Adaptive Lossy Compression of Complex Environmental Indices using Seasonal Auto Regressive Integrated Moving Average Models

Ugur Cayoglu[†*], Peter Braesicke[†], Tobias Kerzenmacher[†], Jörg Meyer[*] and Achim Streit[*]
[*]Steinbuch Centre for Computing
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
Email: {Ugur.Cayoglu, Joerg.Meyer2, Achim.Streit}@kit.edu
[†]Institute of Meteorology and Climate Research - Atmospheric Trace Gases and Remote Sensing
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
Email: {Peter.Braesicke, Tobias.Kerzenmacher}@kit.edu

Abstract—**Significant increases in computational resources have enabled the development of more complex and spatially better resolved weather and climate models. As a result the amount of output generated by data assimilation systems and by weather and climate simulations is rapidly increasing e.g. due to higher spatial resolution, more realisations and higher frequency data. However, while compute performance has increased significantly because of better scaling program code and increasing number of cores the storage capacity is only increasing slowly. One way to tackle the data storage problem is data compression.**

Here, we build the groundwork for an environmental data compressor by improving compression for established weather and climate indices like El Niño Southern Oscillation (ENSO), North Atlantic Oscillation (NAO) and Quasi-Biennial Oscillation (QBO). We investigate options for compressing these indices by using a statistical method based on the Auto Regressive Integrated Moving Average (ARIMA) model. The introduced adaptive approach shows that it is possible to improve accuracy of lossily compressed data by applying an adaptive compression method which preserves selected data with higher precision. Our analysis reveals no potential for lossless compression of these indices. However, as the ARIMA model is able to capture all relevant temporal variability, lossless compression is not necessary and lossy compression is acceptable. The reconstruction based on the lossily compressed data can reproduce the chosen indices to such a high degree that statistically relevant information needed for describing climate dynamics is preserved. The performance of the (seasonal) ARIMA model was tested with daily and monthly indices.

I. Introduction

Climate models simulate, contrary to weather prediction models, decades of weather dynamics and climate influences on a global scale.

The current **E**uropean **ReA**nalysis (ERA5) dataset outputs hourly data starting from 1979 to the present on a 1440 x 721 (about 31 km) horizontal and 137 level vertical (to 0.01 hPa = 80 km) grid [1]. If we assume 16-Bit Integer values for each variable this amounts to 2.26 TiB p.a. per variable[2].

These high resolution climate models are only feasible because of sophisticated and suitable dynamical cores of the new models e.g. with solvers for fluid mechanics. The introduction of non-hydrostatic model equations accompanied by an increase of computational power lead to high resolution models [1]. Unfortunately, the storage capacities did not increase proportionally with the computational power and today it's common practice not to save the data on every timestamp simulated.

One way to solve the data storage problem is to compress the datasets by removing redundant information. Compression reduces the space needed for storing and archiving of the output data. Additionally compression enables the possibility to run simulations with higher resolution, while consuming the same storage.

Here we explore possibilities to improve compression of predicative climate indices using a statistical method based on the Auto Regressive Integrated Moving Average (ARIMA) model [2] with the aim to generalise the knowledge later for more generic data.

The ARIMA model helps identify interdependencies in the dataset. We than take advantage of the interdependencies and improve the correlation between the original and reconstructed index while using negligible more storage.

The remainder of this paper is divided into five sections: Related work is presented in Section 2. Section 3 describes our approach and explains methods and metrics. Finally in Section 4 our results are presented and discussed. In the concluding section we give a short summary and outlook for future work.

II. Related Work

The ARIMA model is being used in environmental research for forecasting individual weather observations [3]–[5]. These studies focus on very narrow timesteps (e.g. hourly) and

[1] European Centre for Medium-Range Weather Forecasts (ECMWF) Newsletter No. 147 – Spring 2016 (p.7)

[2] ERA5 supports circa 120 variables. While some of these variables are simulated, others can be deduced from simulated variables. For reference http://apps.ecmwf.int/codes/grib/param-db

concentrate on specific regional areas. Little attention is being paid for relationships of longer time periods with a larger spatial scope.

A hybrid model for forecasting water resources has been developed by Banihabib et. al [6]. This model uses an ARIMA model with exogenous inputs. After model generation the output is fed to Neural Networks (NN) for the detection of non-linear correlations. Several other studies looked at ARIMA models in connection with NN [4], [7] to improve forecasting of future observations. These studies show that ARIMA models coupled with NN can improve the forecasting ability of models. Since the focus of this paper is not on forecasting possible future values, but finding a better representation of given values, the results from the ARIMA model were satisfactory.

Note that Guenni et. al [8] focuses on the ENSO3[3] index being successfully used for the prediction of precipitation, thus supporting our idea to investigate climate indices for predicative purposes in compression.

The ARIMA model is also being used in other research areas like economics [9], [10], telecommunications and multimedia [11]–[13] and social studies [14], but so far few studies have looked at ARIMA models in connection with compression. Zordan et. al [15] evaluate among others ARIMA models in connection with lossy compression of energy-constrained wireless sensor-networks. While overhead like memory and calculation efficiency during compression is important for energy-constrained sensor-network, for the application described in this paper it is not relevant.

III. Methods

We use two different approaches to obtain compressed indices. Figure 1 illustrates both workflows. Our proposed method using an ARIMA model is depicted as “ARIMA approach”. The second approach illustrates the usual process by applying compression directly on the indices and is described as “Direct approach”.

After calculating the indices we build an ARIMA model for each index. The results from the ARIMA model will then be compressed. After this step several data points will be chosen by the replacement methods defined in Chapter III-E. These data points will then be replaced by ones with higher precision. These replacement methods use information about the model and output the final compressed indices.

The following chapter describes the steps to create the climate indices, the ARIMA model, the compression method used, metrics and the replacement methods.

A. Indices calculation

The data used in this paper was obtained from a reanalysis created by the ECHAM/MESSy[4] Atmospheric Chemistry (EMAC) [16] model. It consisted of a 128x64 (longitude, latitude) grid with six vertical levels (from 1000 hPa to 10 hPa) and spanned a time period from the beginning of 1979 till the end of 2013 with 10 h time steps. The following variables were available as single-precision floating-point values: ozone, pressure, dry air temperature and westerly wind.

The following climate indices have been created for our investigation: El Niño Southern Oscillation 3.4 (ENSO34), North Atlantic Oscillation (NAO), Quasi-Biennial Oscillation at 30 (QBO30) and 50 hPa (QBO50). These indices show high significance in climate research [8], [17]–[19] and help in numerical weather predictions and seasonal forecasting. ENSO34 is being used in forecasting rainfall, NAO in forecasting seasonal temperature for Europe while QBO is being used for predicting monsoon precipitation.

