

# Adaptive Lossy Compression of Environmental Data

by Uğur Çayoğlu, Peter Braesicke, Tobias Kerzenmacher, Jörg Meyer and Achim Streit

STEINBUCH CENTRE FOR COMPUTING (SCC) &  
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

# What is the result?

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

Importance of  
compression for  
climate research



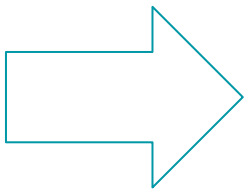
Introduction and  
description of  
proposed  
method

Application on  
environmental  
indices used in  
climate research



# Importance of compression

- Environmental data is 4D (longitude, latitude, altitude, time)
- Weather v climate simulations
- Current European ReAnalysis (ERA5) dataset needs 2.26 TiB p.a. and variable



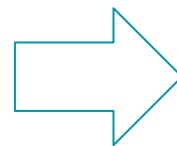
Generate a  
compression method  
specific for  
environmental data

**But how?**

# Compression 101

- 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.

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



Environmental indices

# Environmental indices

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

# What are success criteria?

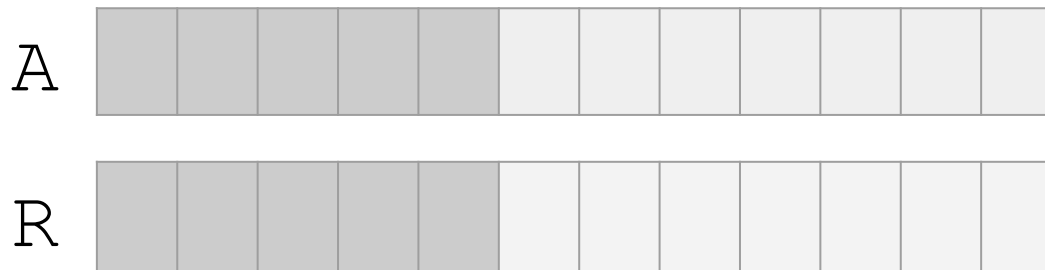
- **Compression ratio**
  - $\text{Filesize (after)} / \text{Filesize (before)}$
- Memory usage
- Compression/Decompression time

Additional criteria in case of lossy compression:

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

# Quality criteria for compressed indices

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



$$\text{Corr}_{s,e}(A, R) = 1.0$$



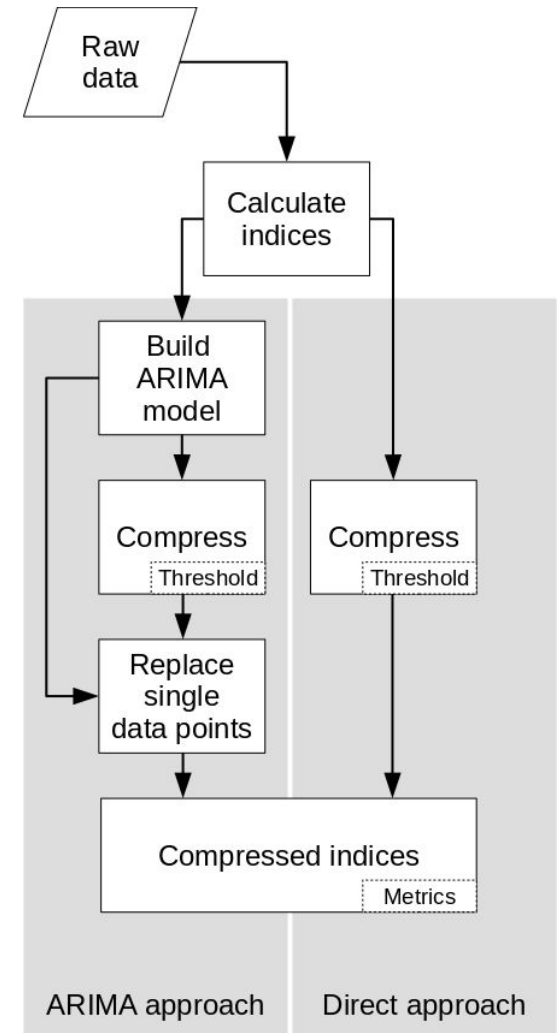
# Direct & proposed approach

## Direct approach

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

## ARIMA approach

From us proposed approach.

