Visual Exploration of Climate Variability Changes Using Wavelet Analysis

Heike Jänicke, Michael Böttinger, Uwe Mikolajewicz, and Gerik Scheuermann, *Member, IEEE*

Abstract—Due to its nonlinear nature, the climate system shows quite high natural variability on different time scales, including multiyear oscillations such as the El Niño Southern Oscillation phenomenon. Beside a shift of the mean states and of extreme values of climate variables, climate change may also change the frequency or the spatial patterns of these natural climate variations. Wavelet analysis is a well established tool to investigate variability in the frequency domain. However, due to the size and complexity of the analysis results, only few time series are commonly analyzed concurrently. In this paper we will explore different techniques to visually assist the user in the analysis of variability and variability changes to allow for a holistic analysis of a global climate model data set consisting of several variables and extending over 250 years. Our new framework and data from the IPCC AR4 simulations with the coupled climate model ECHAM5/MPI-OM are used to explore the temporal evolution of El Niño due to climate change.

Index Terms—Wavelet analysis, multivariate data, time-dependent data, climate variability change visualization, El Niño.

1 Introduction

The El Niño phenomenon is a very strong natural climate fluctuation, characterized by a positive anomaly of the sea surface temperature (SST) in the eastern tropical Pacific Ocean, which occurs in the wintertime at irregular intervals with a mean frequency of 4-7 years. The SST anomaly, the difference between the actual SST and the climatological SST, is correlated with strong weather anomalies around the Pacific and thus has a large impact on the ecology and economy.

Commonly, the Nino3 Index, the SST anomaly averaged over a rectangular area in the equatorial Pacific, is used [32] to determine the state of this natural climate fluctuation. El Ninos occur when the water is much warmer than normal for a sustained period of time. This phenomenon is reflected by a large positive Nino3 index. In the atmosphere, an anomalous pattern of sea level pressure imposing an east west gradient on the equator is associated with this SST anomaly. The Southern Oscillation Index (SOI) reflects the air pressure difference between Tahiti and Darwin (Australia). In the coupled atmosphere-ocean system the oscillation is called El Niño Southern Oscillation (ENSO). The gradient in surface pressure leads to anomalous zonal winds on the equatorial Pacific.

Due to the potential large impact on economy and ecology, the development of the strength and of the frequency of El Niño in a warming world is a quite important aspect of climate change. Although climate projections with different models differed in their results [33], several recent publications [15, 30] described an increase of the amplitude and of the frequency of ENSO as a potential consequence of global warming. We have developed a framework which visually assists the user in the spatio-temporal analysis of climate variability.

Climate change is a very important and widely discussed subject. However, only few visualization papers are directly concerned with climate applications. Looking at summaries on visualization of climate data [12, 18], one can see that most techniques that are commonly used are standard 2D techniques such as colormaps, heightfields, isolines and glyphs. Recent visualization methods for climate data analysis are iconic representation [17], volume rendering [24] or interaction-based frameworks [3, 8, 23]. Visualization of time-varying data in general is covered in [11]. Alternative applications of wavelets in visualization can be found for example in [6, 16]. Most of these techniques aim at a general visualization of the data and do not take domain specific knowledge into account.

One such widely known fact is that the climate system shows high natural variability on different time scales, and scientists have developed a large variety of techniques to reduce the dimensionality of the system and to find the most relevant patterns explaining the variation. The technique most widely used in climate research is empirical orthogonal functions (EOF) [21], commonly known as principle component analysis (PCA). In EOF analysis, the eigenvectors of the covariance matrix of the anomaly data (departure from the climatology) are used to reduce the time dependent 2D data to time series of patterns with greatest variance (eigenvectors with greatest eigenvalue). The eigenvalue gives a measure of the explained variance by its corresponding eigenvector. Projecting the anomaly field onto one of the eigenvectors highlights areas of large variance. EOFs are orthogonal in both space and time, which is not necessarily true for physical modes [26]. Hence, different derivatives of EOF such as rotated EOFs (REOF) and extended EOFs (EEOFs) have been developed.

An alternative approach is taken by wavelet analysis, which does not try to find global patterns of variance but analyzes the data locally and reveals active frequencies in a time series and their changes over time. An excellent introduction to wavelet analysis related to climate research is given by Torrence and Compo [31] (see also references therein). Wavelet analysis has been used to investigate global [10, 27] and local [20] climate changes or the evolution of characteristic parameters such as the Nino3 index used to define El Niño events [14, 31]. All of these studies have in common that they only analyze few time series that are either directly given or defined by averaging larger regions.

In this paper we will explore different techniques to make the wavelet analysis applicable to an entire multivariate two-dimensional climate data set with almost 3000 time steps.