Each index was created with two temporal resolutions: monthly and daily. For the calculation of ENSO34 and QBO*x* a spatial subset of the data according to Table I was selected. Next the zonal and meridional means were calculated. The NAO index was calculated using the surface pressure

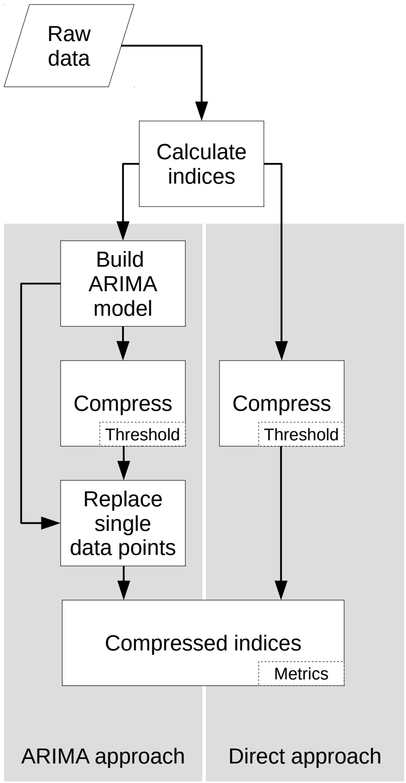

Fig. 1: Flowchart of analysis.

TABLE I: Spatial borders and variables used for calculating indices with temperature (T), pressure (p) and westerly wind (u).

Index	Variable	Lat [N]	Lon [E]	Lev [hPa]
ENSO34	T	-5 to 5	190 to 240	1000
QBO*x*	u	-5 to 5	0 to 360	*indicated by x*
NAO	p	Lisbon and Reykjavík		1000

[3] There are several ENSO*x* indices. The main difference is the spatial area being used for calculation of the index. Although there are subtle differences between each ENSO index, for the purpose of this paper these differences are irrelevant.

[4] ECMWF Hamburg (ECHAM)/Modular Earth Submodel System (MESSy)

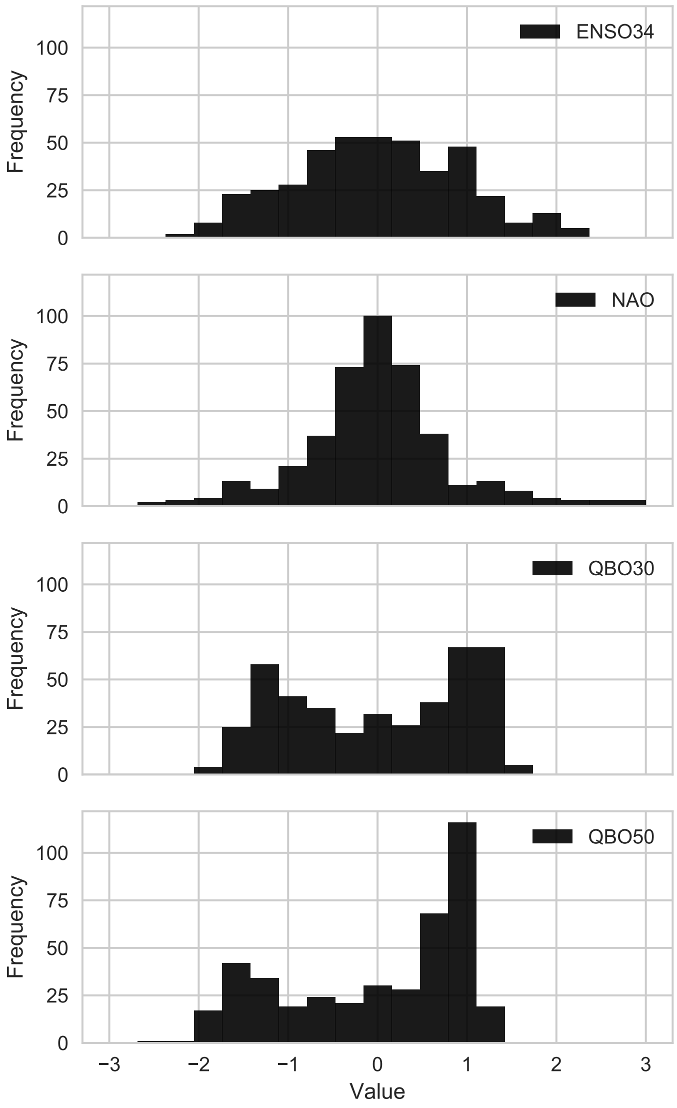

Fig. 2: Histogram of each weather index with monthly resolution. The indices are dimensionless.

difference between Lisbon and Reykjavík. Afterwards yearly monthly mean and multi-year monthly mean for all indices were calculated. The multi-year monthly mean was then subtracted from the corresponding yearly monthly mean and divided by the multi-year standard deviation for each month. This concluded the process for the monthly indices. For the daily indices these steps were repeated with respective daily resolution.

This concluded the preprocessing of the data. A histogram of each index is depicted in Figure 2 and a summary of their characteristics are given in Table II.

B. Model

The Auto Regressive Integrated Moving Average (ARIMA) model was first introduced by Box and Jenkins [2] and has been extended several times [20]. The ARIMA model tries to find interdependencies in the dataset. It uses the premise that a datum in a time series is dependent on its previous values and can be expressed by a function of its former values.

TABLE II: Information about the monthly indices.

	ENSO34	NAO	QBO30	QBO50
count	420	420	420	420
mean	0.000	0.000	0.000	0.000
std	0.936	0.822	0.997	0.995
min	-2.208	-2.626	-1.876	-2.464
25%	-0.656	-0.421	-1.011	-0.934
50%	0.001	-0.016	0.097	0.364
75%	0.693	0.397	0.999	0.869
max	2.273	3.097	1.472	1.383
skew	0.028	0.393	-0.161	-0.565
kurt	-0.481	2.055	-1.474	-1.149

Because of seasonal dependence in weather dynamics we used a seasonal ARIMA model [20] for monthly and the original ARIMA model [2] for daily datasets.

The seasonal ARIMA model is being described by the following notation:

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

with (p,d,q) representing the non-seasonal auto-regressive (p), difference (d) and moving-average (q) order and (P,D,Q) the equivalent seasonal order with period length s.