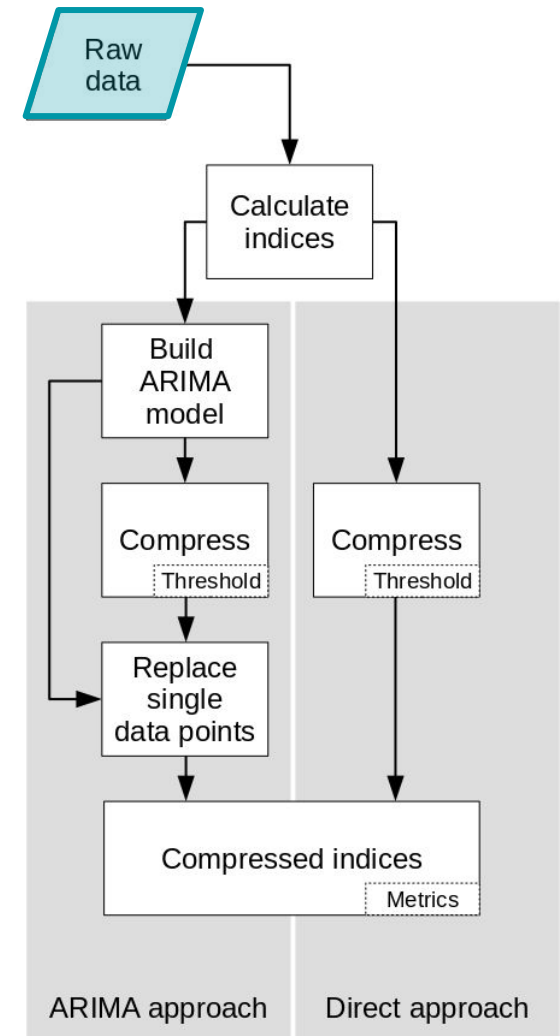


# Raw data

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

↓  
daily

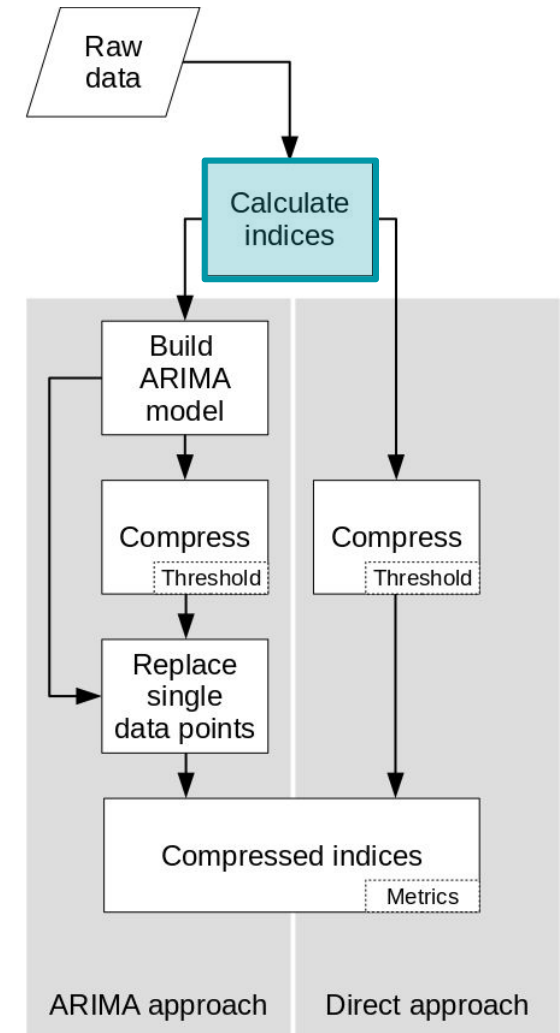
↓  
monthly



# Calculate indices

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

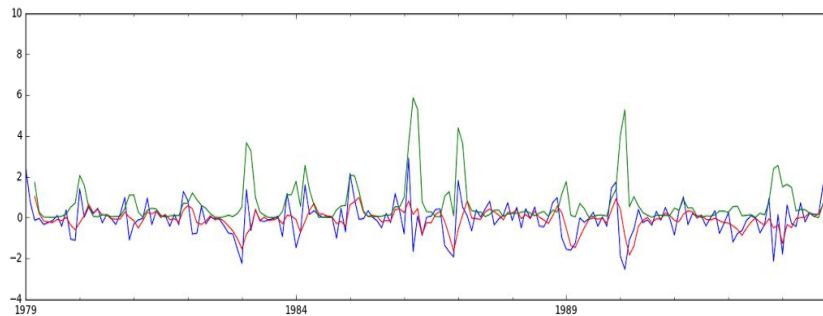
T = temperature, u = westerly wind, p = pressure



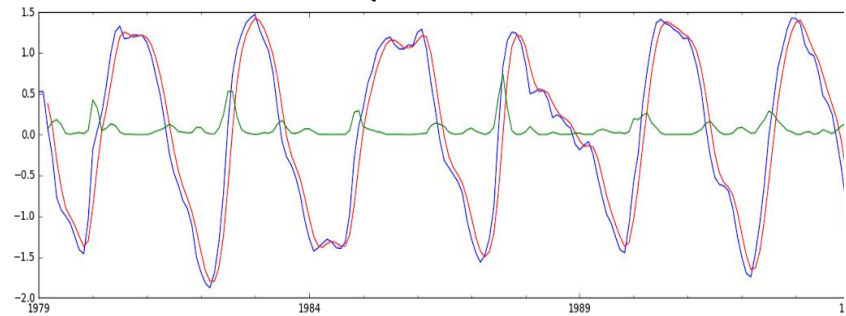
# Calculate indices

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

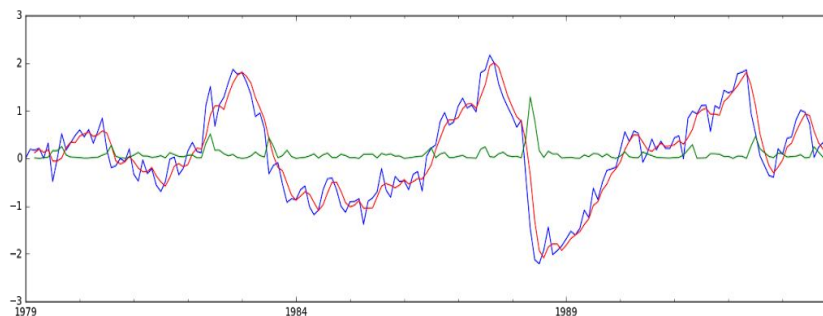
**NAO**



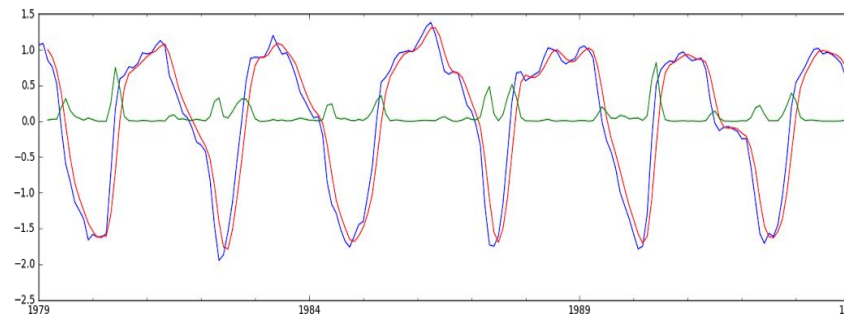
**QBO30**



**ENSO34**



**QBO50**

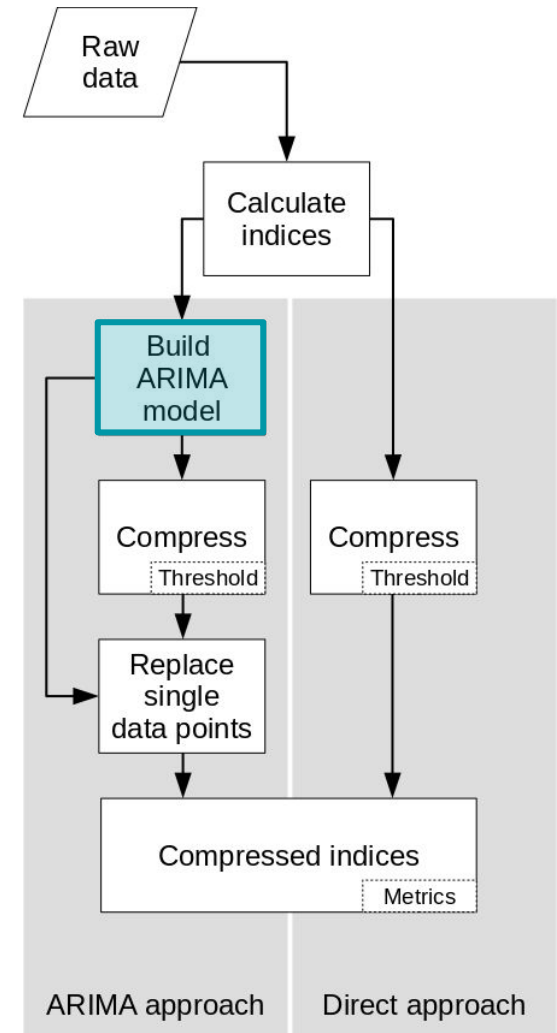


# Build an Auto Regressive Integrated Moving Average model

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:

$$\text{ARIMA}(p, d, q)(P, D, Q)_s$$

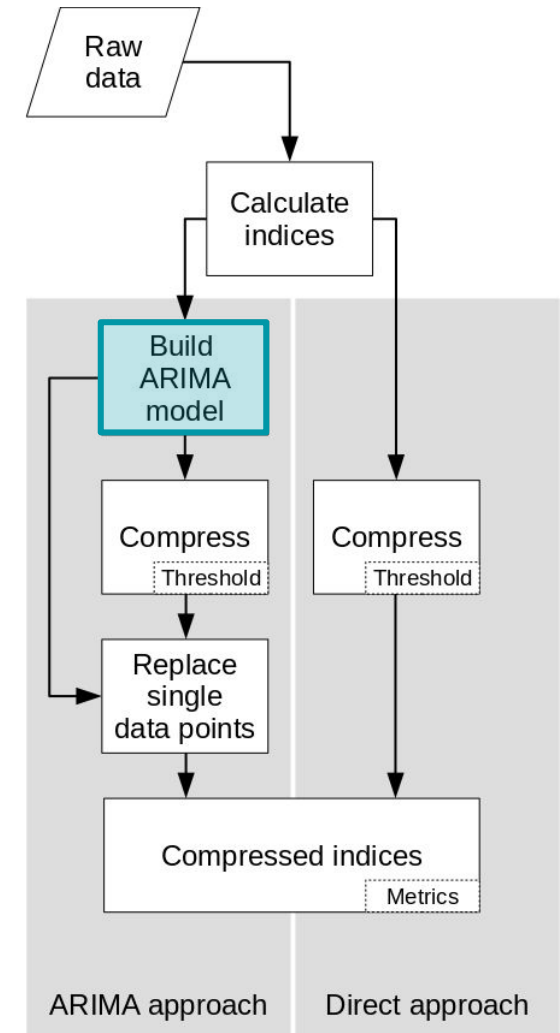
$p$  = autoregressive order

$d$  = differential order

$q$  = moving average order

$s$  = seasonal period

$P, D, Q$  = appropriate seasonal order



# Build an Auto Regressive Integrated Moving Average model

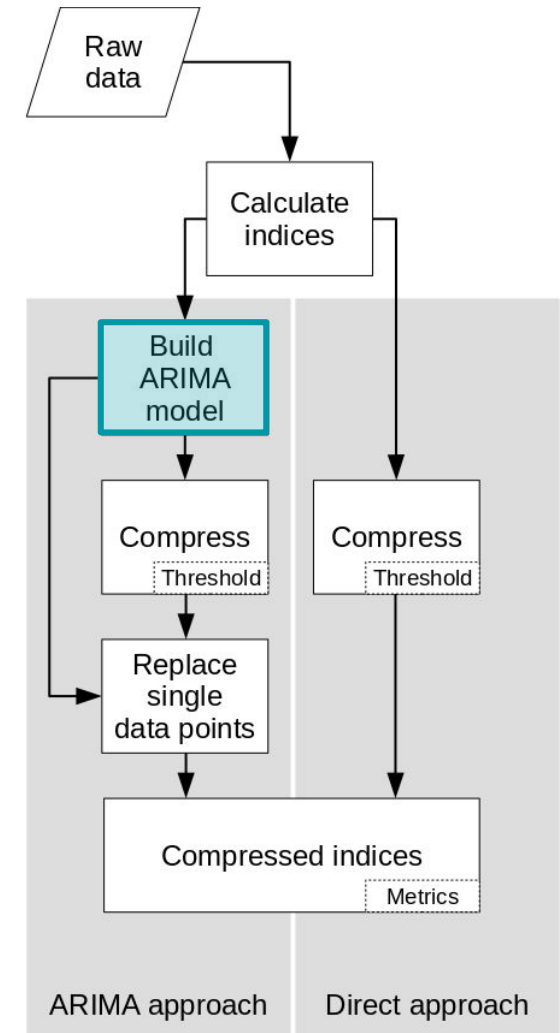
Notation used for the ARIMA model:

$$\text{ARIMA}(p, d, q)(P, D, Q)_s$$

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.

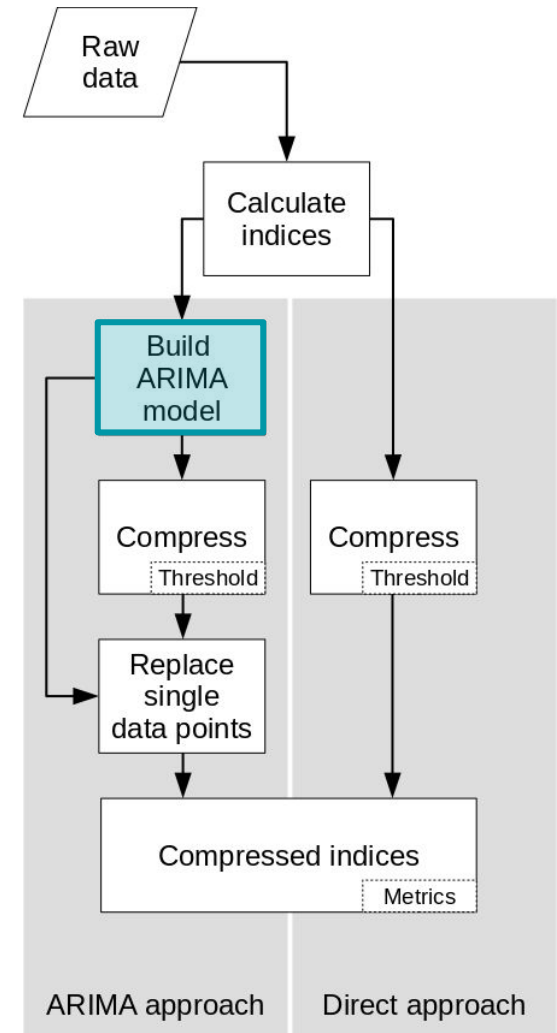


# Build an Auto Regressive Integrated Moving Average model

ARIMA(p, 0, q)(0, 0, 0)<sub>0</sub>

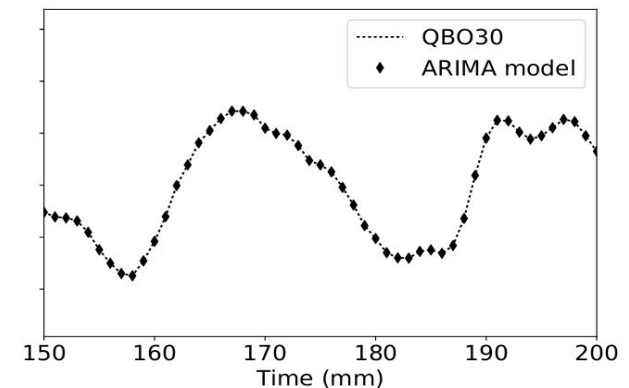
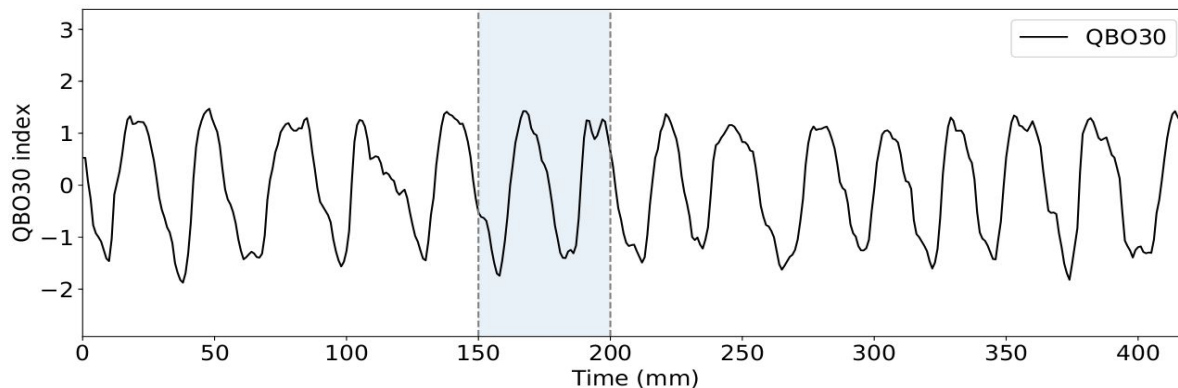
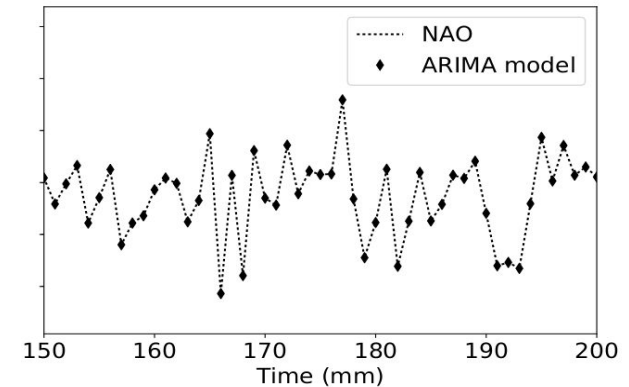
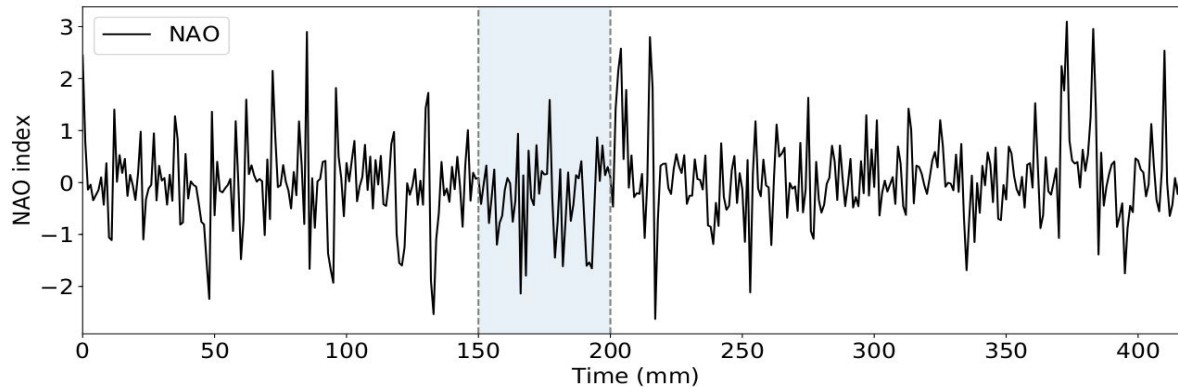
$$x_i = x'_i + \epsilon_i$$

