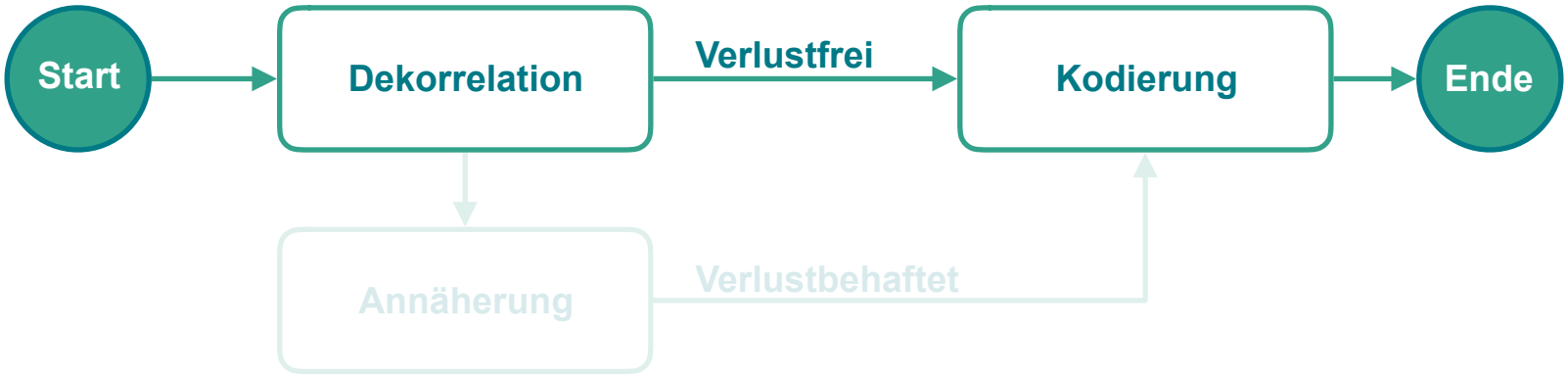






Quelle[1]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



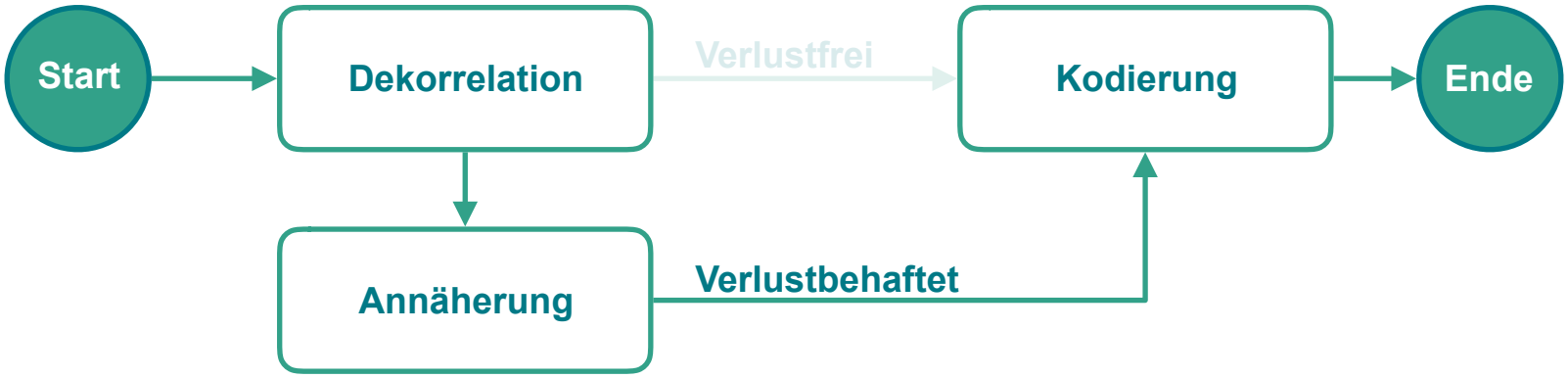
Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

JPEG

Kompressionsarten

- **Verlustbehaftete**
- Verlustfreie

Datentypen

- **Integer**
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- **Bild**
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- **Lauflängenkodierung**
- Wörterbuch-Verfahren
- **Transformation**
- Vorhersagebasierte Verfahren
- **Entropie-basierte Verfahren**

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Beiträge zur Informatik

Karlsruher Institut für Technologie

pzip

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

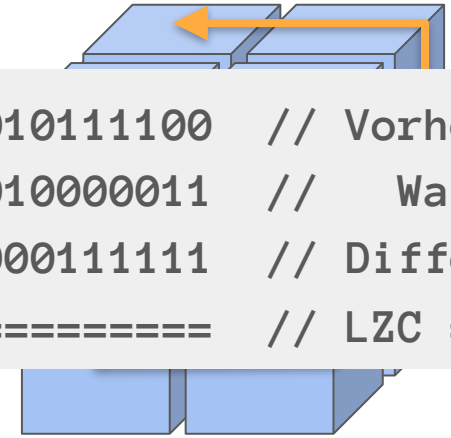
Direkter Beitrag | Einflussbereich

Pascal Zip (pzip)

Vorhersagebasiertes Kompressionsverfahren

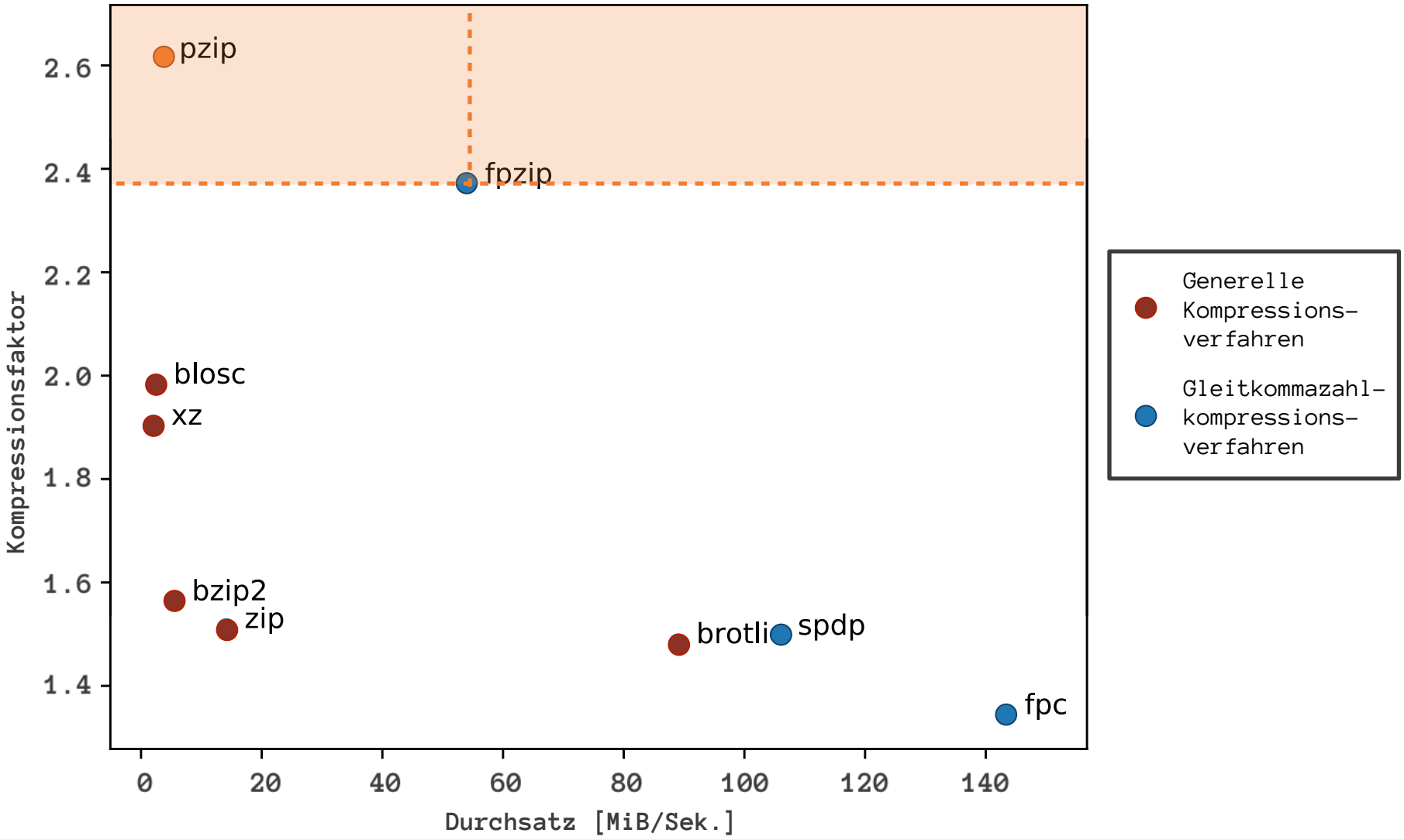
Methode

Für jeden einzelnen Datenpunkt wird (basierend auf vorhergehenden Werten) eine Vorhersage gegeben und die Differenz zum wahren Wert (Residuum) gespeichert

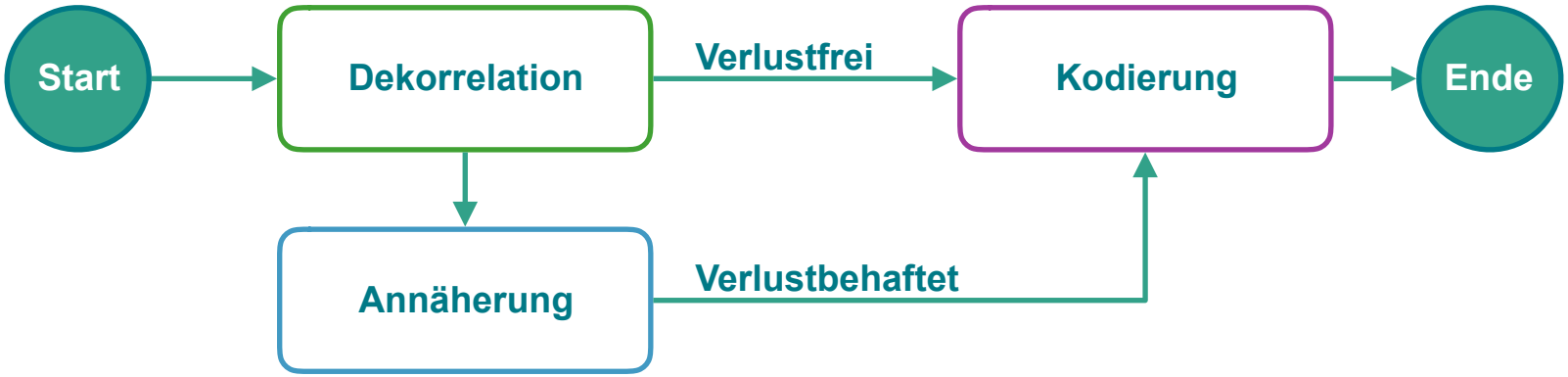


```
01000010100001111011010010111100  // Vorhersage = 67.853
01000010100100101000000010000011  //   Wahrheit = 73.251
00000000000101010011010000111111  // Differenz
          *====================  // LZC = 11 -> 32 - 11 - 1 = 20
```

Kompressionsverfahren im Vergleich



Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

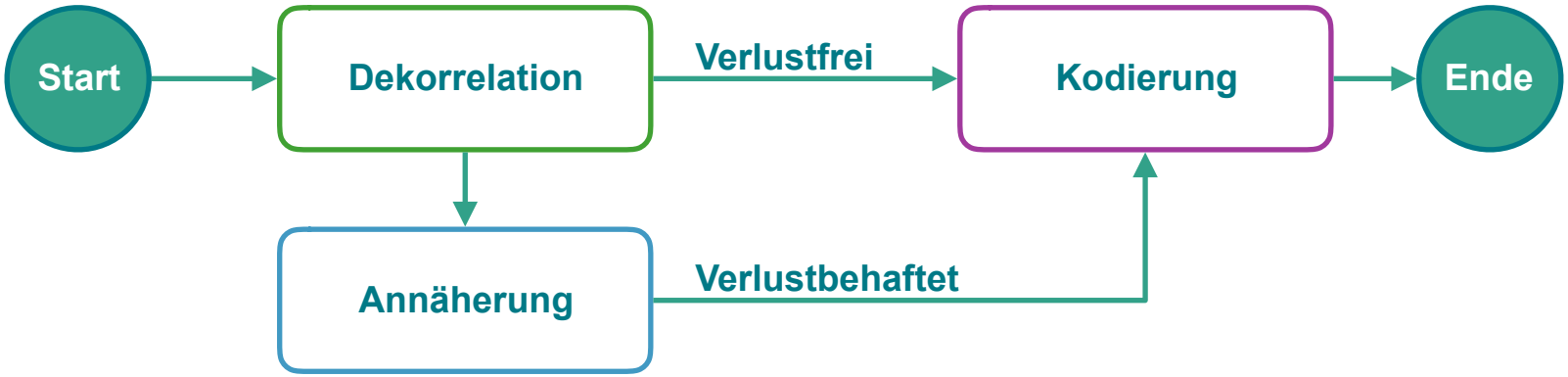
Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