The general equation for seasonal ARIMA is as following:

$$\Phi(B^s)\phi(B)(x_t-\mu)=\Theta(B^s)\theta(B)\varepsilon_t \tag{1}$$

with x_t representing the target value at time t, μ the expected mean value of the data, ε_t the error term of the model, B^k the backpropagation with $B^k x_t = x_{t-k}$ and following components:

$$\begin{aligned}
\text{Seasonal AR}: \Phi(B^s) &= 1-\Phi_1 B^s - \cdots - \Phi_P B^{P\cdot s}\\
\text{AR}: \phi(B) &= 1-\phi_1 B - \cdots - \phi_p B^p\\
\text{Seasonal MA}: \Theta(B^s) &= 1+\Theta_1 B^s + \cdots + \Theta_Q B^{Q\cdot s}\\
\text{MA}: \theta(B) &= 1+\theta_1 B + \cdots + \theta_q B^q
\end{aligned}$$

with i representing the timestep before the target value, $\Phi(B^s)$ the seasonal auto-regressive (AR) parameter, $\phi(B)$ the AR parameter, Φ_i the seasonal AR coefficients, ϕ_i the AR coefficients, $\Theta(B^s)$ the seasonal moving-average (MA) parameter, $\theta(B)$ the MA parameter, Θ_i the seasonal MA coefficients and θ_i the MA coefficients of the model.

For information about the data and choosing the range of AR and MA parameters the (partial) Auto-Correlation Function (ACF) defined as following was used:

$$ACF=\frac{\sum_{t=k+1}^{n}(Y_t-\overline{Y})(Y_{t-k}-\overline{Y})}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2} \tag{2}$$

with $k \in \mathbb{N}$ representing the temporal lag, Y_t time series with start at time t and $\overline{Y}$ the mean value of the time series. The coefficients Φ_i, ϕ_i, Θ_i and θ_i were optimised using the Akaike's Information Criterion (AIC) [21].

C. Compression

For compression we used the zfp compression method introduced in [22]. It has already been applied successfully on climate data [23] and supports lossy as well as lossless data compression. We will use the following notation throughout the paper: *zfpPR*. Here *PR* denotes the precision of the applied compression. In case of single-precision floating-point numbers (32 bits) a lossless compression would be denoted as *zfp32*.

D. Metrics

For evaluating the forecasting models the Root Mean Square Deviation (RMSD) was used. The reconstructed index from the lossy compression was evaluated using the Pearson Correlation coefficient [24]:

$$r_{s,e} = \frac{\sum_{i=s}^{e}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=s}^{e}(x_i - \overline{x})^2}\sqrt{\sum_{i=s}^{e}(y_i - \overline{y})^2}} \tag{3}$$

with x_i representing the original value, $\overline{x}$ original mean value, y_i the reconstructed value, $\overline{y}$ reconstructed mean value, start (s) and end (e) indices of the observed time series. The reason for choosing the Pearson correlation coefficient as a metric was that most of the time the correlation between the index and other weather phenomena is being analysed. Therefore it is of utmost importance to reconstruct an index correlated to the original index.

The compression rate was measured using the so called *bits per float* (bpf) metric (Eq. 4). This metric represents the average number of bits needed to save a floating-point number.

$$\text{bpf} = \frac{\text{Bitsize of file}}{\text{Number of float values represented}} \tag{4}$$

Another metric being used was the compression ratio (cr):

$$\text{cr} = \frac{\text{Size of file after compression}}{\text{Size of file before compression}} \tag{5}$$

A ratio closer to zero would suggest ideal compression and close to one a bad compression.

The introduced ARIMA approach improves compression by replacing several data points by ones with higher precision. Those points are chosen by the replacement methods described in the following chapter.

E. Replacement methods

Let $x^b = \{x_1^b, x_2^b, \dots, x_n^b\}$ be a lossily compressed time series with b representing the bits preserved from the original time series. A lossless compression for single-precision floating-point numbers would be depicted as x^{32} while the most lossy compression would be x^1. Further, let $k \in \mathbb{N}$ be the number of data points we are going to replace, let $l \in \mathbb{N}$ be the number of additional precision bits we want to save and blocksize $bs = \max\{p, q\}$ represent either the auto-regressive or moving-average order of the ARIMA model. The parameter bs helps identify the data contributing to the calculation of a datum x_i^b. The updated time series will be represented by $\hat{x} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$. Further on let $sort(X)$ be the sorted set of X, $argsort(X)$ be the arguments of the sorted set of X and $S(t, X)$ the t previous values of each element of X:

$$sort(X) = \Big\{ x_i \mid x_i \leq x_{i+1} \land x_i \in X \Big\} \tag{6}$$

$$argsort(X) = \Big\{ \arg x_i \mid x_i \leq x_{i+1} \land x_i \in X \Big\} \tag{7}$$

$$S(t, X) = \Big\{ x - j \mid j \leq t \land j \in \mathbb{N} \land x \in X \Big\} \tag{8}$$

The algorithm differentiates between the following methods to choose the data points being replaced:

1) First: The first k values will be replaced.

$$\hat{x}_i = \begin{cases} x_i^{b+l} & \text{if } i \leq k \\ x_i^b & \text{else} \end{cases} \tag{9}$$

2) Even: The k values being replaced are evenly distributed over the whole time series. The time series is split in $bl = \left\lfloor \frac{k}{bs} \right\rfloor + 1$ evenly distributed blocks with size $\left\lfloor \frac{n}{bl} \right\rfloor$ and midpoints $M = \left\{ j \cdot \left\lfloor \frac{n}{bl} \right\rfloor \mid j \in \mathbb{N} \land j \leq bs \right\}$

$$\hat{x}_i = \begin{cases} x_i^{b+l} & \text{if } i \in [m - \left\lfloor \frac{bs}{2} \right\rfloor, \dots, m + \left\lfloor \frac{bs}{2} \right\rfloor) \text{ with } m \in M \\ x_i^b & \text{else} \end{cases} \tag{10}$$

3) Special: The cumulative correlation (Eq. 11) of the time series will be calculated, the results sorted and those data replaced, which contribute to the calculation of the data with the lowest correlation.

$$C = \Big\{ r_{1,j} \mid j \in \mathbb{N} \land j \leq n \Big\} \tag{11}$$

$$C' = argsort(C)$$

$$\hat{x}_i = \begin{cases} x_i^{b+l} & \text{if } \arg i \leq k \text{ with } i \in S(bs, C') \\ x_i^b & \text{else} \end{cases} \tag{12}$$

4) Rolling: The rolling correlation (Eq. 13) with window size bs will be calculated, the results sorted and those data replaced which contribute to the calculation of the data with the lowest correlation.

$$R = \Big\{ r_{j-bs,j} \mid j \in \mathbb{N} \land bs < j \leq n \Big\} \tag{13}$$

$$R' = argsort(R)$$

$$\hat{x}_i = \begin{cases} x_i^{b+l} & \text{if } \arg i \leq k \text{ with } i \in S(bs, R') \\ x_i^b & \text{else} \end{cases} \tag{14}$$

5) Cumcorr: The cumulative correlation of the time series will be calculated (Eq. 11) and the datum identified which is followed by the biggest consecutive drop in correlation. The data responsible for the calculation of this datum will then be replaced. Afterwards the process will be repeated until k data points have been replaced.

TABLE III: Results of (seasonal) ARIMA model run for monthly and daily data.