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





# How good is the reconstruction of the ARIMA model?



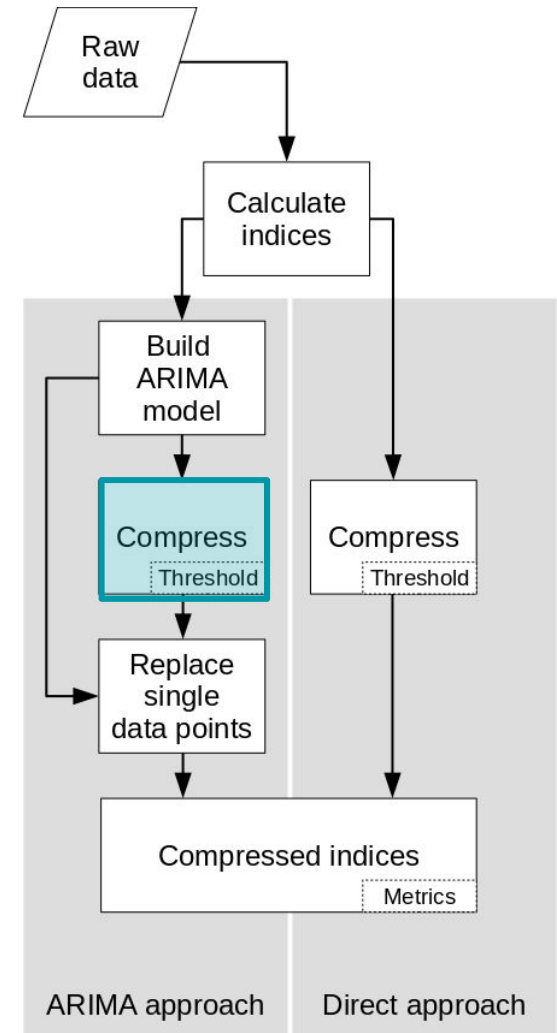
# Build an Auto Regressive Integrated Moving Average model

Timeline	ARIMA Model			
	Monthly		Daily	
	Model	RMSD	Model	RMSD
<b>ENSO34</b>	ARIMA(3,0,2)(1,0,0) <sub>12</sub>	5.067e-8	ARIMA(5,2,4)(0,0,0) <sub>0</sub>	4.686e-4
<b>NAO</b>	ARIMA(1,0,0)(1,0,0) <sub>12</sub>	8.195e-9	ARIMA(2,0,2)(0,0,0) <sub>0</sub>	1.440e-7
<b>QBO30</b>	ARIMA(2,0,3)(1,0,0) <sub>12</sub>	1.0877e-7	ARIMA(5,0,4)(0,0,0) <sub>0</sub>	1.084e-7
<b>QBO50</b>	ARIMA(1,1,1)(1,0,1) <sub>12</sub>	2.909e-6	ARIMA(5,0,4)(0,0,0) <sub>0</sub>	4.488e-8



# 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
<b>ENSO34</b>	1.043	1.024	1.036	1.009
<b>NAO</b>	1.043	1.033	1.033	1.026
<b>QBO30</b>	1.038	1.005	1.032	0.961
<b>QBO50</b>	1.045	1.014	1.033	0.969

## Exp. 2:

### Lossy compression with threshold $T = 1e-5$

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

## Exp. 2:

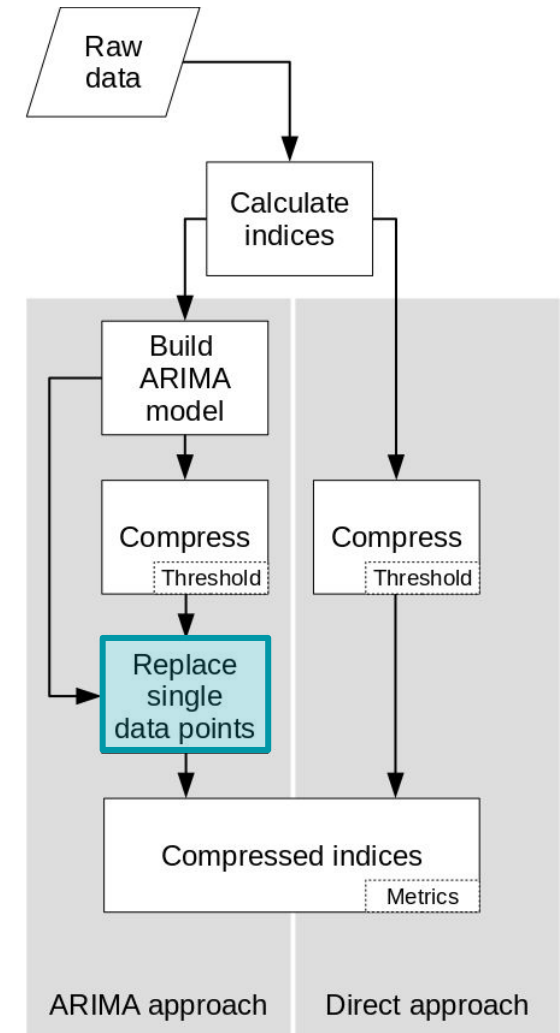
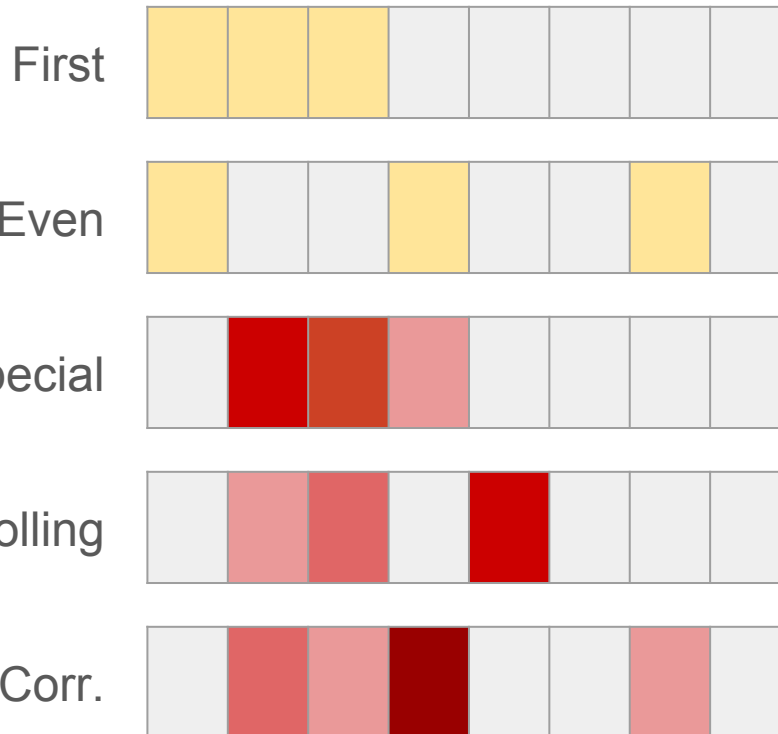
### Lossy compression with threshold $T = 1e-5$