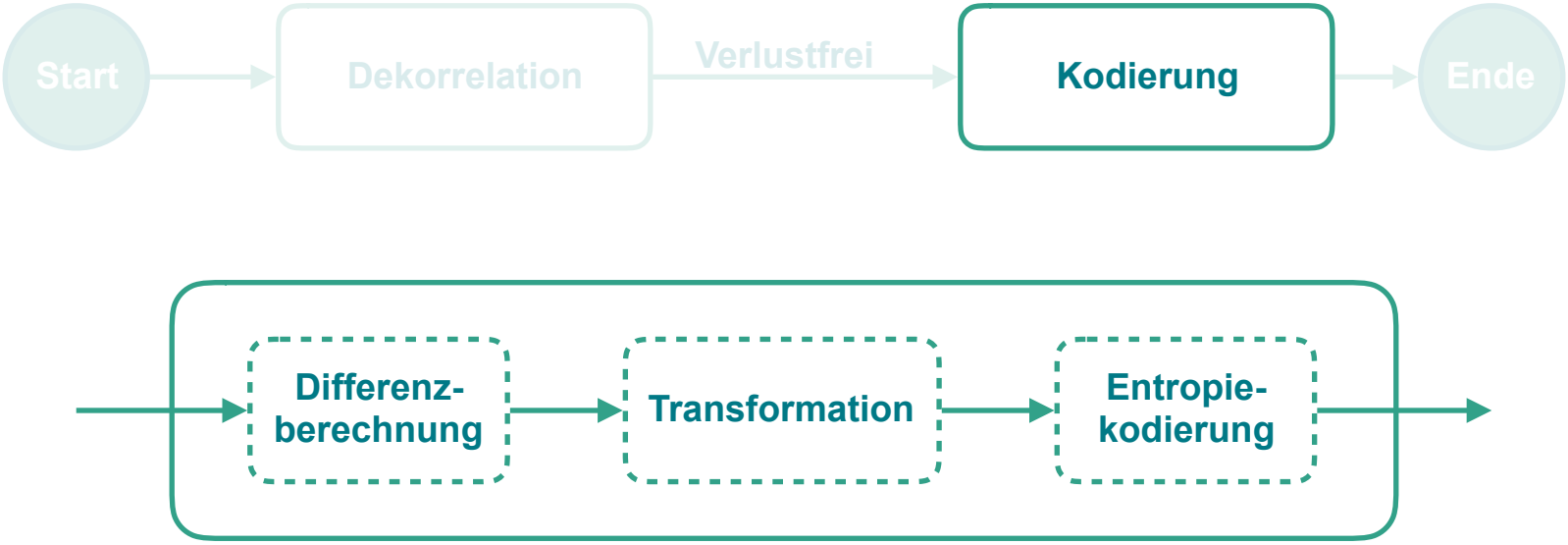
Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

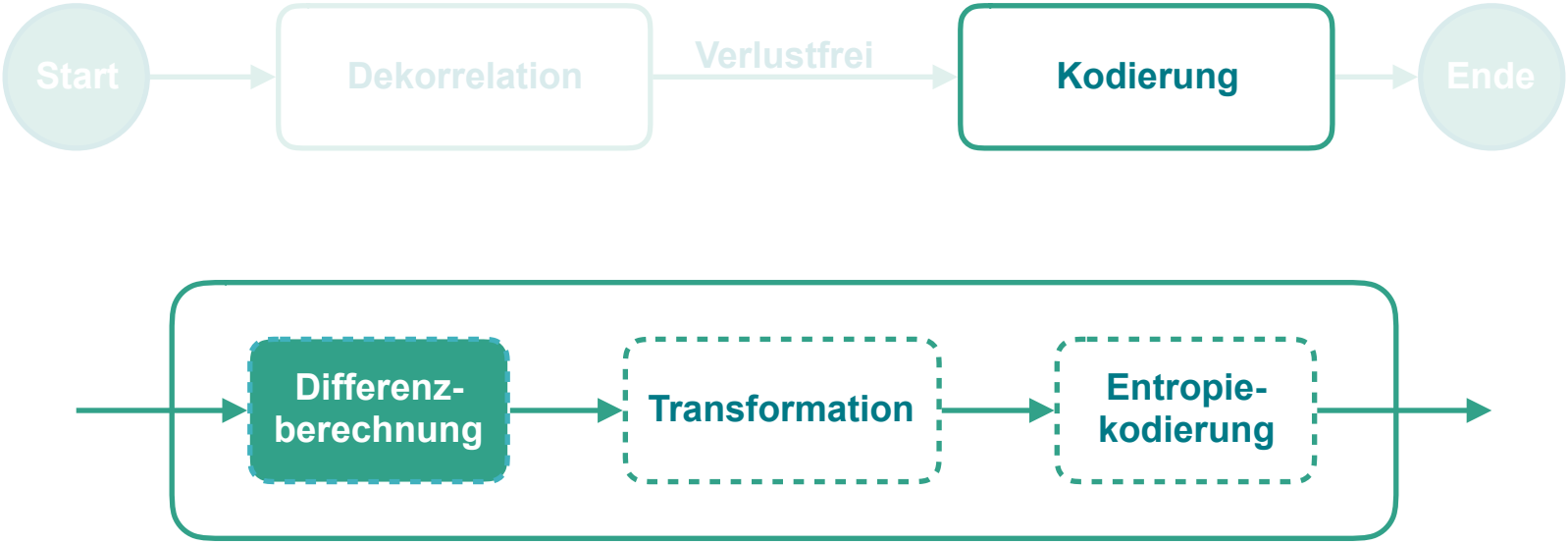
Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Datenkodierung bei der verlustfreien vorhersagebasierten Kompression

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Zwei Arten der Berechnung von Residuen

Karlsruher Institut für Technologie

Abs. Differenz

$$d = |v - w|$$

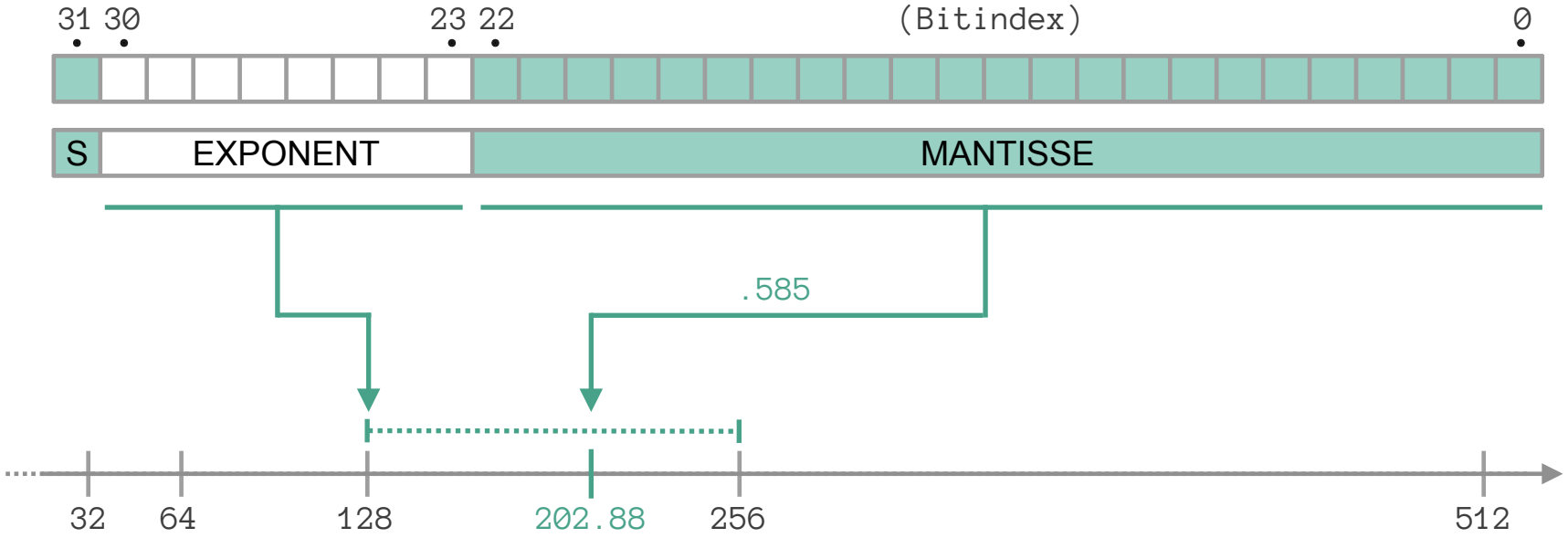
- ⊕ Kleine Residuen
- ⊖ Underflow
- ⊖ Zwei Operationen
- ⊖ Bit für Vorzeichen