2 Climate Simulation Data

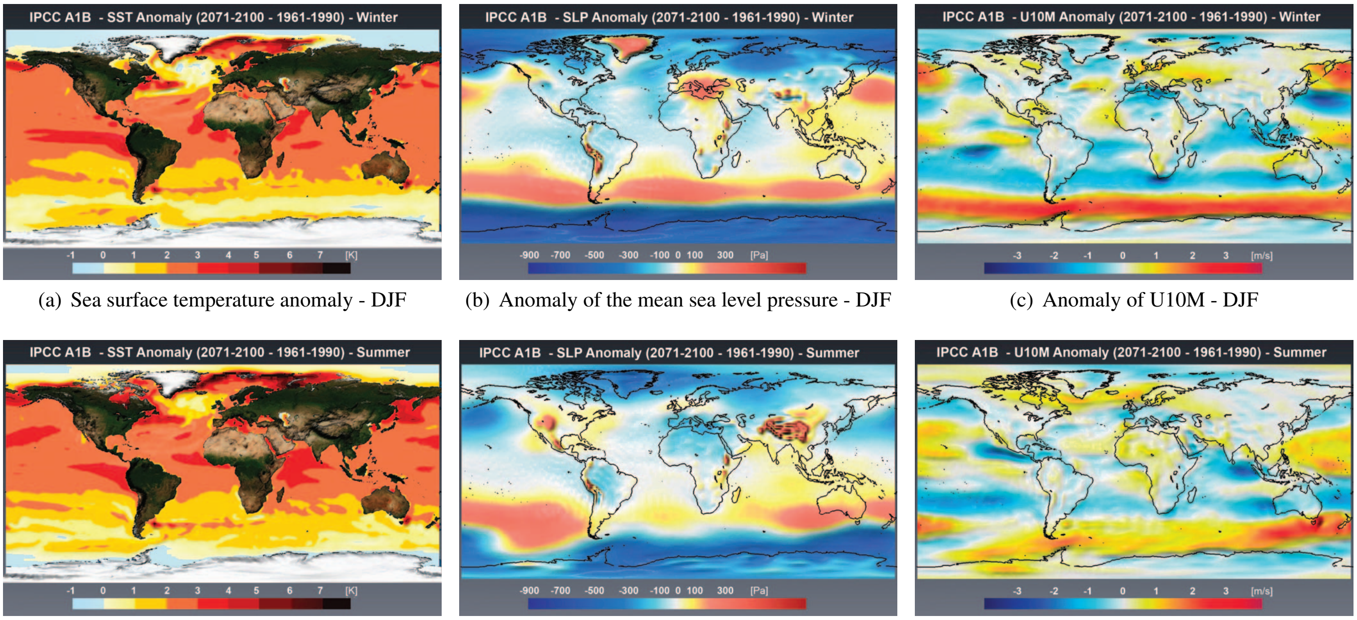
For our work we have used results of the Intergovernmental Panel on Climate Change (IPCC) AR4 [33] simulations carried out at DKRZ by the Max-Planck-Institute for Meteorology [25]. The simulation, which starts at 1860, is first forced by observed greenhouse gas concentrations until the year 2000. For the time from 2001 to 2100 we have chosen the IPCC A1B scenario results [33]. For control purposes we have also used the CTL experiment, where the greenhouse gas concentrations have been kept frozen at the preindustrial level of 1860.

The atmospheric component of the coupled atmosphere-ocean model ECHAM5/MPI-OM has a horizontal resolution of approxi-

- *Heike Jänicke and Gerik Scheuermann are with University of Leipzig, E-mail: {jaenicke,scheuermann}@informatik.uni-leipzig.de.*
- *Michael Böttinger is with German Climate Computing Center (DKRZ), E-mail: boettinger@dkrz.de.*
- *Uwe Mikolajewicz is with Max Planck Institute for Meteorology (MPI-M), E-mail: uwe.mikolajewicz@zmaw.de.*

Manuscript received 31 March 2009; accepted 27 July 2009; posted online 11 October 2009; mailed on 5 October 2009.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org .

Published by the IEEE Computer Society



IPCC A1B - SST Anomaly (2071-2100 - 1961-1990) - Winter

IPCC A1B - SLP Anomaly (2071-2100 - 1961-1990) - Winter

IPCC A1B - U10M Anomaly (2071-2100 - 1961-1990) - Winter

IPCC A1B - SST Anomaly (2071-2100 - 1961-1990) - Summer

IPCC A1B - SLP Anomaly (2071-2100 - 1961-1990) - Summer

IPCC A1B - U10M Anomaly (2071-2100 - 1961-1990) - Summer

(a) Sea surface temperature anomaly - DJF

(b) Anomaly of the mean sea level pressure - DJF

(c) Anomaly of U10M - DJF

(d) Sea surface temperature anomaly - JJA

(e) Anomaly of the mean sea level pressure - JJA

(f) Anomaly of U10M - JJA

Fig. 1. Mean northern hemisphere summer (**J**une**J**uly**A**ugust) and winter (**D**ecember**J**anuary**F**ebruary) anomalies of sea surface temperature, mean sea level pressure and the west-east component of the 10m wind (U10M) for 2071-2100 (IPCC A1B) relative to 1961-1990. The anomalies show the projected change of the seasonally averaged fields due to global warming relative to the according seasonally averaged fields of the undisturbed climate. The regional impact varies for the different seasons.