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



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

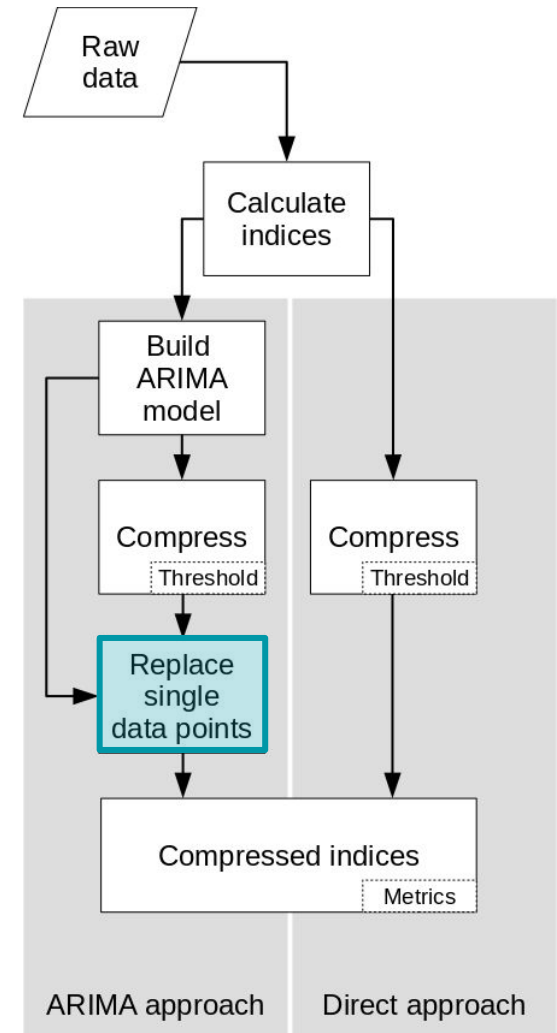
Methods for finding data points to be replaced:



# Exp. 3:

## Lossy compression (w/ replacement)

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





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

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

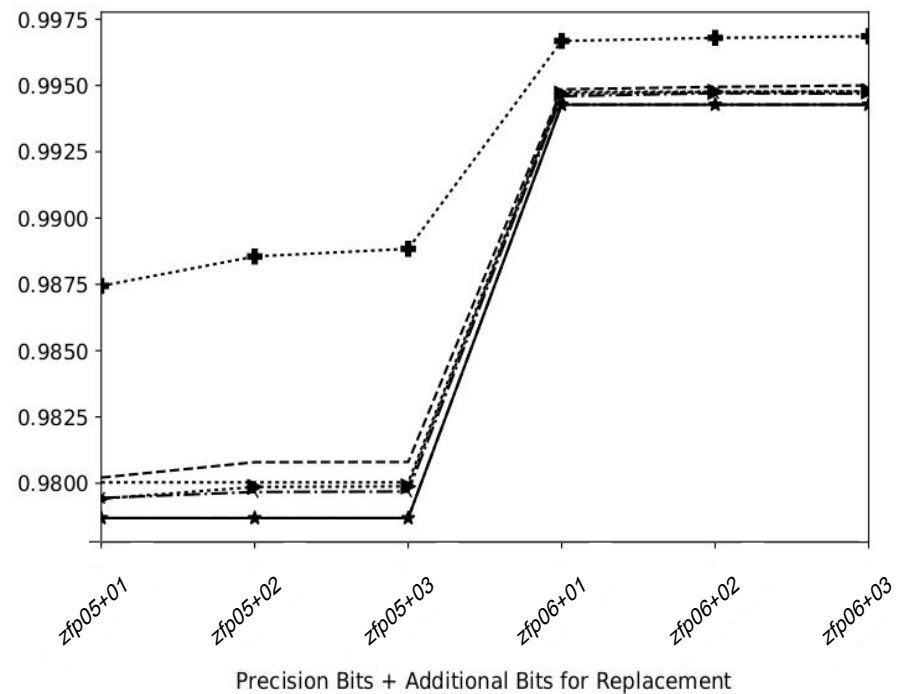
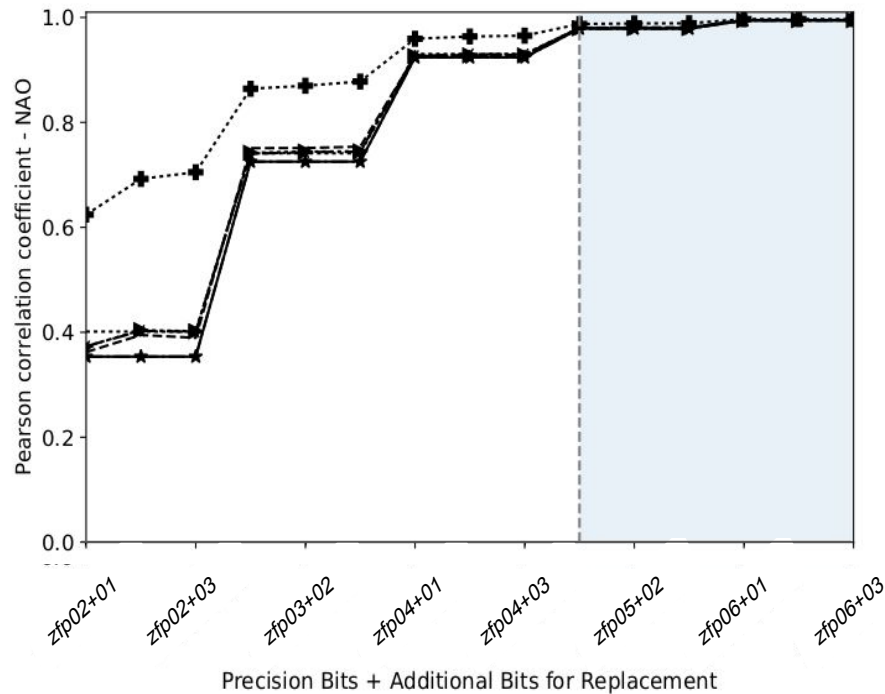
# Exp. 3: Lossy compression (w/ replacement)