XOR

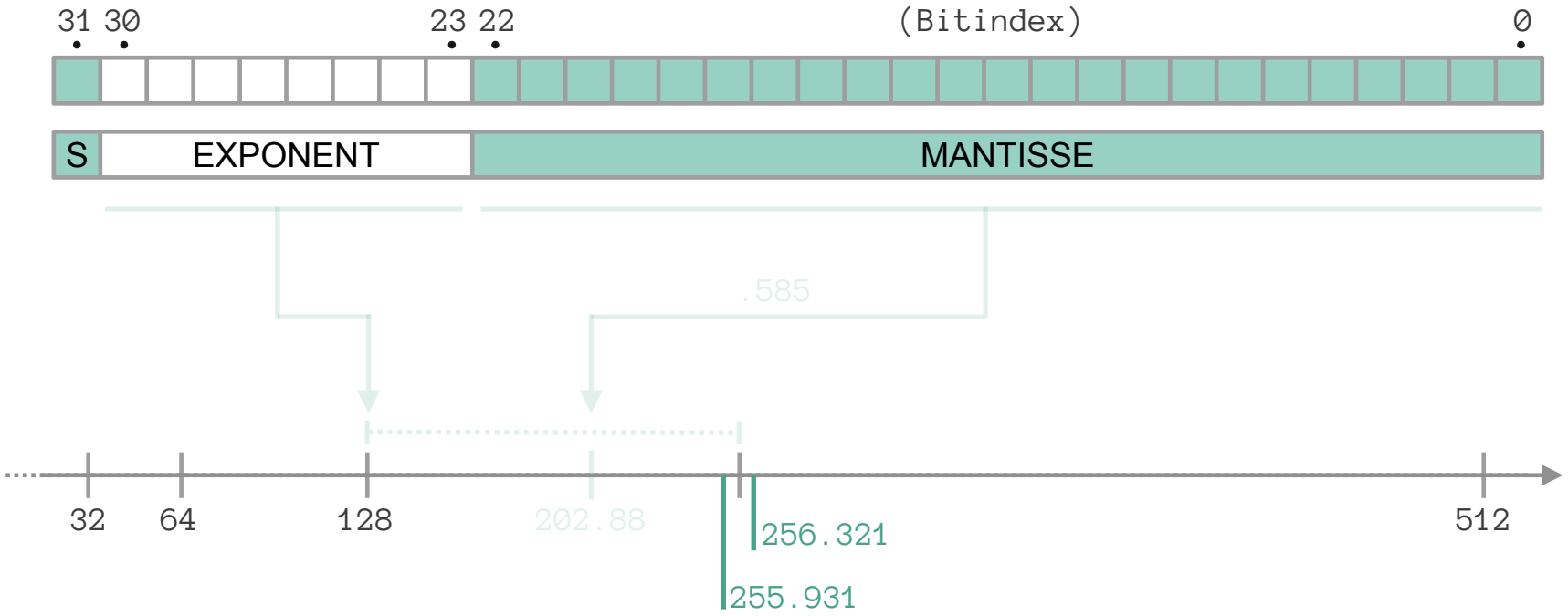
$$d = v \oplus w$$

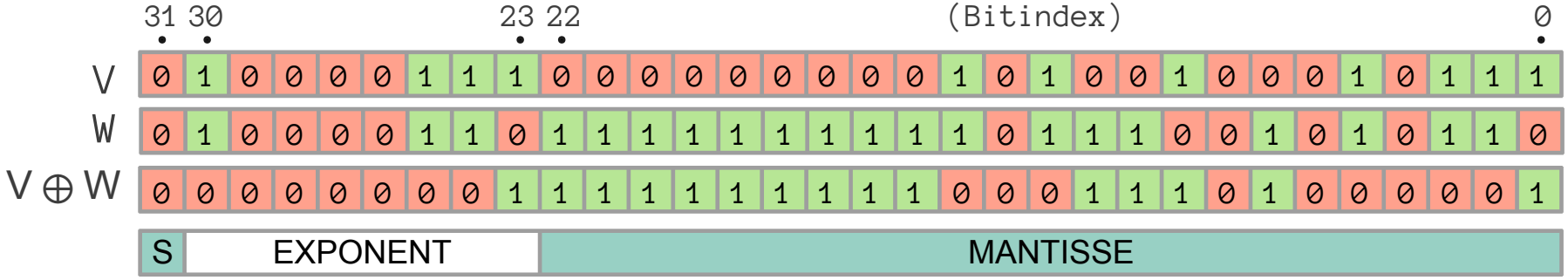
- ⊕ Eine Operation
- ⊕ Kein Underflow
- ⊖ Bitflip-Problem (große Residuen)

Gleitkommazahlen und der Bitflip

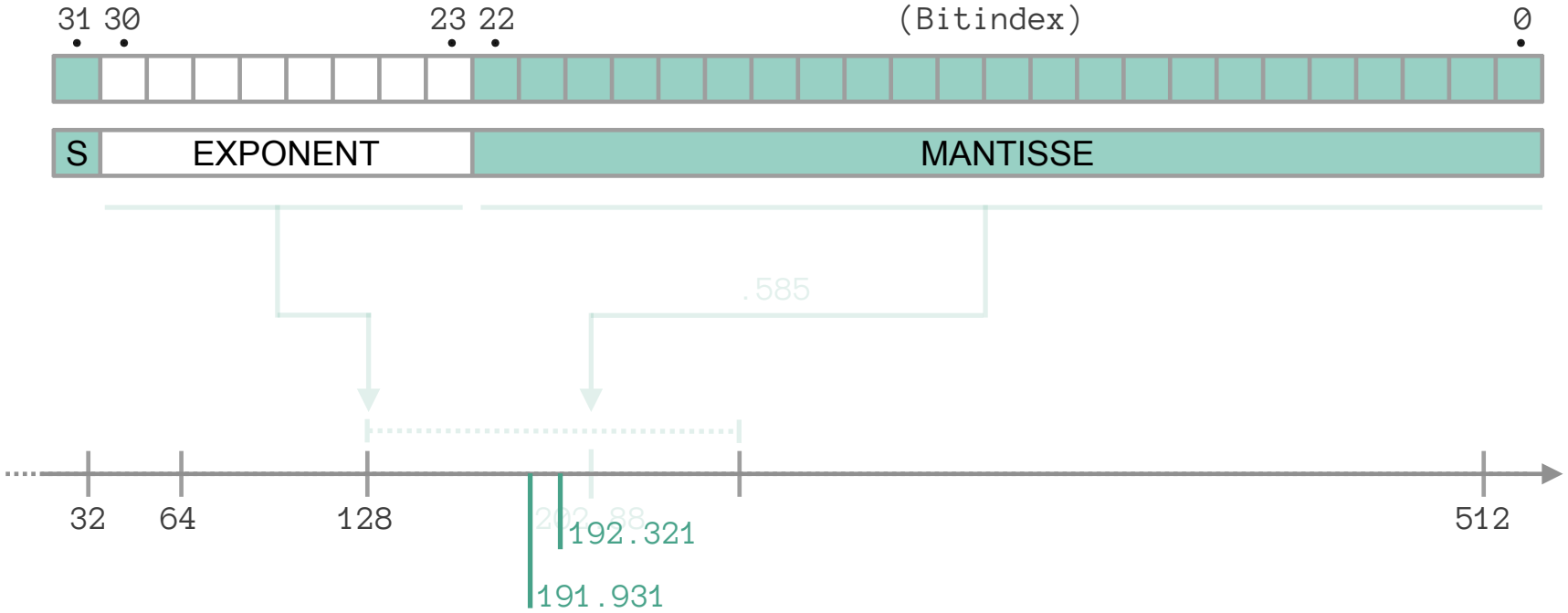


Gleitkommazahlen und der Bitflip

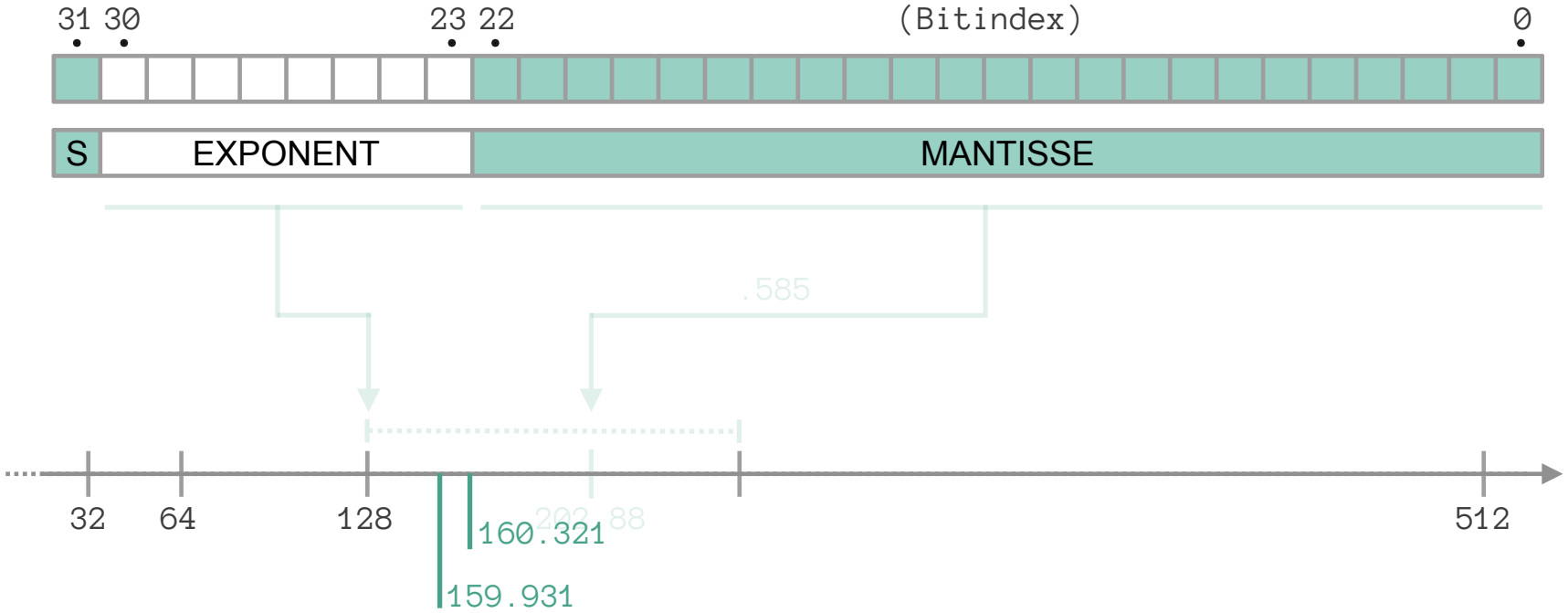




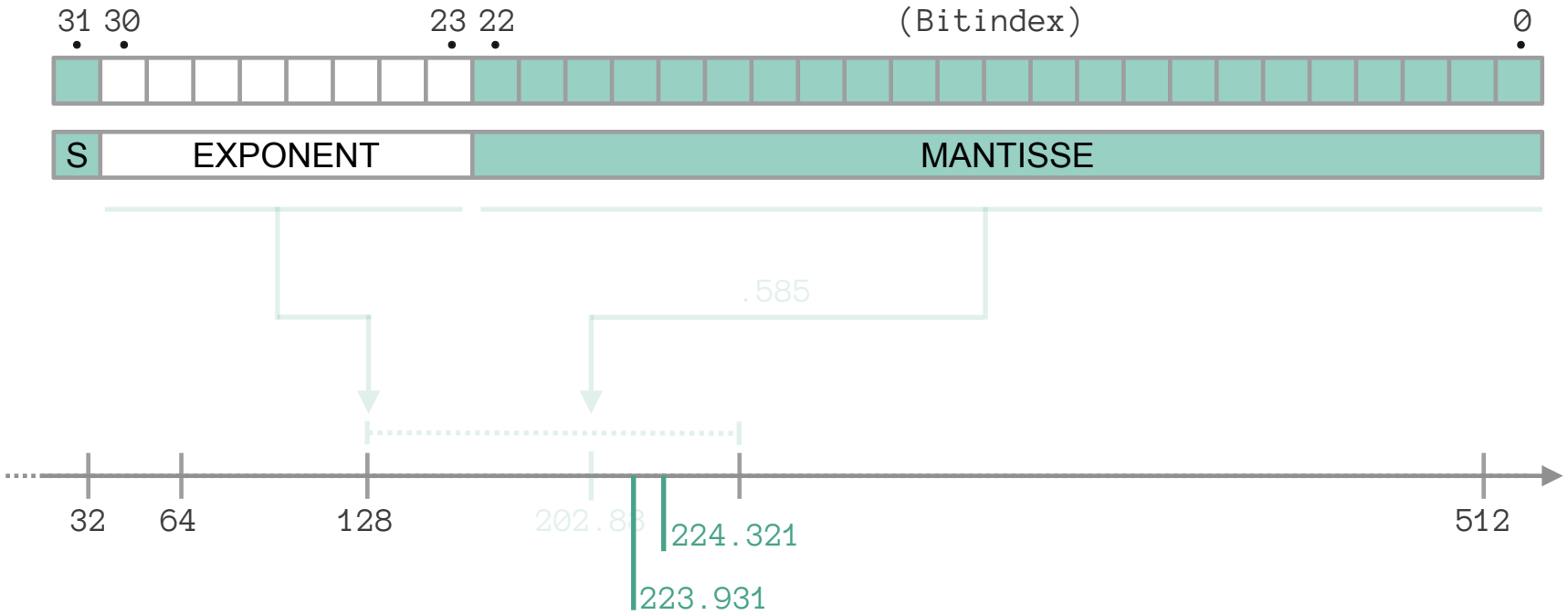
Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