mately 200 km, yielding a grid size of 192x96 on 31 vertical levels. Since for this work we were mainly interested in the variability range from months to a few decades, we have only used a small subset of the original data: the monthly mean data of some typical surface fields (sea surface temperature (SST), 2m temperature (T2M), mean sea level pressure (MSLP), 10 m wind components (U10M, V10M), and precipitation rate (PREC)).

Most quantities show a strong annual cycle. In order to differentiate between the ”normal” variability and the variability induced by global warming, we have used the anomalies of these quantities relative to the simulated corresponding multiyear monthly means of the normal period 1961-1990. By building the difference between the monthly mean fields and the means of the corresponding 30 monthly mean fields of the reference period, the mean annual cycle is removed from the data. In Figure 1, we show the seasonally averaged northern hemisphere winter (**D**ecember**J**anuary**F**ebruary) and northern hemisphere summer (**J**une**J**uly**A**ugust) anomalies of the SST (a and d), MSLP (b and e) and U10M (c and f) for 2071-2100 relative to 1961-1990. For summer and winter, all mean anomalies show regionally different responses to the increase of greenhouse gases. Regionally different responses imply that the mean seasonal cycle is going to change differently for distinct locations in a warmer world. In Figure 1(a), a strong positive SST anomaly can be observed in the eastern tropical Pacific for DJF. This is exactly the area and season where positive SST anomalies are observed in ElNiño years. Further analysis is needed to investigate if changes in the variability are associated with the shift in the mean climate.

3 Wavelet Analysis

Wavelet analysis is a method to investigate signals that change over time. By decomposing the time series into time-frequency space, one can determine which frequencies are dominant in certain time intervals. Figure 2 illustrates this concept. While the upper image gives the original time series, the lower one shows a visual representation of the wavelet analysis, the power spectrum. The x-axis corresponds to time and the y-axis to frequency. Dark regions indicate periods in the data, where it oscillates with the corresponding frequency. The bathtub-shaped curve separates between valid (above) and invalid (below) areas. Between 2026 and 2046, for example, the data features a strong underlying oscillation at a period of 2 to 8 years. Using this visualization, changes in frequency can be observed easily. In the following, we will first explain the wavelet transform and afterwards the computation of the power spectrum and a significance test to detect relevant structures in the power spectrum.

3.1 Wavelet Transform

Assume that one has a time series, x_n, consisting of $n = 0,...,N-1$ time steps with equal spacing as given in Figure 2(a). The goal is to determine at which frequencies this time series oscillates on a local scale. Therefore, the data is filtered locally with an oscillating kernel (cf. Fig. 3(a)). The stronger the result of this filtering is, the stronger the data oscillates at the frequency encoded by the kernel. We require a local measure as the frequency of events might change due to climate change. Global techniques, such as the Fourier transform, are not able to detect such time-dependent characteristics.

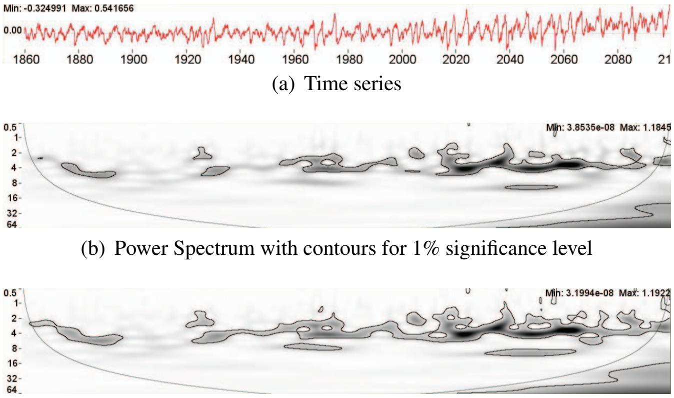
In wavelet analysis, the wavelet function forms the local kernel. A wavelet function, $\Phi_0(\eta)$, can be an arbitrary function, but has to fulfill two requirements: It must have zero mean and must be localized in both time and frequency space [4]. A wavelet widely used in climate research is the Morlet wavelet as depicted in Figure 3, consisting of a plane wave modulated by a Gaussian [31]:

$$\Phi_0(\eta) = \pi^{-1/4} e^{i\omega_0\eta} e^{-\eta^2/2} \tag{1}$$

where ω_0 is the non-dimensional frequency, here taken to be 6 to satisfy the admissibility condition [4]. As the Morlet wavelet is a complex function, it is better adapted for capturing oscillatory behavior as it will return information about both amplitude and phase [31].