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

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

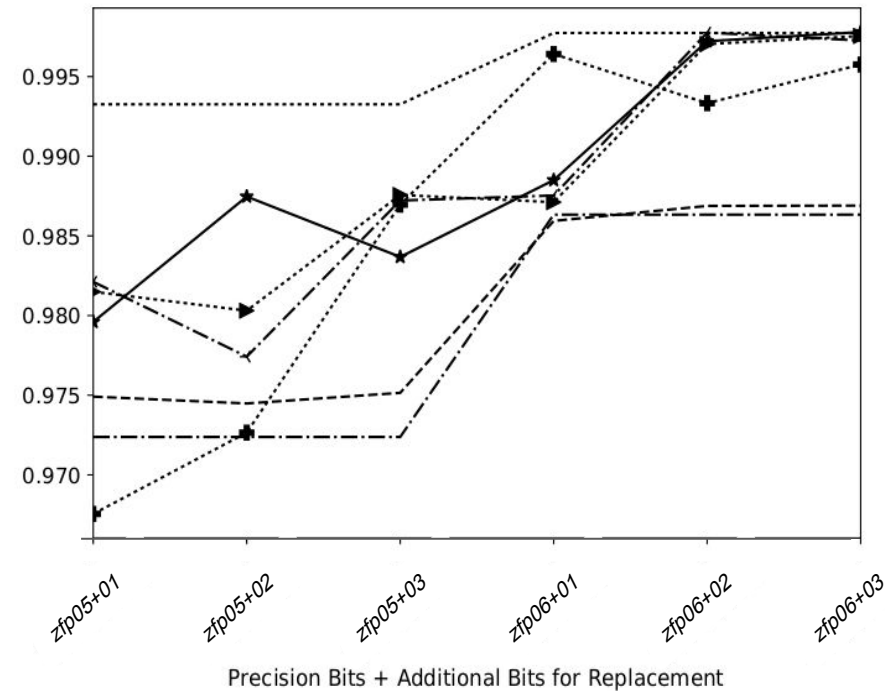
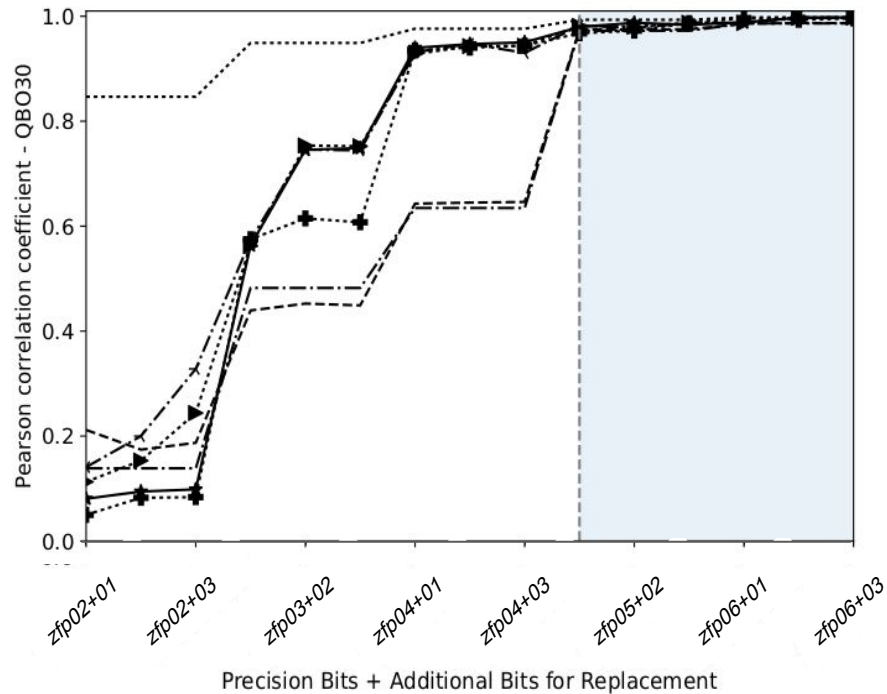
	Correlation Coefficient (Monthly, 10%, Method: Special)				
	zfp02	zfp03	zfp04	zfp05	zfp06
NAO	.354 ~25%	.725 ~15%	.924	.979	.994
+1 Bit	.624	.864	.959	.987	.997
+2 Bit	.692	.870	.964	.989	.997
+3 Bit	.705	.878	.965	.989	.997
QBO30	.139 -10%	.482 ~10%	.635 ~30%	.972 -0.4%	.986
+1 Bit	.050	.575	.935	.968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



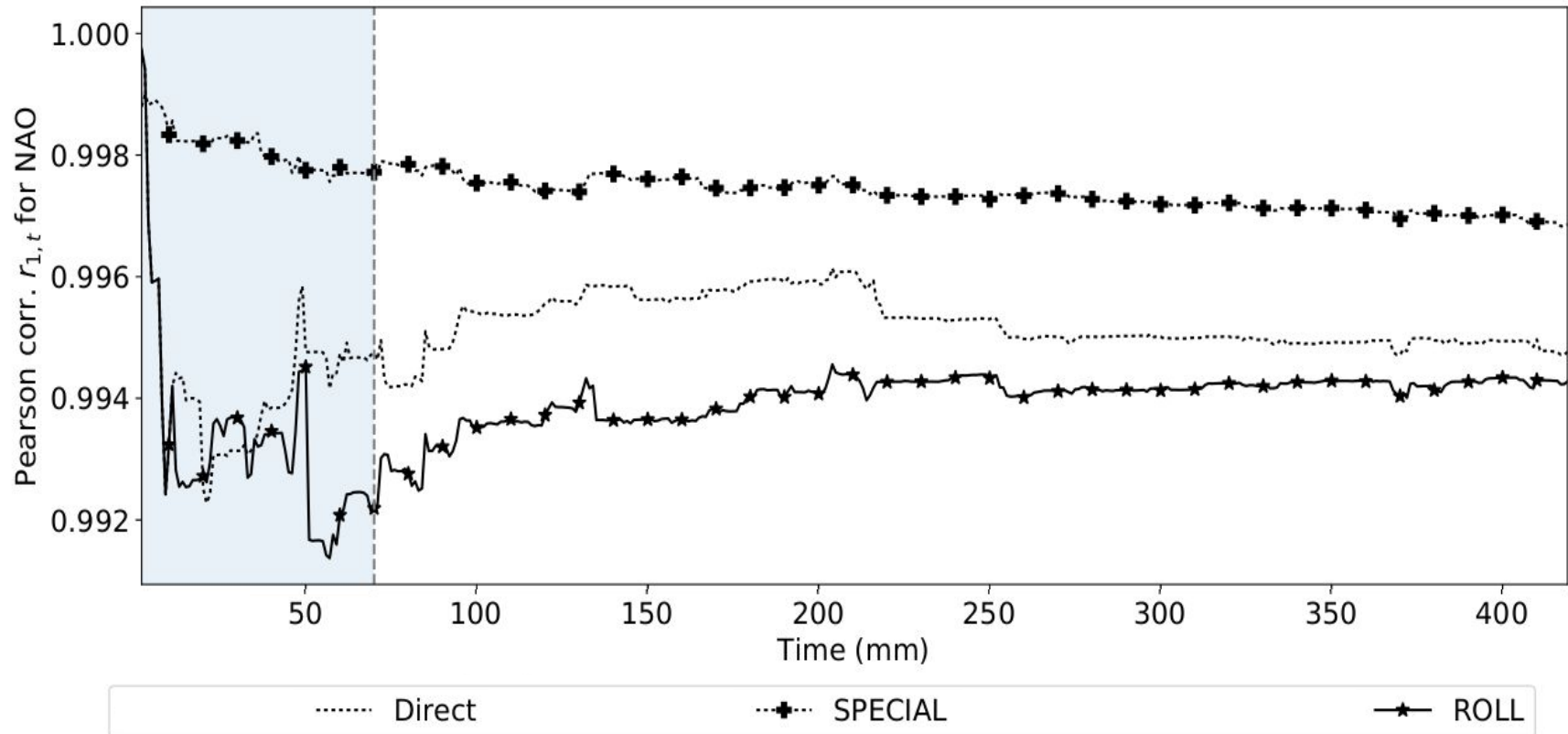
-.- ARIMA      ..... Direct      -<- CUMCORR      -+ SPECIAL      -> FIRST      --- EVENLY      -★ ROLL