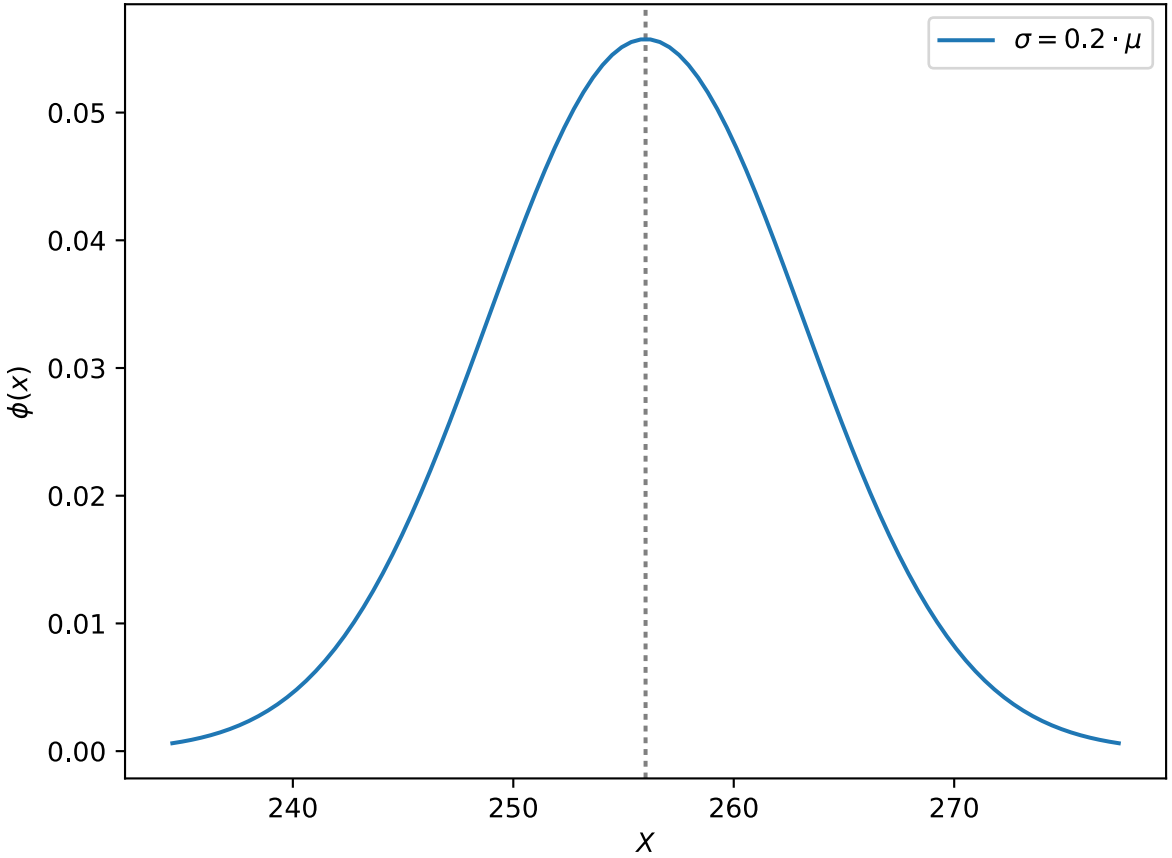
Frage

Kann ich im Vorfeld bestimmen wie stark die Kompression vom Bitflip betroffen sein wird?

Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

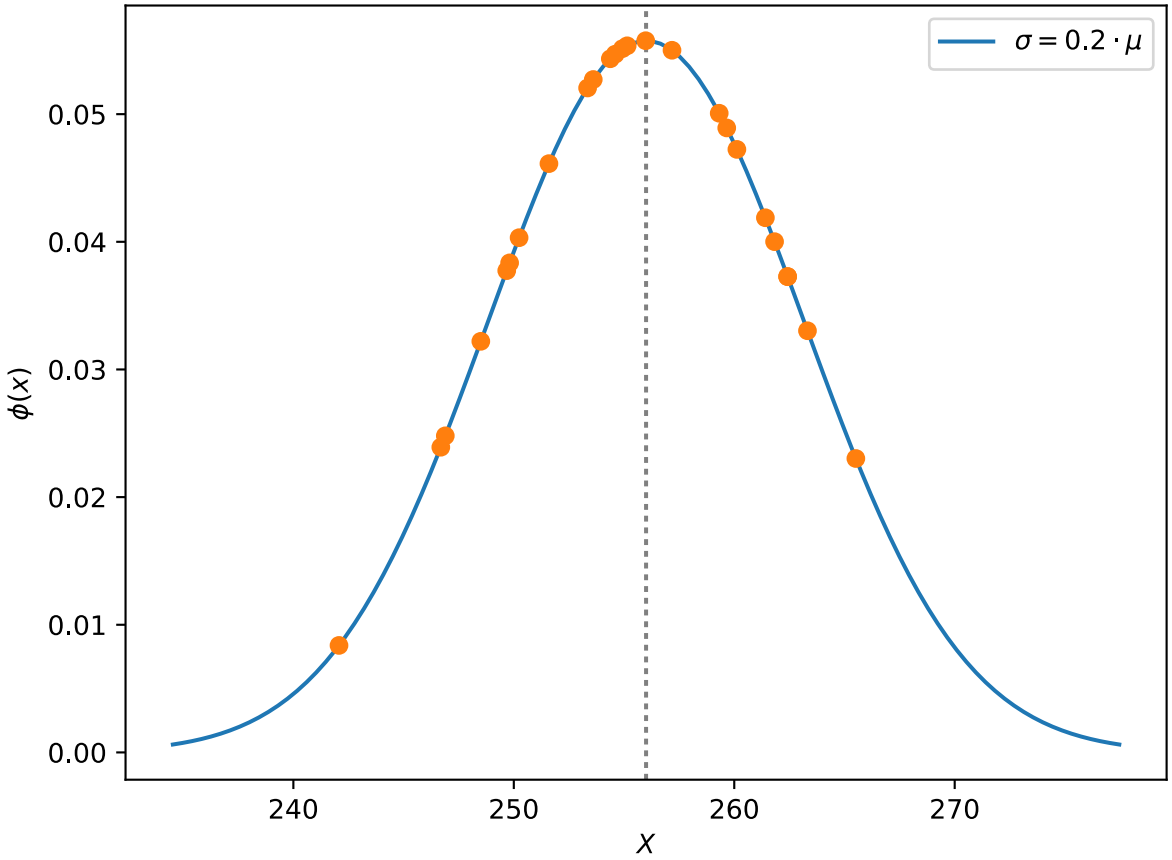
Prämisse: Normalverteilung der Vorhersagen



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

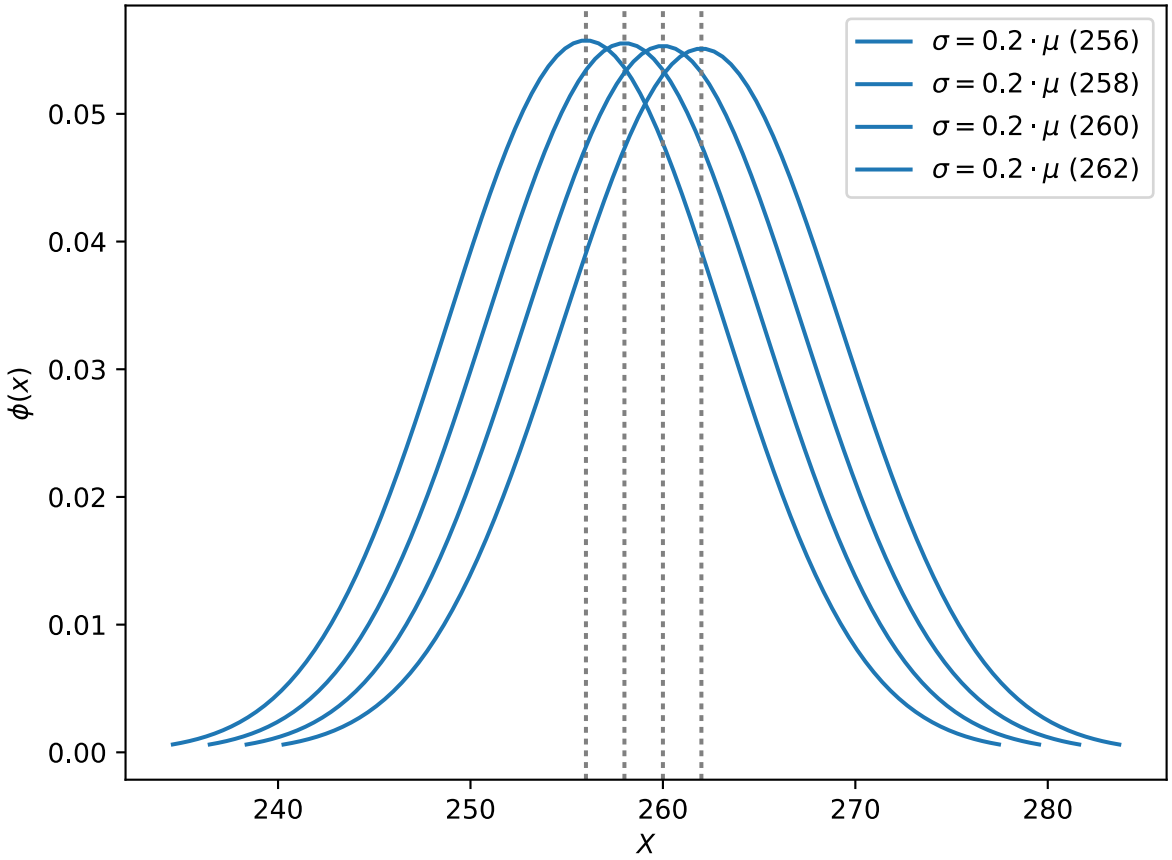
Prämisse: Normalverteilung der Vorhersagen



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

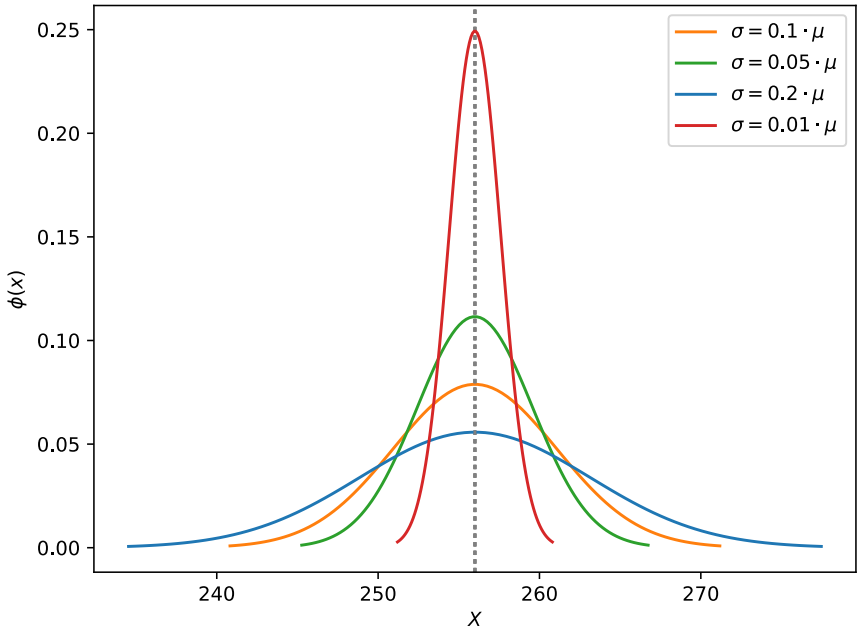
Prämisse: Normalverteilung der Vorhersagen



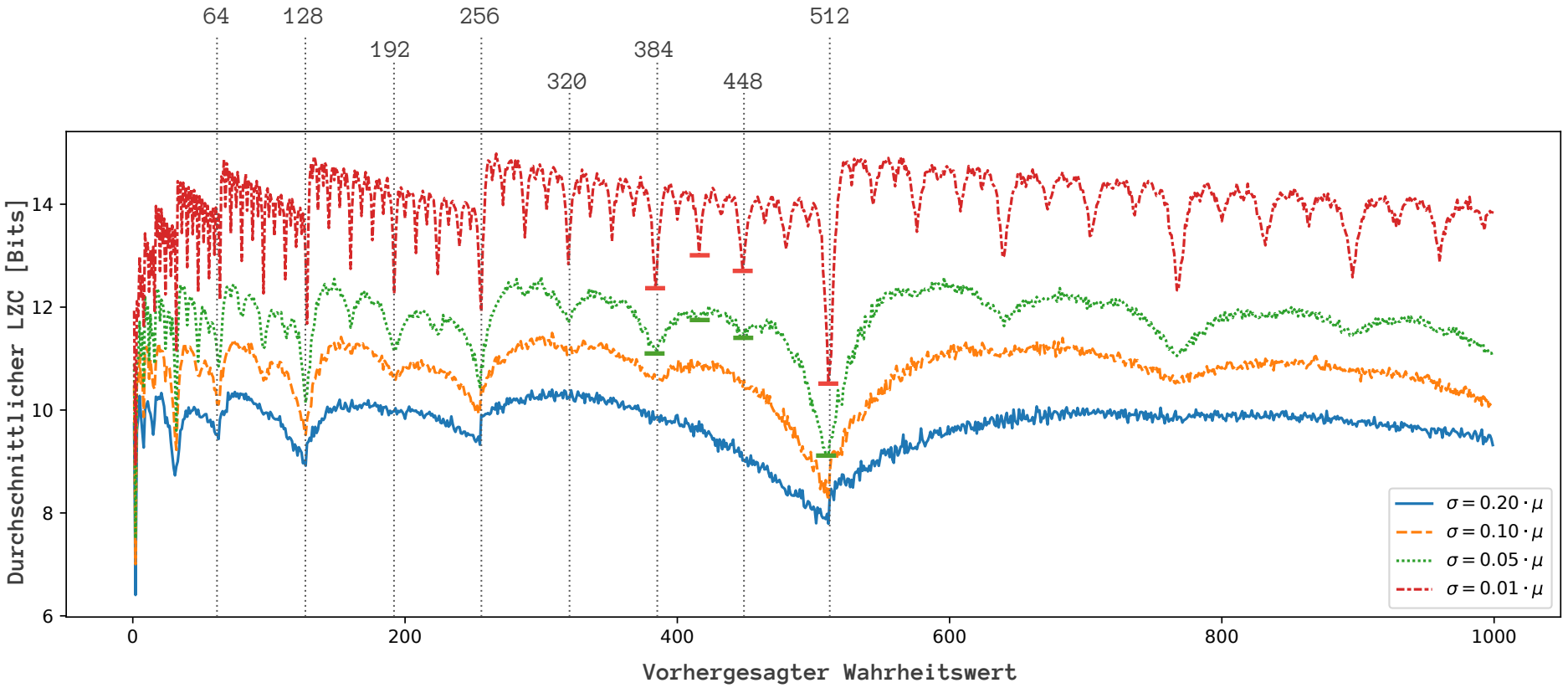
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