Index	Model	AIC	RMSD
Monthly			
ENSO34	$ARIMA(3,0,2)(1,0,0)_{12}$	290.164	5.067e−8
NAO	$ARIMA(1,0,0)(1,0,0)_{12}$	1020.352	8.195e−9
QBO30	$ARIMA(2,0,3)(1,0,0)_{12}$	-456.730	1.0877e−7
QBO50	$ARIMA(1,1,1)(1,0,1)_{12}$	-164.427	2.909e−6
Daily			
ENSO34	$ARIMA(5,2,4)(0,0,0)_{0}$	-10245.442	4.686e−4
NAO	$ARIMA(2,0,2)(0,0,0)_{0}$	31267.670	1.440e−7
QBO30	$ARIMA(5,0,4)(0,0,0)_{0}$	-54091.415	1.084e−7
QBO50	$ARIMA(5,0,4)(0,0,0)_{0}$	-52112.790	4.488e−8

$$C = \left\{ r_{1,j} \mid j \in \mathbb{N} \land j \leq n \right\} \tag{15}$$

$$C' = \begin{cases} c_i & \text{if } c_{i+1} \geq 0 \\ \displaystyle\sum_{j=0}^{b} c_{i+j} & \text{else with } c_{i+j} < 0 \land b \in \mathbb{N} \end{cases} \tag{16}$$

$$C'' = argsort(C')$$

$$\hat{x}_i = \begin{cases} x_i^{b+l} & \text{if } \arg i \leq k \text{ with } i \in S(bs, C'') \\ x_i^{b} & \text{else} \end{cases} \tag{17}$$

F. Experiments

Several tests were carried out to investigate possible compression methods. First we focused on lossless compression. Since the datasets were single-precision floating-point numbers we used *zfp32* for compression.

Further we analysed a lossy compression with the goal to achieve a deviation as small as possible for a given error bound. For this experiment we choose the error bound $\tau = 1\text{e}{-5}$ so that $r_{1,n} \geq\ 1.0 - \tau$ with $r_{1,n}$ representing the Pearson Correlation coefficient (details in following section).

A third experiment was conducted to see what effect a gradual decline in precision from *zfp32* to *zfp01* has on the correlation coefficient and if replacing several data points with a higher precision would improve the correlation coefficient. These indices with updated data will be described by the following notation: *zfpPR+l* with l representing the number of additional precision bits. The notation *zfp06+02* depicts a lossy compression method with six precision bits where several data points have additional two bits of precision. For the following experiments we replaced five and ten percent of the data with $l \in \{1, 2, 3\}$.

In the following chapter we will evaluate and discuss our findings and applied methods.

IV. Evaluation

A. Model

The (seasonal) ARIMA model can reconstruct all indices with good accuracy. The RMSD of the reconstructed index for monthly data is better than the one for the daily dataset. The ARIMA models with differentiation step, QBO50 for monthly data and ENSO34 for daily data, perform worst in their respective group. Detailed results are described in Table III.

TABLE IV: Results of the DF-Test for stationariness.

	ENSO34	NAO	QBO30	QBO50
DFT Test Statistic	-5.341	-17.571	-7.447	-9.257
Critical Value (1%)	-3.446	-3.446	-3.447	-3.447
Critical Value (5%)	-2.869	-2.868	-2.869	-2.869
Critical Value (10%)	-2.571	-2.570	-2.571	-2.571

TABLE V: Results for lossless compression of daily (dm) and monthly (mm) datasets for the residuals of the ARIMA model and directly on the dataset. Header files are excluded.

	ARIMA	Direct	ARIMA	Direct
Index	mm	mm	dm	dm
ENSO34	33.371	32.762	33.072	32.300
NAO	33.371	33.067	33.071	32.821
QBO30	33.219	32.152	33.031	30.753
QBO50	33.451	32.457	33.051	31.009

The Pearson correlation coefficient $r_{1,n}$ for all indices is $1.0 \pm 2\text{e}{-12}$. Figure 3 illustrates the ARIMA model for NAO and QBO30. It can be seen, that the reconstructed index defined by the ARIMA model represents the original index very well.

Since ARIMA models can only be applied to stationary data we conducted the Dickey-Fuller-Test (DF-Test) [25] to test for stationariness. All indices are stationary with a confidence level of 99 %. The results of the DF-Test are represented in Table IV.

B. Compression

In this section we will compare the ARIMA approach without replacements with the direct approach. This first comparison builds the groundwork for further comparisons. Afterwards in section IV-C we will compare the results of the replacement methods with the original ARIMA results and the direct approach.

1) Lossless: Our results show that lossless compression of the ARIMA output is resulting into bigger files than without compression. A lossless compression applied directly on the indices returns similar results. The only exception being the QBO30 and QBO50 indices with a daily resolution. The filesize of the QBO30 and QBO50 daily dataset is slightly decreasing by four percent for QBO30 and three percent for QBO50. Detailed results are presented in Table V.

2) Strict lossy compression: A lossy compression with $\tau = 1\text{e}{-5}$ achieves in most cases a compression ratio of ~.4. The ARIMA approach (monthly and daily) achieved an average compression ratio of 0.381. The only exceptions were QBO50 (monthly) and ENSO34 (daily) which were only compressed with a ratio of 0.663 until the boundary condition τ was met. Detailed results are presented in Table VI.

This deviation is due to the differentiation step during model building. This additional calculation step increases error