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

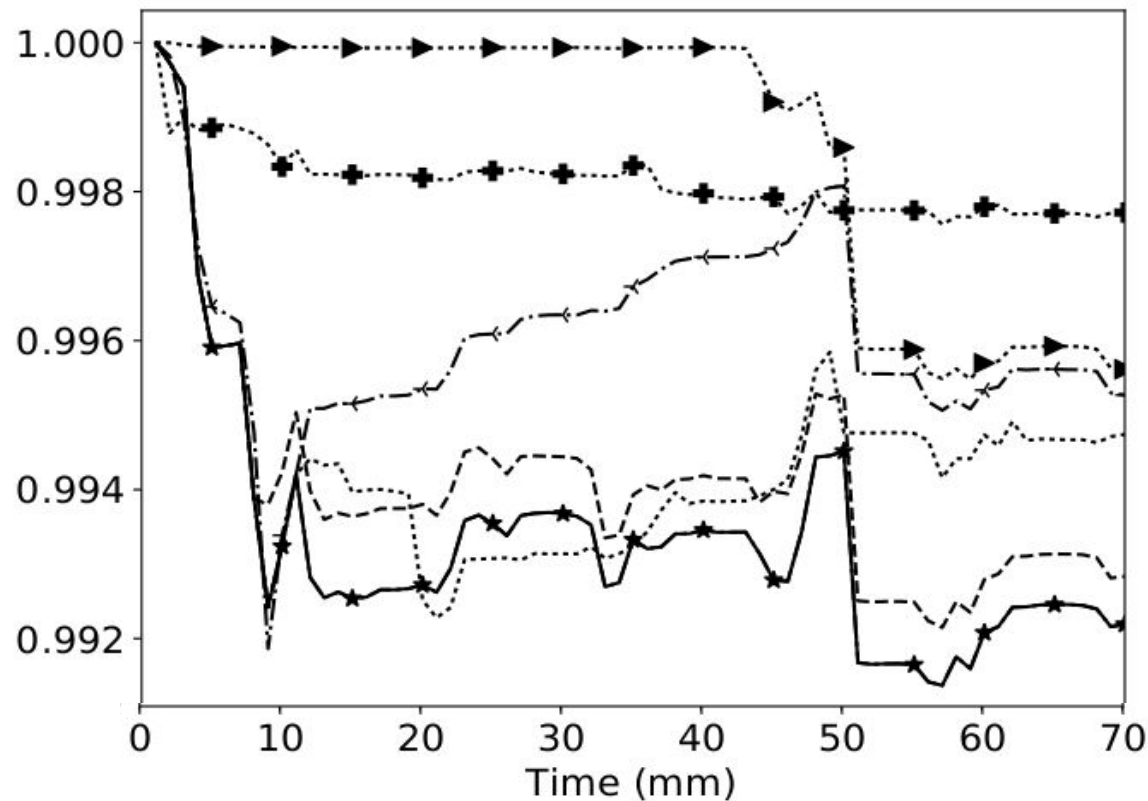


- - - ARIMA      ..... Direct      - - - CUMCORR      - - - SPECIAL      - - - FIRST      - - - EVENLY      - - - ROLL

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



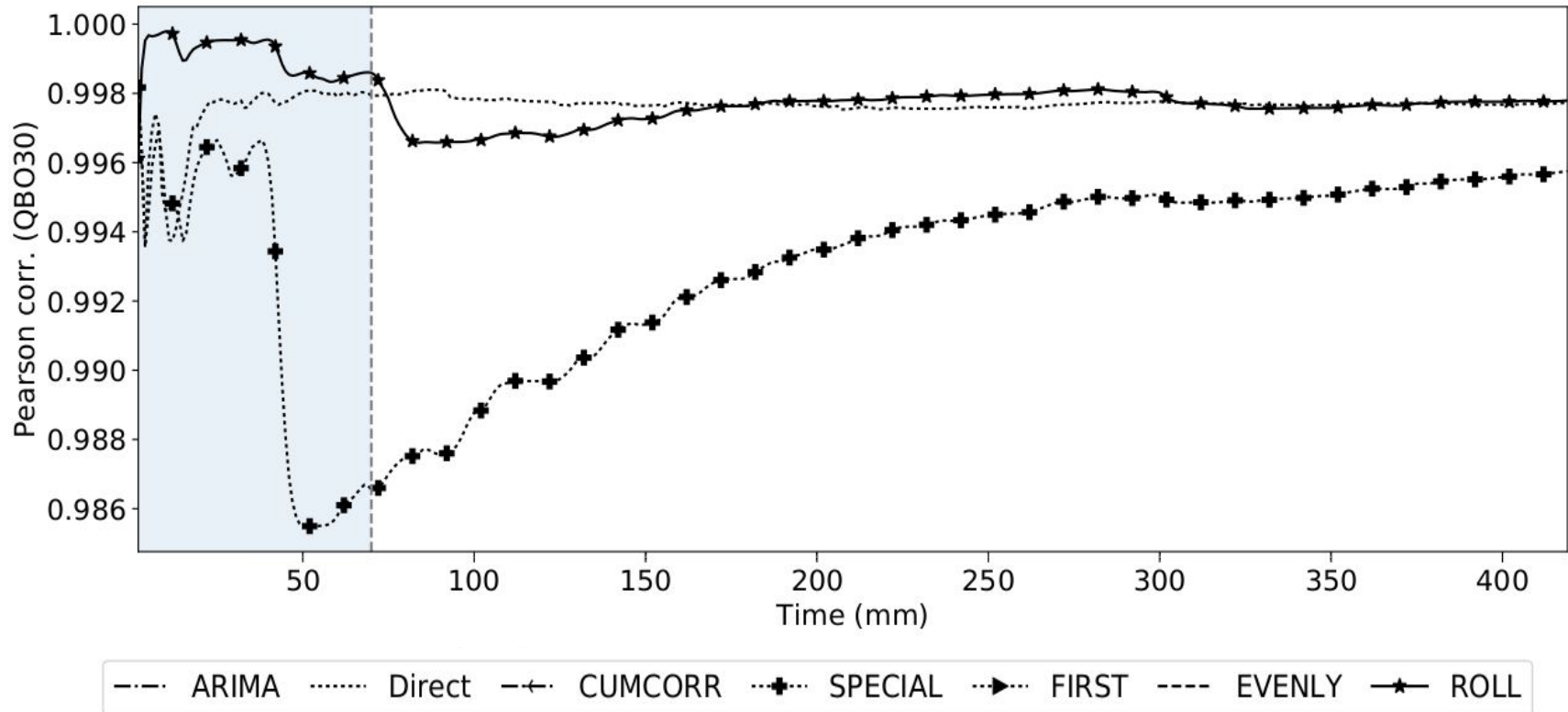
# Exp. 3: Lossy compression (w/ replacement) over time series NAO