- Verschiedene normalverteilte Datensätze (100 Datenpunkte)
  - Erwartungswert (Wahrheit) $\mu \in [0; 1000]$ mit $\mu \in \mathbb{R}$
  - Standardabweichung (Vorhersagen) $\sigma \in \{0.2\mu, 0.1\mu, 0.05\mu, 0.01\mu\}$
- Berechnen des durchschnittlichen LZC



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor



Lösung zum Bitflip-Problem

Karlsruher Institut für Technologie

- Verschiebung der Differenzberechnung in einen anderen Wertebereich

```
Vorhersage = V        V + Shift = V'
  Wahrheit = W        W + Shift = W'        V + Shift = ?
     V ⊕ W = R          V' ⊕ W' = R'
```

- Eigenschaften vom Zielwert
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

Lösung zum Bitflip-Problem

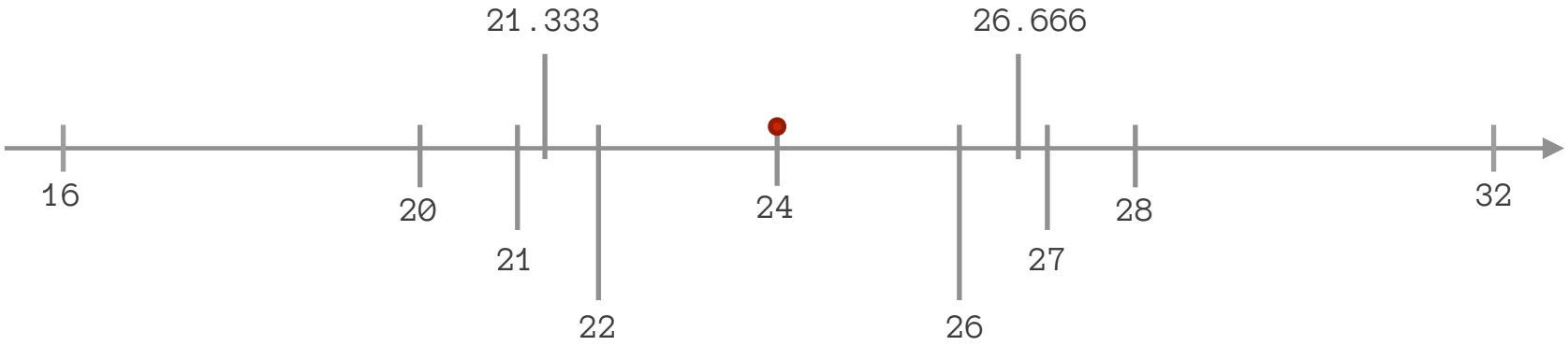
Karlsruher Institut für Technologie

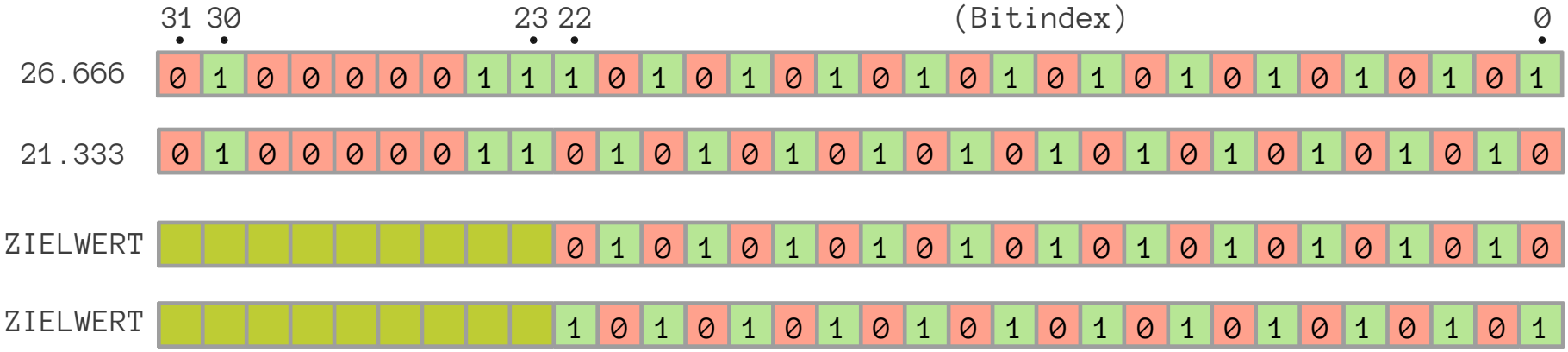
- Verschiebung der Differenzberechnung in einen anderen Wertebereich

```
Vorhersage = V        V + Shift = V'
  Wahrheit = W        W + Shift = W'        V + Shift = ?
     V ⊕ W = R          V' ⊕ W' = R'
```

- Eigenschaften vom Wertebereich
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

Vermeidung des Bitflip-Problems

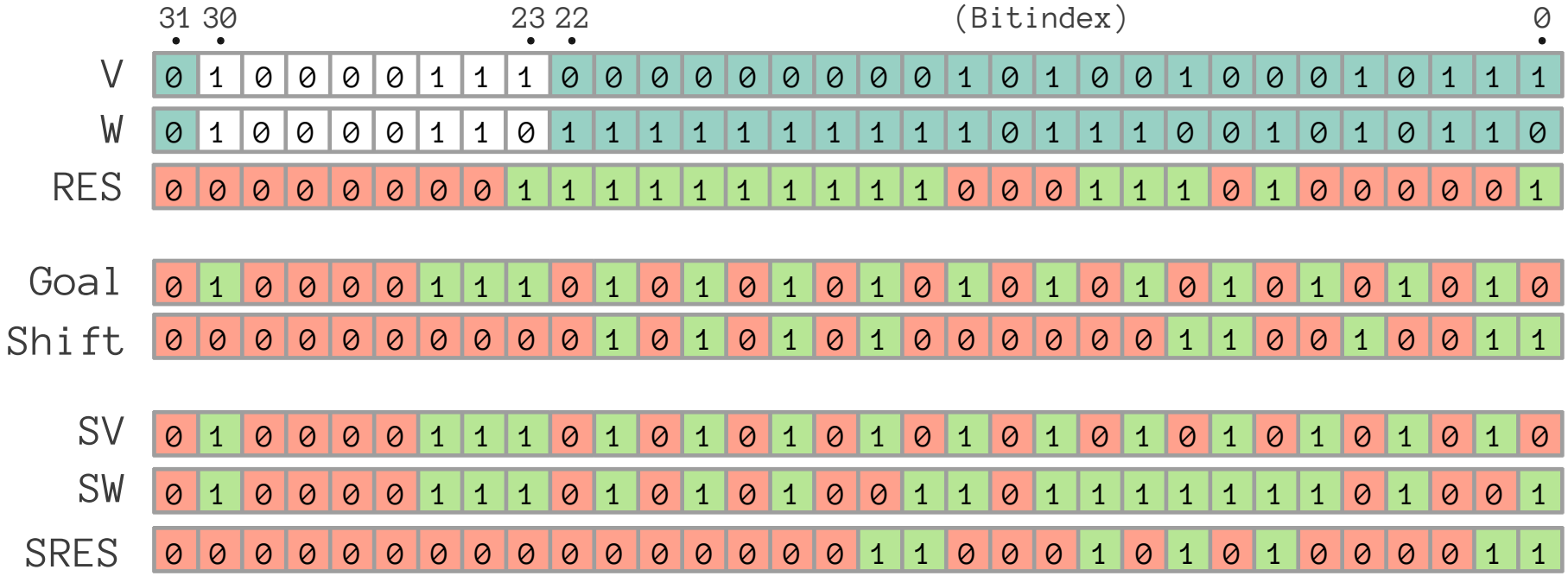




Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
Vorhersage (V):   256.321
Wahrer Wert (W):  255.931
```



LZC: 8 -> 16

Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
 Vorhersage (V):  256.321
Wahrer Wert (W):  255.931          LZC: 8 -> 16
```

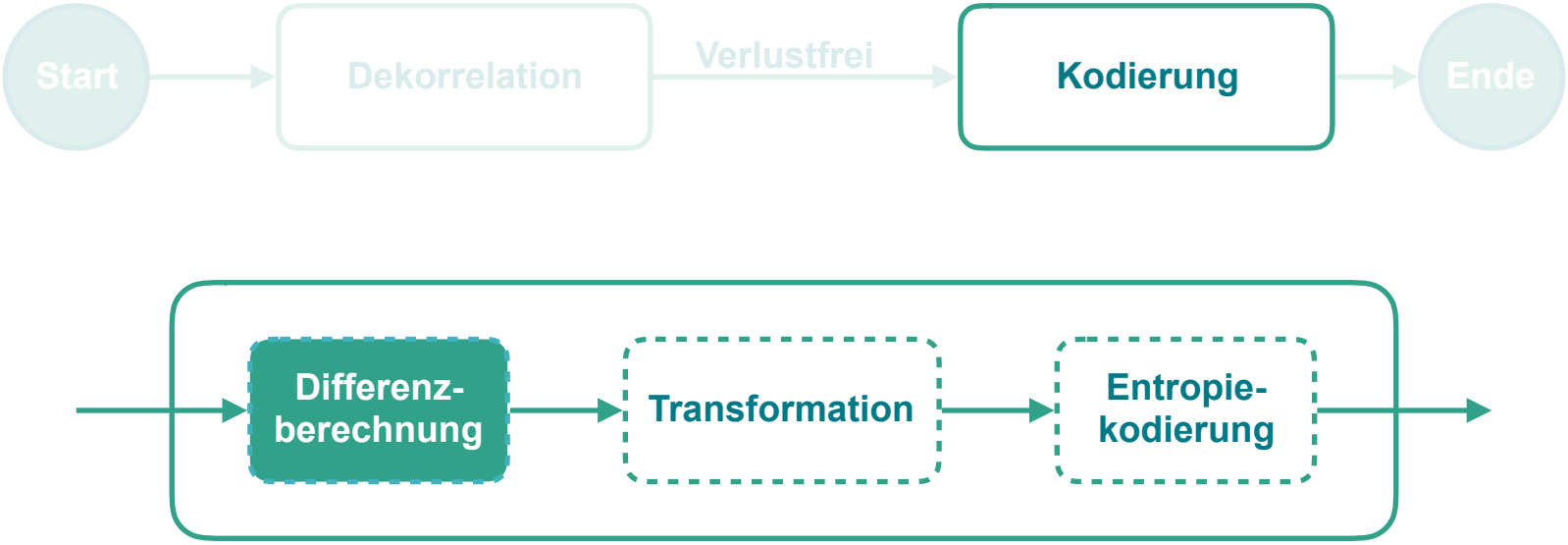
Es funktioniert besser, je näher die Zahlen an Zweierpotenzen liegen

```
 Vorhersage (V):  256.002
Wahrer Wert (W):  255.991          LZC: 8 -> 21
```