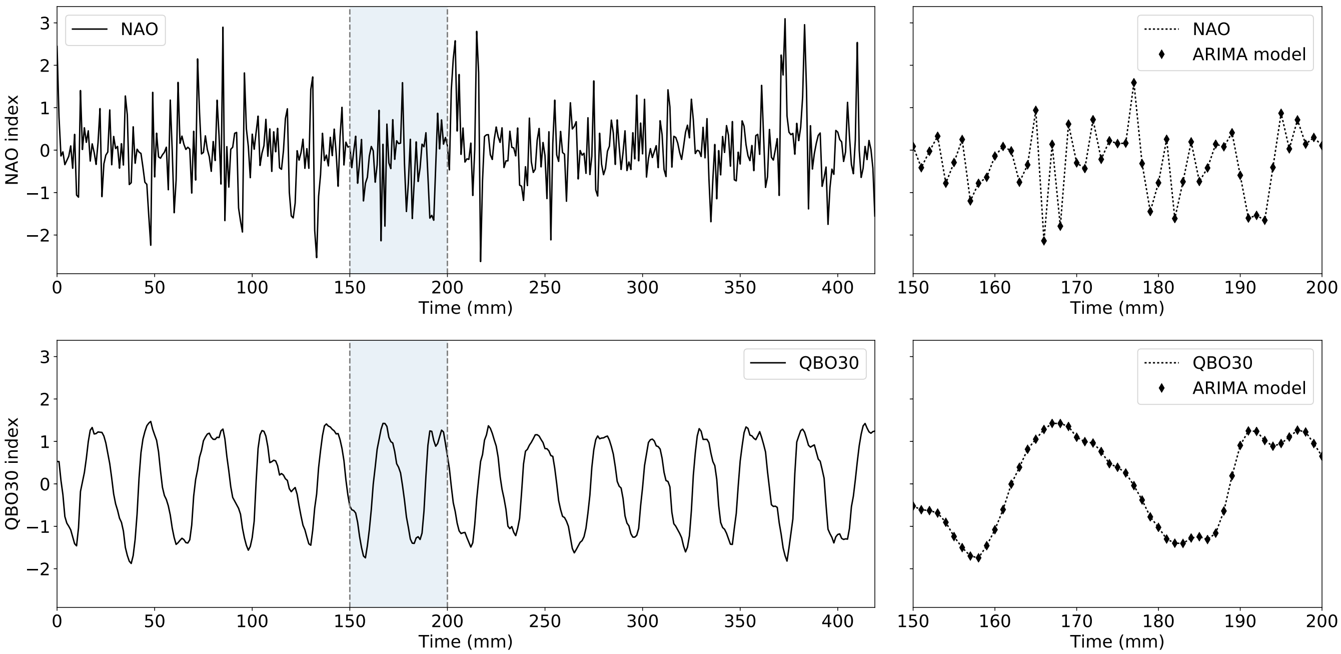

Fig. 3: This Figure illustrates the NAO (top) and QBO30 (bottom) indices and their respective reconstruction via the ARIMA model. Error bars have been omitted for plotting purposes.

TABLE VI: Ratios for lossy compression of daily (dm) and monthly (mm) datasets with $\tau = 1\text{e}{-5}$ as error threshold. Header files are excluded.

Index	ARIMA mm	Direct mm	ARIMA dm	Direct dm
ENSO34	0.386	0.371	0.658	0.322
NAO	0.386	0.386	0.377	0.370
QBO30	0.381	0.357	0.376	0.273
QBO50	0.668	0.362	0.377	0.281

propagation and results in additional precision bits needed to meet the threshold τ.

It is no surprise that the direct approach has improved the cr of these two indices. The compression ratio of QBO50 (monthly) went down from 0.668 to 0.362 and ENSO34 (daily) from 0.658 to 0.322. While applying lossy compression directly on the monthly indices did not improve the NAO index, the compression ratio for QBO30 and ENSO34 slightly improved by three respectively one percent.

The daily datasets show better results on average (from 0.37 to 0.30). Only the NAO (daily) index did not show such a decrease in compression ratio and gained no percentage points.

3) Lossy compression with gradual decline: The third and last experiment was conducted to analyse the effects of a more and more aggressive lossy compression on the Pearson correlation coefficient (Eq. 3). The precision level of the lossy compression algorithm was gradually reduced from *zfp32* to *zfp01*.

Applying lossy compression on daily and monthly data directly showed that the NAO index was performing worst regarding both metrics $r_{1,n}$ and cr. The results of the ARIMA approach was similar to the strict lossy compression (described in IV-B2). The indices NAO (monthly) and ENSO34 (daily) performed worst in every step regarding cr. It looks like the difference step of the ARIMA models (Table III) has a big effect on the correlation levels.

The results we gathered until now suggests that additional calculation steps needed for generating the ARIMA model has a negative effect on the compressed indices. This effect was expected because of the interdependence and hence error propagation in the ARIMA model. The accuracy of each datum depends on all calculations done until that point. The later the datum is in the time series, the greater is the effect of calculation errors. This effect is even more significant if an ARIMA model with a differentiation step is being used. This can be seen in the cr of QBO50 in the monthly dataset and ENSO34 in the daily dataset (see Table VI). Otherwise, the ARIMA approach (without replacement) is causing a 1-3 % loss in storage space for the monthly indices and ten percent for the daily indices.

Most interesting are the results for the NAO index. While the other indices show similar behaviour in gain and loss of cr with both approaches, the NAO index does not. The direct and ARIMA approach have no effect on the cr of the monthly indices and only negligible effect on the daily indices with 0.007 difference in compression ratio. A closer look at the index (Tab II and Fig. 2) reveals properties which may be reasons for the difficulties in compression. The standard deviation of the NAO index is the lowest with

0.822 and the first and third quartile are the closest to the mean with -0.421 and 0.397. Additionally the NAO index has several outliers. The minimum and maximum have the highest absolute distance to mean of all indices. This properties are supported by the unbiased skewness and kurtosis of the NAO index given in Table II. The NAO index is heavy tailed with a slight asymmetry on the right tail.

This behaviour can also be observed in the replacement methods. The ENSO34, QBO50 and QBO30 indices behave similar to each other. For reasons of brevity and because of these similarities only the replacement results for NAO and QBO30 will be presented in the next chapter.

C. Replacement methods

The results of the former chapter show that cr and $r_{1,n}$ for lossy compression with the ARIMA approach is worse than the direct approach. This is due to the interdependence of the data. In the ARIMA model a single datum x_i^b is dependent on previous data. If one or several of these data points deviate too far from its original datum, then it will negatively effect the calculation of the dependent data. The consequence of this is error propagation.

But there is the possibility to use this interdependence to our advantage. We can identify those data points which have a negative impact on the reconstruction of the index. These can then be replaced by ones with higher precision. In the following we will first compare the indices reconstructed by the different replacement methods with the original ARIMA output and afterwards with the directly compressed indices.

1) Replacement of 5% and 10% of data: Several tests were carried out to see how many data points needed to be replaced to see an effect on the correlation coefficient $r_{1,n}$. Table VII illustrates this effect for the monthly indices.

Most of the time the gain in correlation by replacing ten instead of five percent of the data is small. There were two exceptions to this: The increase in correlation from 0.468 to 0.624 with *zfp02+01* on the NAO index and an increase from 0.691 to 0.935 with *zfp04+01* on the QBO30. It should be pointed out that the correlation value of 0.935 with *zfp04+01* is almost as good as using *zfp05* for the whole index which has a correlation coefficient of 0.972.

A more striking and disappointing finding was that replacing data with higher precision did not always increase the correlation coefficient. The NAO index showed no decline, but the correlation coefficient of QBO30 dropped in several cases. While most of the time the drops where $< .01$, the most significant drop was from 0.139 to 0.027 with *zfp02+01*. Further research is needed to analyse why these drops occurred in the lowest precision level. Figure 4 shows the correlation coefficient of each replacement method from *zfp02+01* to *zfp06+03*.

In the next section the replacement methods will be compared with each other.

2) Replacement methods: Figure 5 illustrates the correlation coefficient $r_{1,t}$ at month t for *zfp06+03* and ten percent replacement. The NAO index is best represented by the special method described in III-E3. The reason for this seems to be twofold.