--- ARIMA    ..... Direct    -+- CUMCORR    -+ SPECIAL    -+ FIRST    ---- EVENLY    -★ ROLL

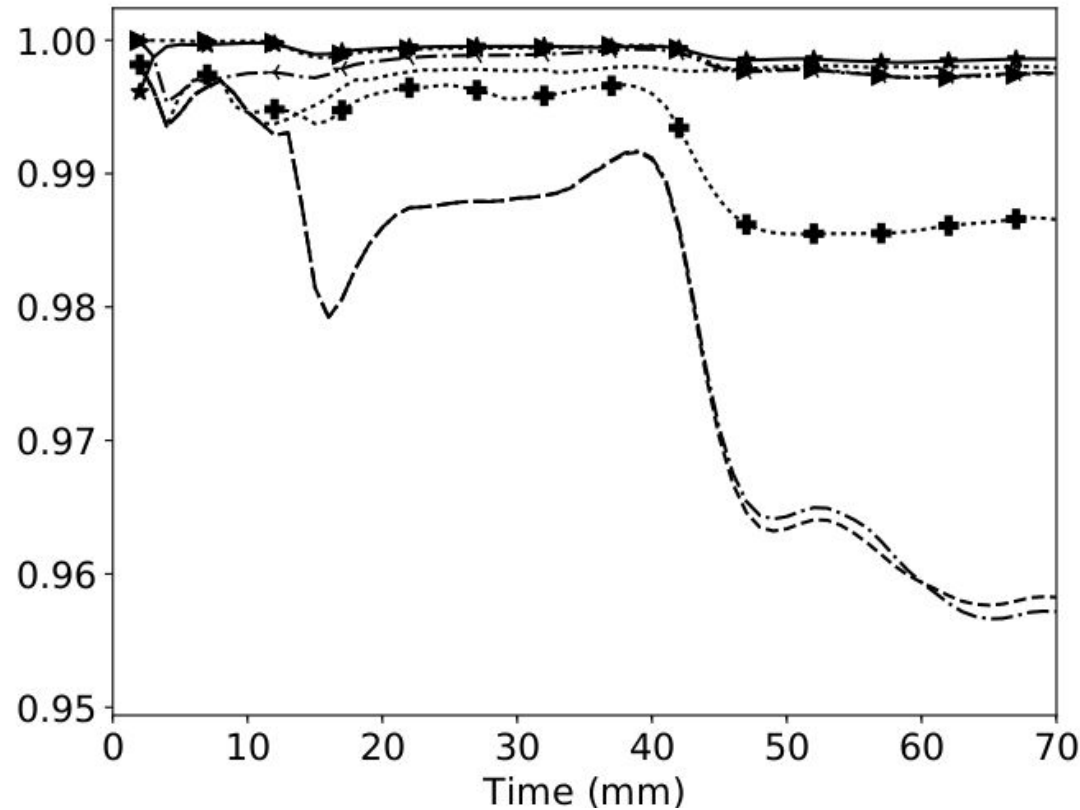


# Exp. 3: Lossy compression (w/ replacement) over time series QBO30



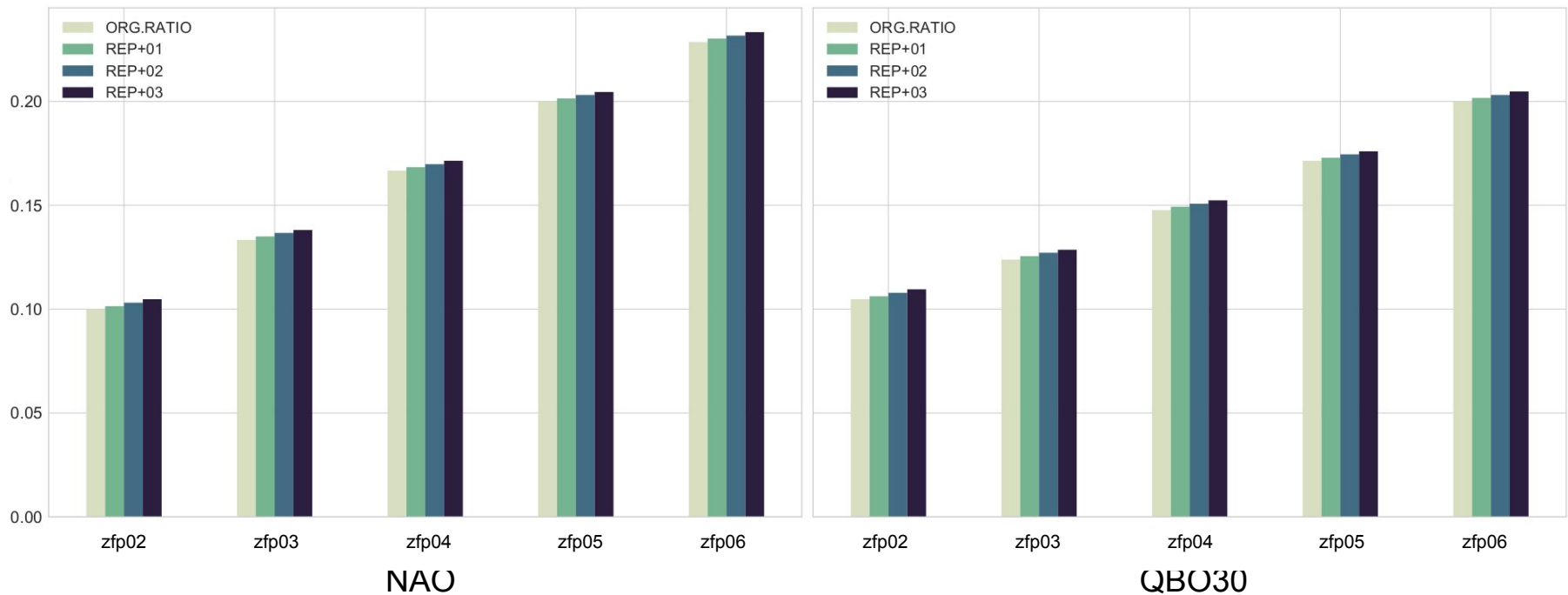


# Exp. 3: Lossy compression (w/ replacement) over time series QBO30



--- ARIMA    ..... Direct    -+- CUMCORR    -+ SPECIAL    -+ FIRST    ---- EVENLY    -+ ROLL

# 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*.

# Thank you for your attention

- Data and Code available (GPL-v3)
  - [github.com/ucyo/adaptive-lossy-compression](https://github.com/ucyo/adaptive-lossy-compression)
- Contact
  - [Cayoglu@kit.edu](mailto:Cayoglu@kit.edu)

- Thanks to Peter Braesicke,  
Tobias Kerzenmacher,  
Jörg Meyer and Achim Streit