Es funktioniert besser, je größer die Zahlen sind

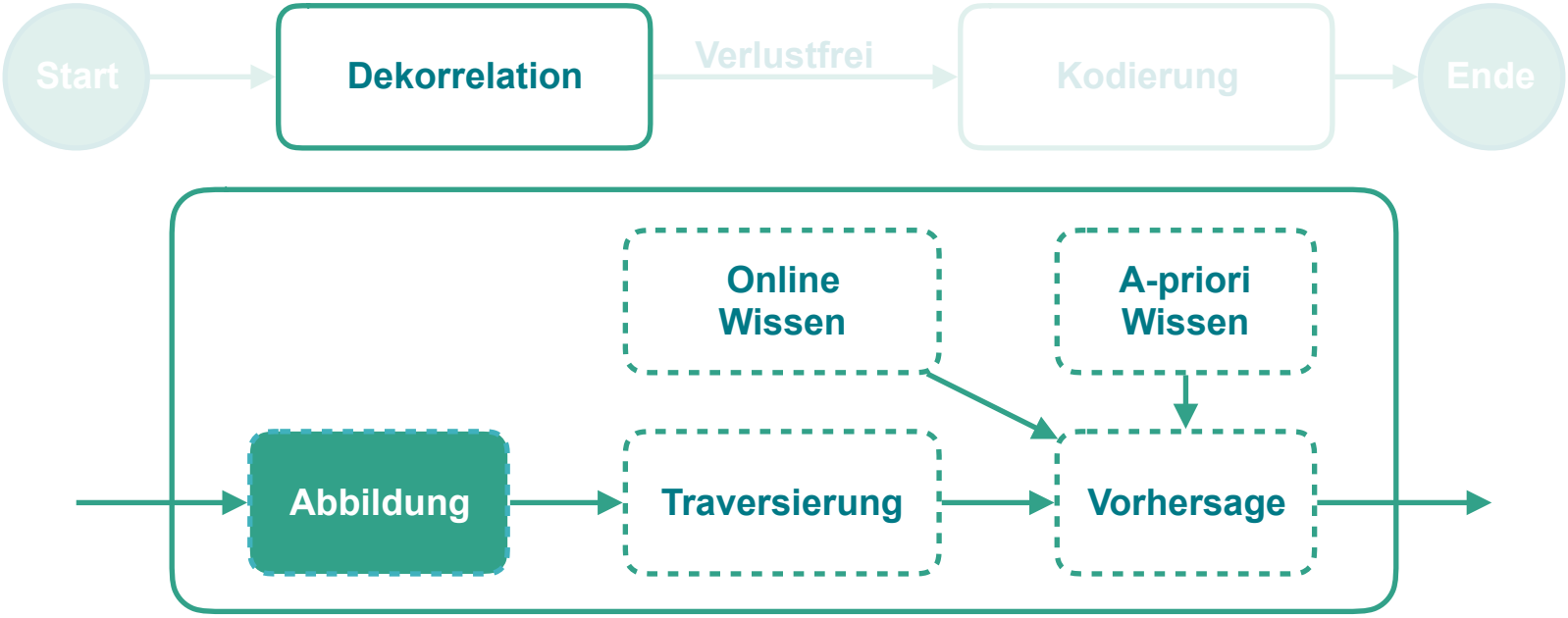
```
 Vorhersage (V): 1024.002
Wahrer Wert (W): 1023.991          LZC: 8 -> 24
```

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Pascal Zip (pzip)

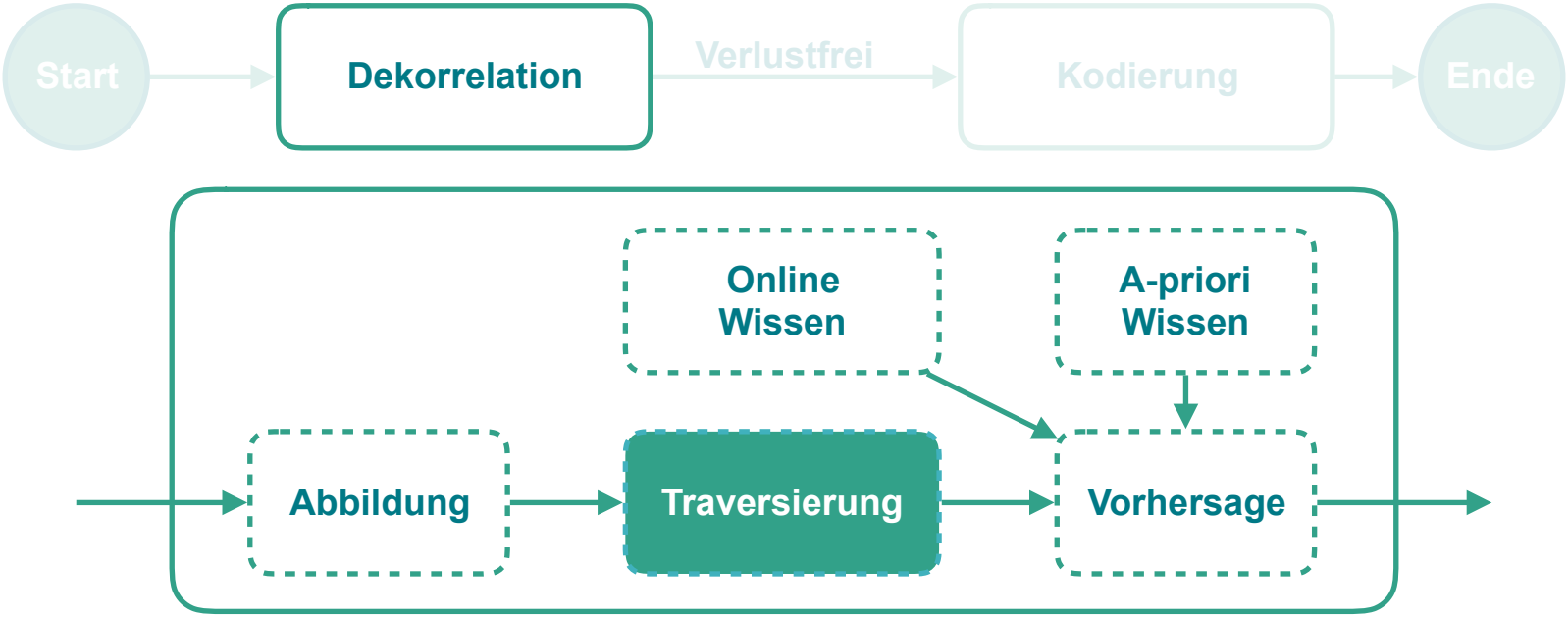
Pascal Zip (pzip)



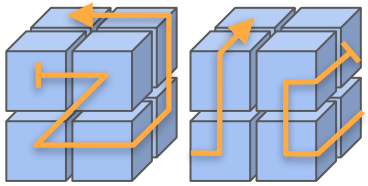
$$m : \mathbb{R} \rightarrow \mathbb{N}$$

```
256.321 -> 1132472599
255.931 -> 1132457558
    . . .
```

Pascal Zip (pzip)

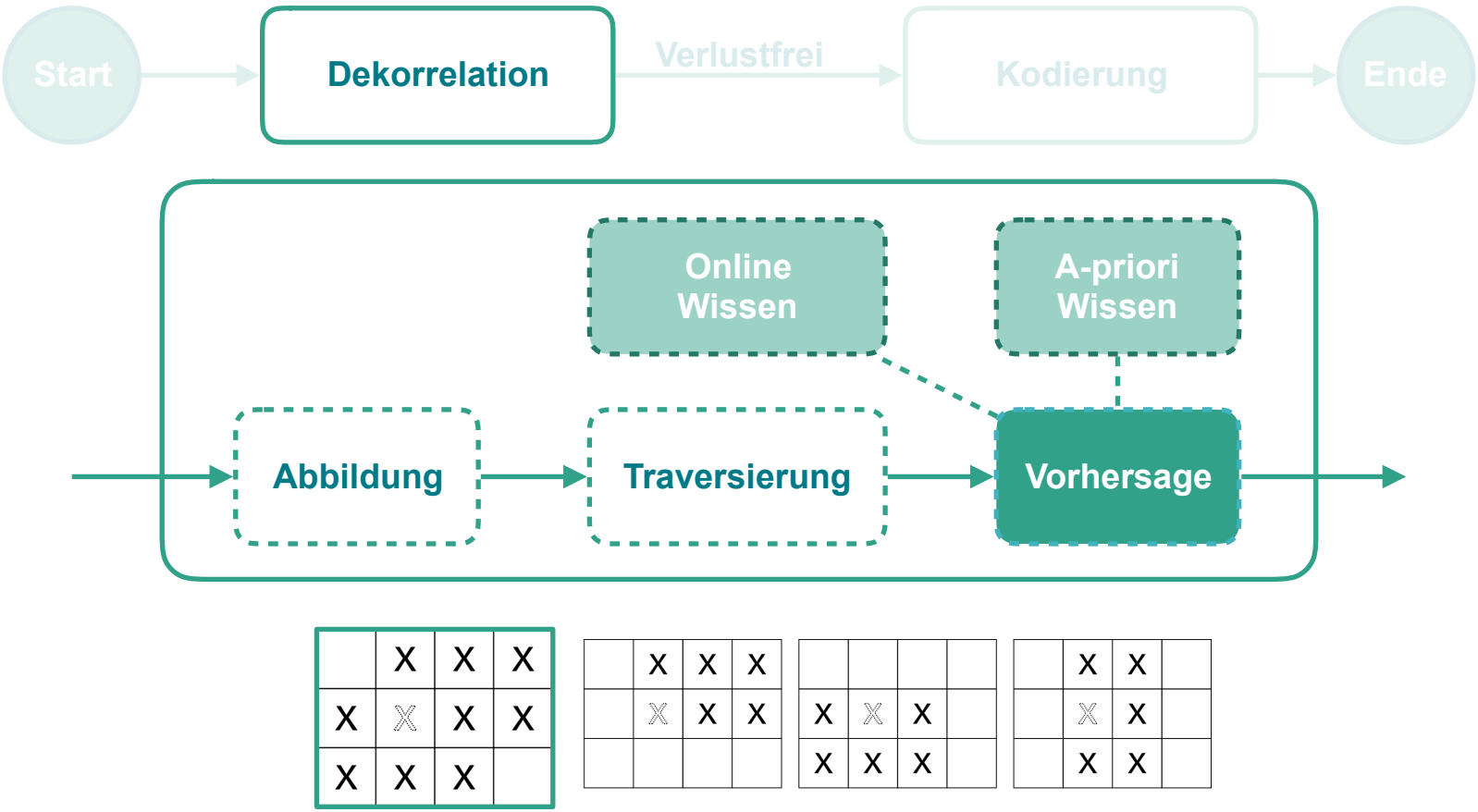


$$t : \mathbb{N} \rightarrow \mathbb{N}$$



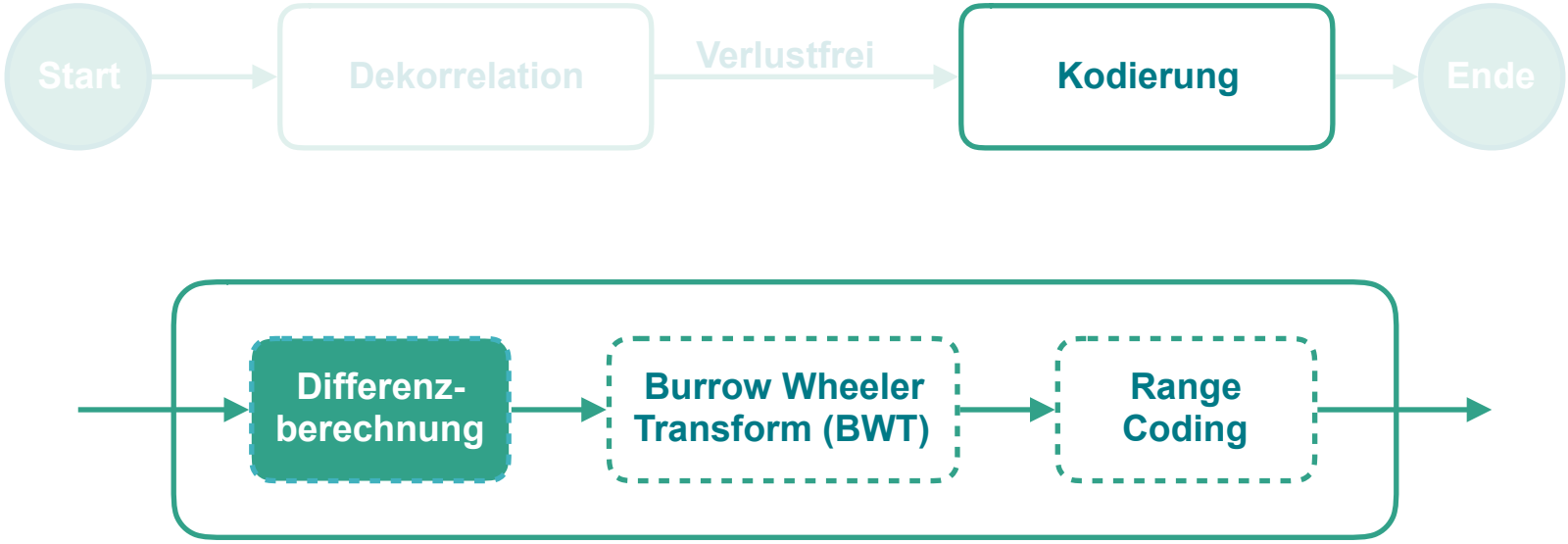
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**
Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression **GI INFORMATIK 2019**