TABLE VII: Pearson correlation coefficient by replacing five and ten percent of the monthly indices. The replacement method being used is "Special" (see section III-E3 for details).

	zfp02	zfp03	zfp04	zfp05	zfp06
NAO (5%)					
$l = 0$	0.354	0.725	0.924	0.979	0.994
$l = 1$	0.468	0.825	0.950	0.983	0.996
$l = 2$	0.506	0.826	0.952	0.984	0.996
$l = 3$	0.519	0.831	0.953	0.984	0.996
NAO (10%)					
$l = 0$	0.354	0.725	0.924	0.979	0.994
$l = 1$	0.624	0.864	0.959	0.987	0.997
$l = 2$	0.692	0.870	0.964	0.989	0.997
$l = 3$	0.705	0.878	0.965	0.989	0.997
QBO30 (5%)					
$l = 0$	0.139	0.482	0.635	0.972	0.986
$l = 1$	0.027	0.566	0.691	0.967	0.996
$l = 2$	0.039	0.591	0.692	0.973	0.993
$l = 3$	0.042	0.596	0.677	0.985	0.995
QBO30 (10%)					
$l = 0$	0.139	0.482	0.635	0.972	0.986
$l = 1$	0.050	0.575	0.935	0.968	0.996
$l = 2$	0.082	0.615	0.940	0.973	0.993
$l = 3$	0.084	0.607	0.944	0.987	0.996

First, the special method decides which datum to replace depending on the lowest correlation coefficient. The correlation coefficients at each timestep are being sorted and those data replaced, which contribute to the calculation of the lowest correlation coefficient. With this property the special method can compensate for sudden changes in the index. Especially the first drop at the beginning of the NAO index and at $t = 50$ is not having as big of an impact on the correlation coefficient with the replacements defined by the special method. The closeup on the right of Figure 5 illustrates this well.

Second, the model being used for the NAO index is $ARIMA(1,0,0)(1,0,0)_{12}$. Every single datum is only depending on its immediate predecessor and the one from last year. A single datum is only depending on two previous values. This small dependence helps correcting more data points and error propagation has not as much of an impact.

The ARIMA approach improves the reconstruction of the NAO index. The reconstruction has on each time step t a better correlation coefficient $r_{1,t}$ than the direct approach with using only negligible more storage (see Table IX).

For the QBO30 index the rolling method described in III-E4 has the highest correlation coefficient (see Figure 5). The rolling method calculates the rolling correlation coefficient with window size $bs = max\{p, q\}$ where p describes the autoregressive and q the moving-average of the ARIMA model. The coefficients will then be sorted and those data replaced which contribute to the calculation of the data with the lowest correlation.

Unfortunately, in the case of the QBO30 index the ARIMA approach is not consistently better. In the beginning of the time series with $t < 50$ it performs significantly better. The

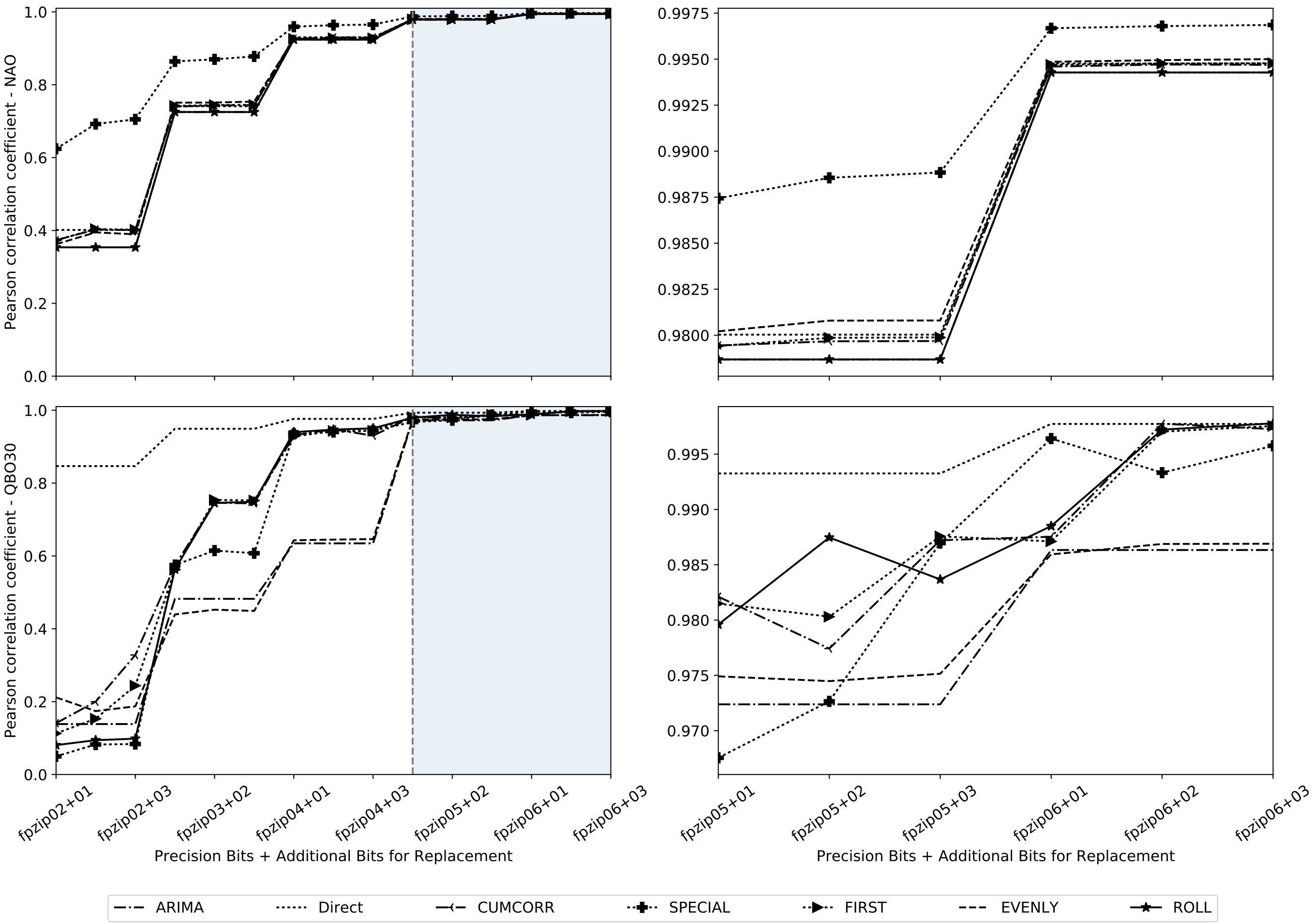

Fig. 4: Pearson correlation coefficient based on replacement method for ten percent replacement. Left: Replacement methods from *zfp02+01* to *zfp06+03*. Right: A zoom on the replacement methods with $r_{1,n} > .97$.

direct approach drops to 0.991 while the ARIMA approach stays constantly above 0.999. After this both methods behave similar.