Pascal Zip (pzip)



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output **EGU 2018**
Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using SARIMA **IEEE eScience 2017**
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**

Pascal Zip (pzip)

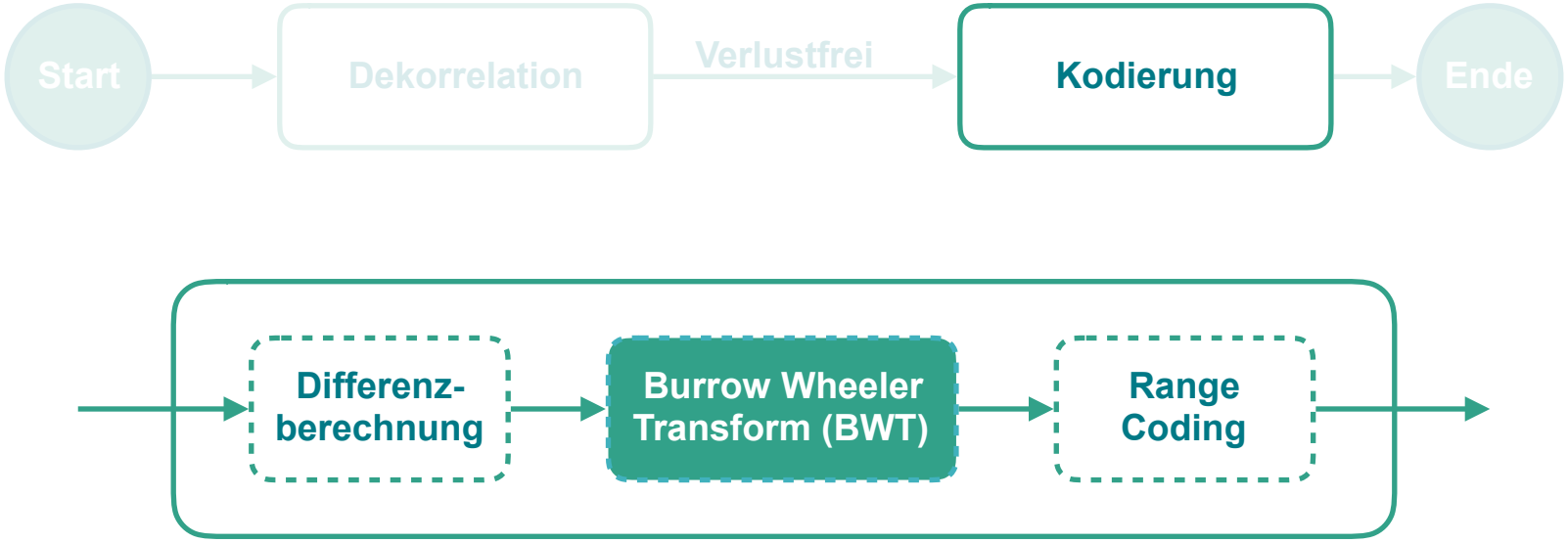



```
V + Shift = V'
W + Shift = W'
  V' ⊕ W' = R'
```

```
R' = 00000000000000111010111001000 1011
LZC = 14, FOC = 03, RES = 101110010001011
```

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms **IEEE eScience 2019**

Pascal Zip (pzip)



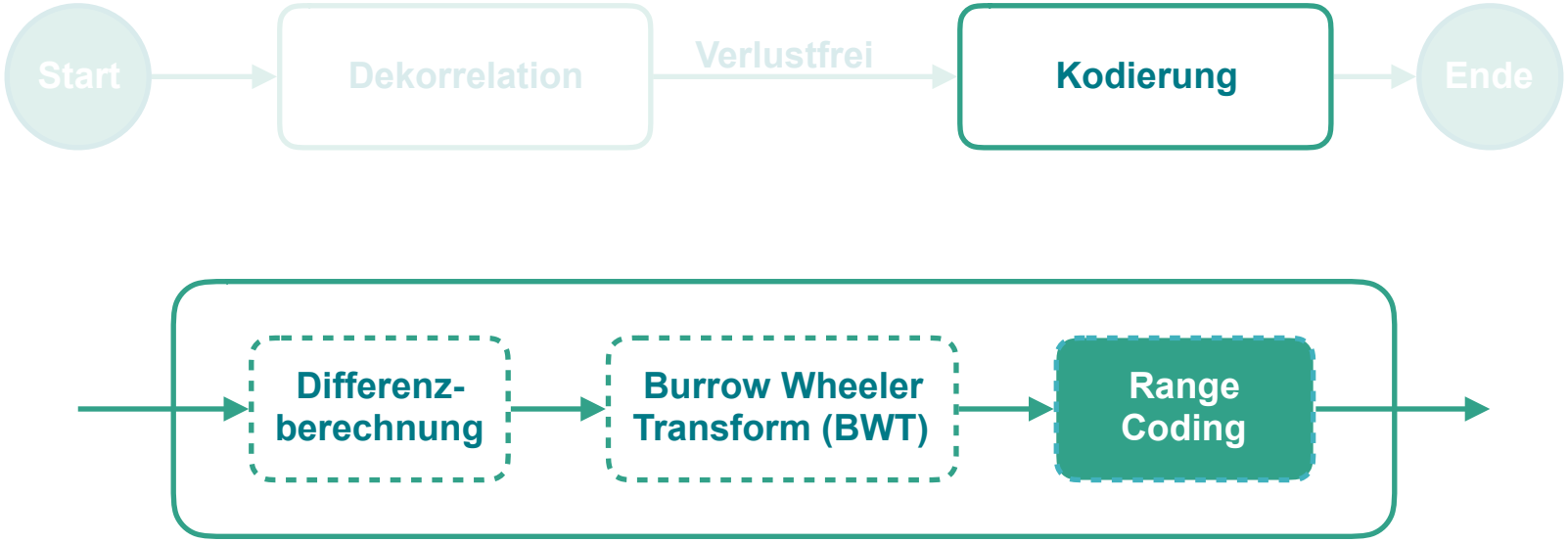
^BANANA$

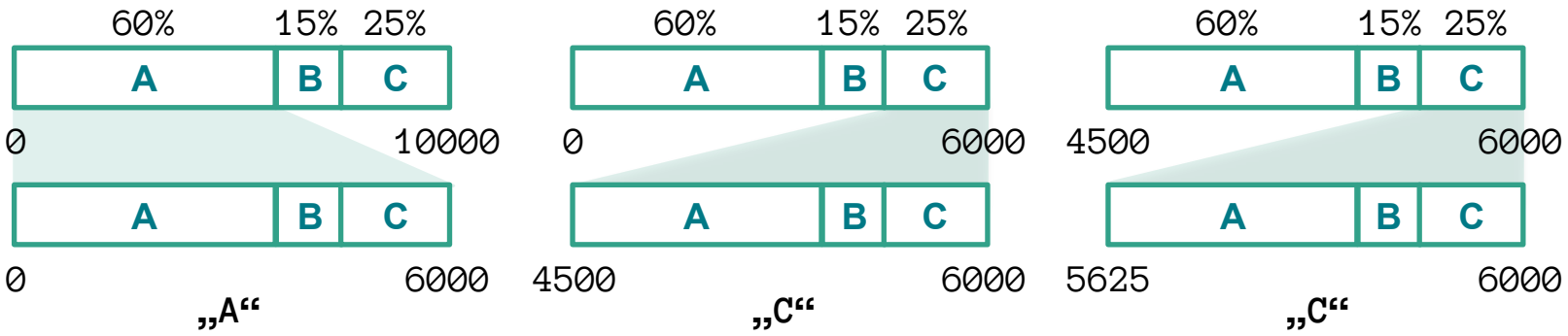
^BANANA$	**A**NANA$^B	ANANA$^**B**
$^BANANA	**A**NA$^BAN	ANA$^BA**N**
A$^BANAN	**A$**^BANAN	A$^BANA**N**
NA$^BANA	**B**ANANA$^	BANANA$^
ANA$^BAN	**N**ANA$^BA	NANA$^B**A**
NANA$^BA	**N**A$^BANA	NA$^BAN**A**
ANANA$^B	^BANANA$	^BANANA**$**
BANANA$^	**$**^BANANA	$^BANAN**A**

BNN^AA$A

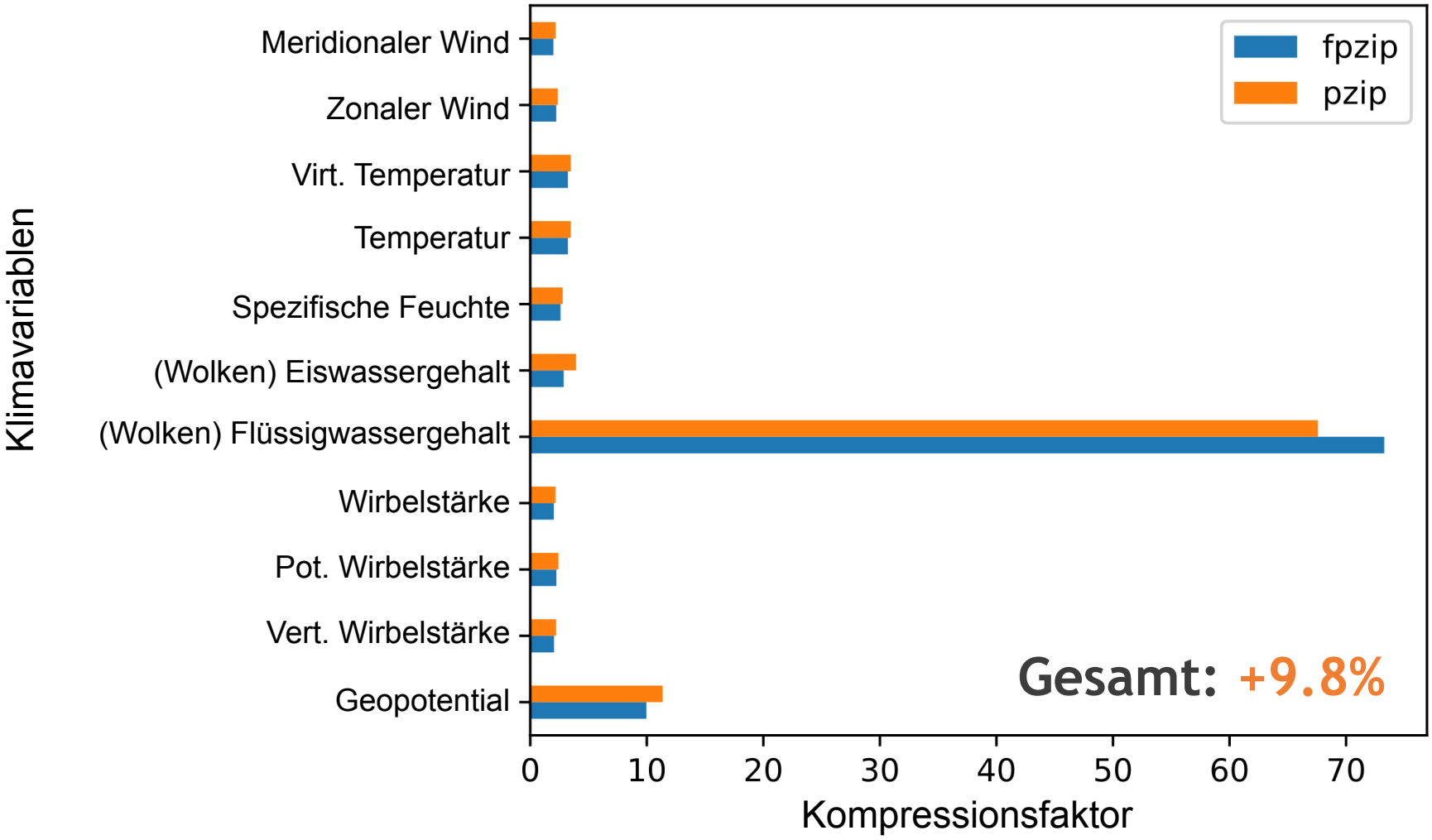
Quelle[3]