For the daily indices the results are different. The increased number of calculation steps and error propagation has a more severe impact. While the correlation coefficient for the direct approach is at 0.999 for the QBO30 and 0.997 for the NAO index, the best ARIMA approach can achieve 0.994 for QBO30 and 0.996 for the NAO index. Table VIII show the results for *zfp06+03* on daily and monthly data.

3) Effects on storage space: Until now we analysed the impact of the replacement methods on the Pearson correlation coefficient (see Eq. 3). The additional precision bits used by the replacement methods have a negative impact on the compression ratio. The cr after using the replacement methods is depicted in Table IX where l is the number of additional precision bits.

The introduced replacement methods were conceptualised to use only a certain amount of additional storage space. They were designed to use only l additional precision bits for k data points of the indices (see Section III-E for details).

TABLE VIII: Correlation coefficient for *zfp06+03* for daily and monthly data.

	NAO monthly	QBO30 monthly	NAO daily	QBO30 daily
CUMCORR	0.99469	0.99726	0.99598	0.98400
EVENLY	0.99500	0.98690	0.99611	0.98814
FIRST	0.99478	0.99755	0.99600	0.99404
ROLL	0.99428	0.99779	0.99608	0.99409
SPECIAL	0.99686	0.99575	0.99598	0.98899
ARIMA	0.99428	0.98633	0.99583	0.98825
Direct	0.99476	0.99774	0.99669	0.99938

This design decision allowed an upper limit on how much additional storage space was being used by the method. This precaution reflects in Table IX. In the worst case we need one percent more storage space. This occurred when using *zfp06+03* and replacing ten percent of the data. The compression ratio increased from 0.200 to 0.210.

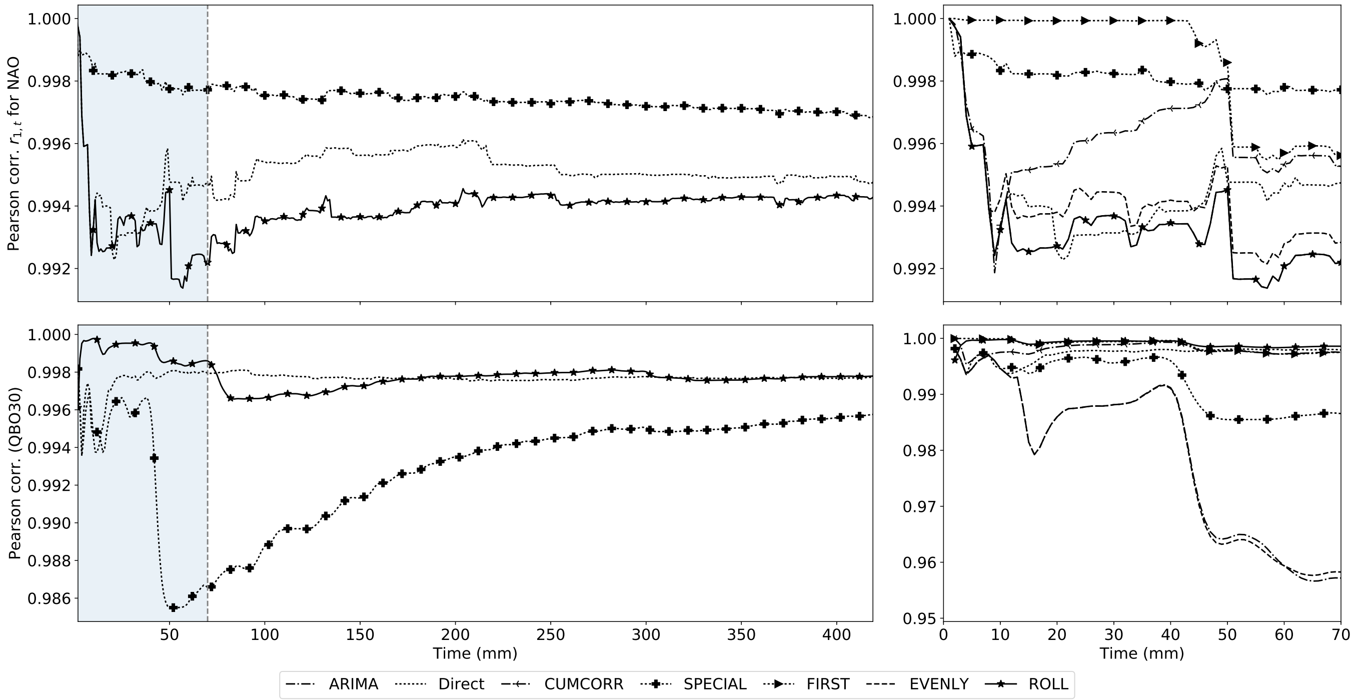

Fig. 5: Pearson correlation coefficient $r_{1,t}$ for NAO (top) and QBO30 (bottom) with *zfp06+03* lossy compression on ten percent of the data. The left part depicts the whole index, while the right part only the first 70 months. For illustration purposes the left figures only show the direct approach and ARIMA approach improved by the rolling and special replacement methods. The right figure illustrates all replacement methods.

TABLE IX: Compression ratio cr for NAO and QBO30 index after invocation of the ARIMA approach and replacing ten percent of the data. Header files are excluded.

	zfp02	zfp03	zfp04	zfp05	zfp06
NAO (daily)					
$l = 0$	0.094	0.120	0.150	0.182	0.213
$l = 1$	0.097	0.123	0.153	0.185	0.217
$l = 2$	0.100	0.126	0.156	0.188	0.220
$l = 3$	0.104	0.129	0.159	0.191	0.223
NAO (monthly)					
$l = 0$	0.100	0.133	0.167	0.200	0.229
$l = 1$	0.103	0.137	0.170	0.203	0.232
$l = 2$	0.106	0.140	0.173	0.206	0.235
$l = 3$	0.110	0.143	0.176	0.209	0.238
QBO30 (daily)					
$l = 0$	0.099	0.117	0.134	0.151	0.169
$l = 1$	0.103	0.121	0.137	0.154	0.172
$l = 2$	0.106	0.124	0.140	0.157	0.176
$l = 3$	0.109	0.127	0.143	0.160	0.179
QBO30 (monthly)					
$l = 0$	0.105	0.124	0.148	0.171	0.200
$l = 1$	0.108	0.127	0.151	0.174	0.203
$l = 2$	0.111	0.130	0.154	0.178	0.206
$l = 3$	0.114	0.133	0.157	0.180	0.210

V. Summary and Outlook

We investigate the efficiency of compression algorithms for environmental data. We have developed a test framework for the compression of climate indices based on a statistical method known as the Auto Regressive Integrated Moving Average (ARIMA) model. The indices examined are the El Niño Southern Oscillation (ENSO), North Atlantic Oscillation (NAO) and Quasi-Biennial Oscillation (QBO). Each index describes a different aspect of large-scale atmospheric dynamics and shows different variance.

To improve the lossily compressed indices we have introduced an adaptive approach. This approach shows that it is possible to improve accuracy of the reconstructed data by replacing several data points with slightly higher precision. The improved reconstruction based on lossy compressed data can reproduce the chosen indices to such a high degree that statistically relevant information needed for describing climate dynamics is preserved. The compressed indices have the same diagnostic performance than the original indices.

The study showed that ARIMA models using a differentiation step have difficulties and performed worse than other models. Our findings indicate that time series which can be expressed with small auto-regressive and moving-average order can be improved significantly.

Further analysis will focus on the aspect why certain time series like the QBO30 do not show the same improvement in reconstruction like the NAO index.

The same way ENSO indices are used to predict and diagnose climate dynamics ([8], [26]), these reconstructed indices will be used to improve the compression of environmental data.

Code and Data availability

The data of the environmental indices and an implementation of the replacement methods described above will be made available under GNU GPLv3 license at https://github.com/ucyo/adaptive-lossy-compression.

References

[1] G. Zängl, D. Reinert, P. Rípodas, and M. Baldauf, "The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core," *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 687, pp. 563–579, jan 2015. [Online]. Available: http://doi.wiley.com/10.1002/qj.2378

[2] G. E. Box and G. M. Jenkins, "Time series analysis, control, and forecasting," *San Francisco, CA: Holden Day*, vol. 3226, no. 3228, p. 10, 1976.

[3] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, "Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir," *Journal of Hydrology*, vol. 476, pp. 433–441, jan 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S002216941200981X

[4] H. Hippert, C. Pedreira, and R. Souza, "Combining neural networks and ARIMA models for hourly temperature forecast," *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000 (IJCNN 2000) , Vol. IV*, vol. 1, no. 2, pp. 414–419, 2000.

[5] J. Palomares-Salas, J. G. de la Rosa, J. Ramiro, J. Melgar, A. Aguera, and A. Moreno, "ARIMA vs. Neural networks for wind speed forecasting," in *2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. IEEE, may 2009, pp. 129–133. [Online]. Available: http://ieeexplore.ieee.org/document/5069932/

[6] M. E. Banihabib, A. Ahmadian, and F. S. Jamali, "Hybrid DARIMA-NARX model for forecasting long-term daily inflow to Dez reservoir using the North Atlantic Oscillation (NAO) and rainfall data," *GeoResJ*, vol. 13, pp. 9–16, jun 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S2214242816300584

[7] G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, jan 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0925231201007020

[8] L. B. de Guenni, M. García, Á. G. Muñoz, J. L. Santos, A. Cedeño, C. Perugachi, and J. Castillo, "Predicting monthly precipitation along coastal Ecuador: ENSO and transfer function models," *Theoretical and Applied Climatology*, pp. 1–15, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00704-016-1828-4

[9] K. R. French, G. W. Schwert, and R. F. Stambaugh, "Expected stock returns and volatility," *Journal of financial Economics*, vol. 19, no. 1, pp. 3–29, 1987.

[10] P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[11] A. K. Al Tamimi, R. Jain, and C. So-In, "SAM: A simplified seasonal ARIMA model for mobile video over wireless broadband networks," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. IEEE, 2008, pp. 178–183.

[12] L. De la Cruz, E. Pallarès, J. J. Alins, and J. Mata, "Self-similar traffic generation using a fractional ARIMA model. Application to the VBR MPEG video traffic," in *Telecommunications Symposium, 1998. ITS'98 Proceedings. SBT/IEEE International*. IEEE, 1998, pp. 102–107.

[13] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *ACM SIGCOMM computer communication review*, vol. 24, no. 4. ACM, 1994, pp. 269–280.

[14] P. Chen, H. Yuan, and X. Shu, "Forecasting Crime Using the ARIMA Model," in *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5. IEEE, oct 2008, pp. 627–630. [Online]. Available: http://ieeexplore.ieee.org/document/4666600/

[15] D. Zordan, B. Martinez, I. Vilajosana, and M. Rossi, "On the performance of lossy compression schemes for energy constrained sensor networking," *ACM Transactions on Sensor Networks (TOSN)*, vol. 11, no. 1, p. 15, 2014.

[16] P. Jöckel, H. Tost, A. Pozzer, C. Brühl, J. Buchholz, L. Ganzeveld, P. Hoor, A. Kerkweg, M. G. Lawrence, R. Sander, B. Steil, G. Stiller, M. Tanarhte, D. Taraborrelli, J. van Aardenne, and J. Lelieveld, "The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere," *Atmospheric Chemistry and Physics*, vol. 6, no. 12, pp. 5067–5104, 2006. [Online]. Available: http://www.atmos-chem-phys.net/6/5067/2006/

[17] P. J. Nowack, P. Braesicke, N. Luke Abraham, and J. A. Pyle, "On the role of ozone feedback in the ENSO amplitude response under global warming," *Geophysical Research Letters*, 2017.

[18] J. W. Hurrell and H. Van Loon, "Decadal variations in climate associated with the North Atlantic Oscillation," in *Climatic change at high elevation sites*. Springer, 1997, pp. 69–94.

[19] M. M. Hurwitz, P. Braesicke, and J. A. Pyle, "Sensitivity of the mid-winter Arctic stratosphere to QBO width in a simplified chemistry–climate model," *Atmospheric Science Letters*, vol. 12, no. 3, pp. 268–272, 2011.

[20] J. D. Cryer and K.-S. Chan, *Time Series Analysis*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2008, no. January.

[21] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[22] P. Lindstrom, "Fixed-Rate Compressed Floating-Point Arrays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2674–2683, dec 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6876024/

[23] A. H. Baker, D. M. Hammerling, S. A. Mickleson, H. Xu, M. B. Stolpe, P. Naveau, B. Sanderson, I. Ebert-Uphoff, S. Samarasinghe, F. De Simone, F. Carbone, C. N. Gencarelli, J. M. Dennis, J. E. Kay, and P. Lindstrom, "Evaluating Lossy Data Compression on Climate Simulation Data within a Large Ensemble," *Geoscientific Model Development Discussions*, no. July, pp. 1–38, jul 2016. [Online]. Available: http://www.geosci-model-dev-discuss.net/gmd-2016-146/

[24] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896. [Online]. Available: http://www.jstor.org/stable/90707

[25] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[26] P. Braesicke, O. Morgenstern, and J. Pyle, "Might dimming the sun change atmospheric ENSO teleconnections as we know them?" *Atmospheric Science Letters*, vol. 12, no. 2, pp. 184–188, 2011. [Online]. Available: http://dx.doi.org/10.1002/asl.294