Pascal Zip (pzip)

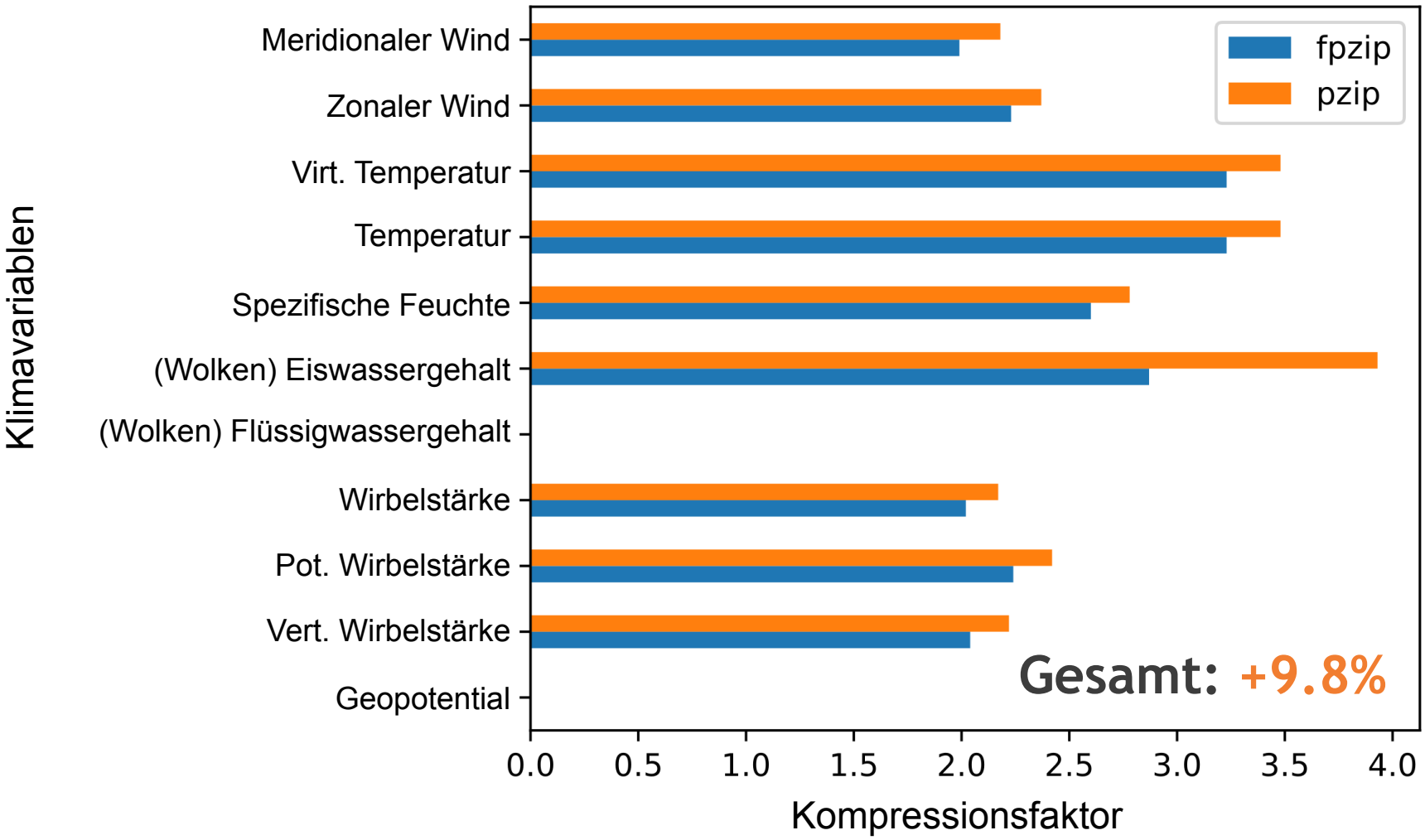




Kompressionsfaktor

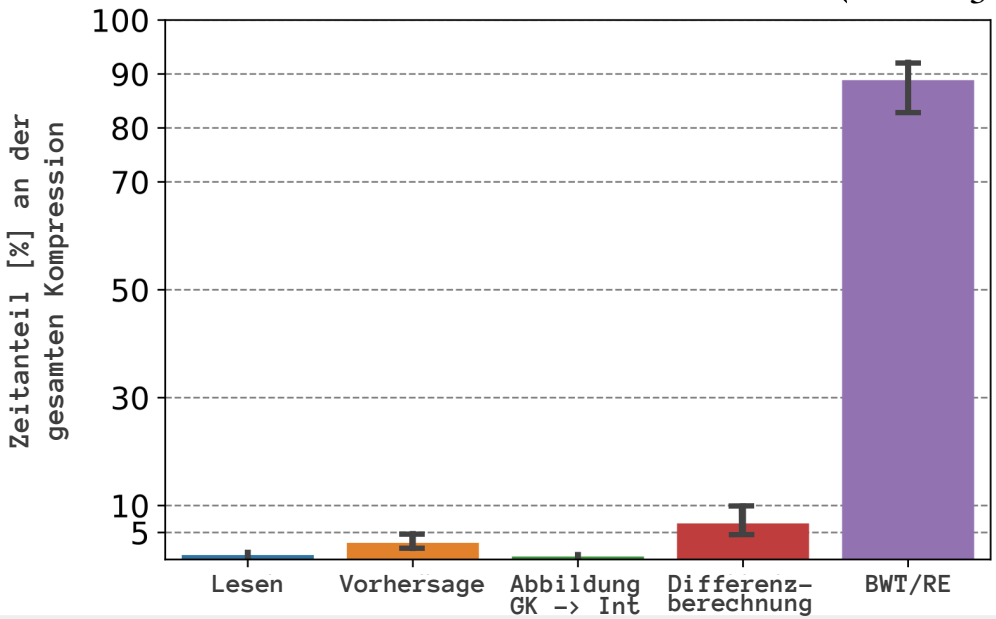


Kompressionsfaktor



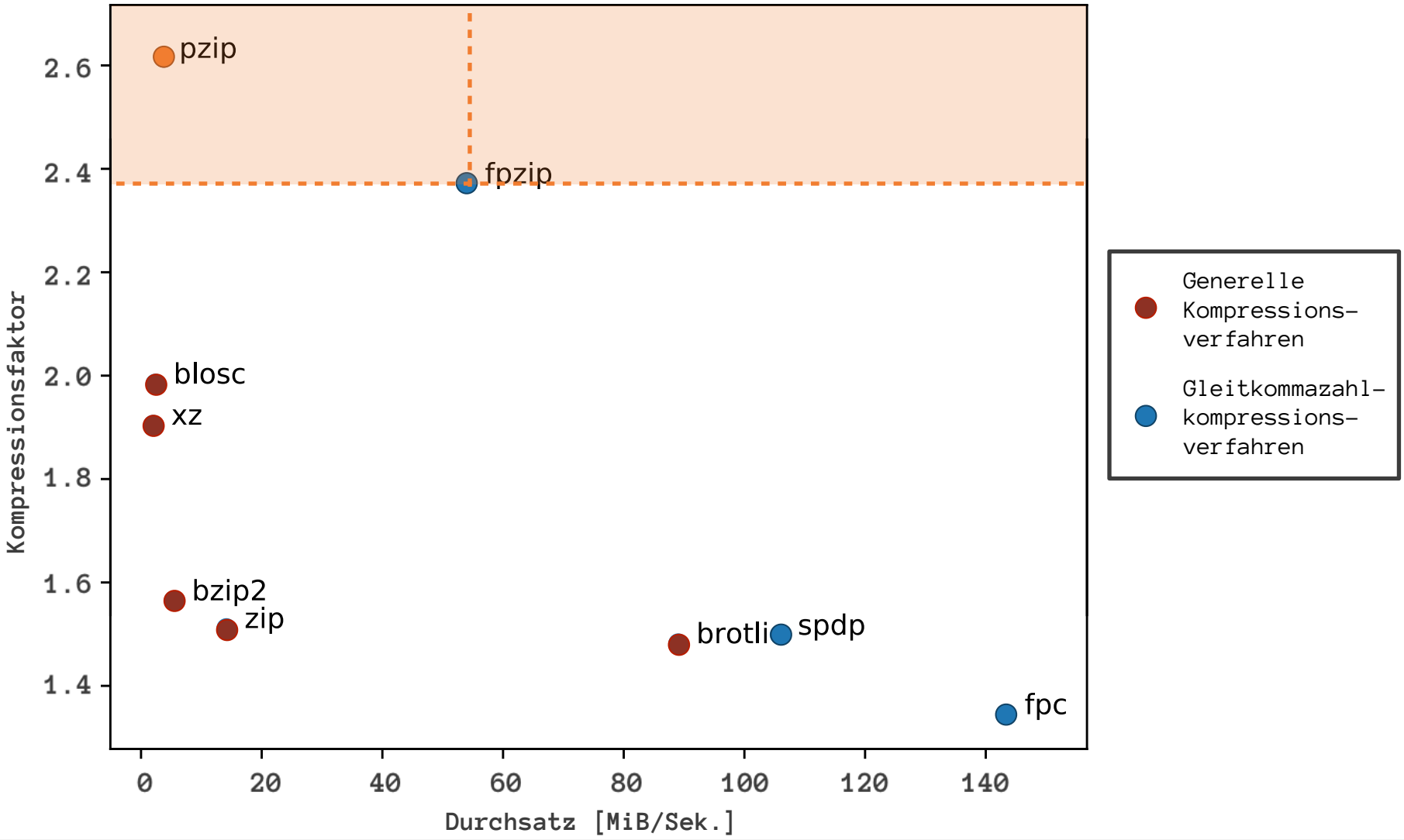
Durchsatz und Komplexität

- Engpass in der aktuellen Implementierung ist BWT/RE
- Laufzeit- und Speicherplatzkomplexität von fpzip $\mathcal{O}(n)$
- Laufzeitkomplexität $\mathcal{O}(n + 4 \cdot \tau \cdot n + n)$
- Speicherkomplexität $\mathcal{O}(\tau \cdot \left(1 + \frac{n}{d_3}\left(\frac{1}{d_2}\left(\frac{1}{d_1} + 1\right) + 1\right)\right) + n \log \sigma)$



τ = Nachbarschaft
$n = d_0 d_1 d_2 d_3$
σ = |Alphabet|

Kompressionsverfahren im Vergleich



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Reduktion

		~2.6	
ERA5:	10.89 PiB	⇨	4.19 PiB
IMK-ASF:	770 TiB	⇨	296 TiB

Open Source

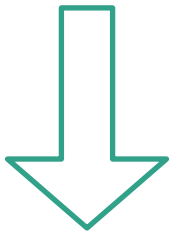
Alle Programmbeispiele und Daten sind öffentlich zugänglich (github.com, GPLv3)

Beitrag

Andere Kompressionsverfahren können einzelne Entwicklungen aus der Arbeit aufgreifen und einbauen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor erfüllt ✔



Verlustfreie Kompression von Klimadaten

Vielen Dank