The filtering of the time series with the wavelet could be easily performed in time space. However, this procedure would be very slow. From digital signal processing, it is well known that the computational intensive convolution in the time domain is a simple point-wise multiplication in the frequency domain (convolution theorem, e.g. [7]). Hence, we will first Fourier transform the time series and the wavelet, then multiply both transformed signals and finally perform an inverse transformation of the multiplied signals to receive the filtered signal. The discrete Fourier transform of the input time series x_n is given by

$$\mathscr{F}\{X\}_k = \hat{x}_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}, \tag{2}$$



(a) Time series

(b) Power Spectrum with contours for 1% significance level

(c) Power Spectrum with contours for 5% significance level

Fig. 2. Wavelet Transform: (a) Original time series. (b,c) Wavelet power spectrum of signal given in (a) with different levels of significance. The time axis is the same for all three illustrations.

where $k = 0,\ldots,N-1$ is the frequency index [19]. The Fourier transform of the Morlet wavelet is

$$\mathscr{F}\{\Phi_0\}_{s,\omega} = \hat{\Phi}_o(s\omega) = \pi^{-1/4}H(\omega)e^{-(s\omega-\omega_0)^2/2}, \tag{3}$$

where s is the scale and ω_k the angular frequency

$$\omega_k = \begin{cases} \frac{2\pi k}{N\delta t} & \text{for } k \leq \frac{N}{2} \\ -\frac{2\pi k}{N\delta t} & \text{for } k > \frac{N}{2} \end{cases} \tag{4}$$

and $H(\omega)$ is the Heaviside step function, $H(\omega) = 1$ if $\omega > 0$, $H(\omega) = 0$ otherwise. To ensure that the wavelet transforms at different scales are comparable, the wavelet function at each scale s is normalized to have unit energy:

$$\hat{\Phi}(s\omega_k) = \sqrt{\frac{2\pi s}{\delta t}}\,\hat{\Phi}_0(s\omega_k) \tag{5}$$

Now that we have transformed both signals to frequency space, they are multiplied point-wise. The wavelet transform is the inverse Fourier transform of the multiplied signal:

$$W_n(s) = x_n * \Phi_0 = \frac{1}{N}\sum_{k=0}^{N-1}\hat{x_k}\hat{\Phi}^*(s\omega_k)exp^{i\omega_k n\delta t}. \tag{6}$$

For further details on wavelet theory see [4, 19, 31].

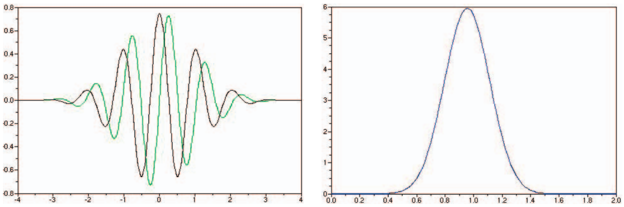
3.2 Power Spectrum

The wavelet transform as given in Equation (6) computes the filtered signal for a specific frequency given by the scale s of the wavelet. Hence, the result is a single line of the image in Figure 2(bottom). To provide the big picture about the changes in frequency over time, the analysis has to be performed on different levels/scales. Combining these different scales in a single image gives the wavelet power spectrum. The question that remains is: How many scales and how many samples at each scale are required? In orthogonal wavelet analysis, a "wavelet basis" is used, which is an orthogonal set of functions. The number of convolutions at each scale is proportional to the width of the wavelet basis at that scale. This most compact representation of the signal is useful for signal processing, but is rather difficult to interpret visually as it produces discrete blocks and an aperiodic shift in the time series results in a different image [31]. Thus, nonorthogonal wavelet analysis is commonly used which provides a smoother (arbitrary number of scales and number of samples per scale) and more intuitive representation of the data.

Torrence and Compo [31] suggest the following formula to compute the scale parameter:

$$s_j = s_0 2^{j\delta j} \qquad j = 0,\ldots,J \tag{7}$$
$$J = \frac{log_2(N\delta t/s_0)}{\delta j},$$

where s_0 is the smallest resolvable scale and J determines the largest scale. The choice of a sufficiently small δj depends on the width in spectral-space of the wavelet function. For the Morlet wavelet, 0.5 is the largest value for which δj gives adequate sampling. For more details see [31].



(a) Time domain

(b) Frequency domain

Fig. 3. Morlet Wavelet: (a) Plot of the complex Morlet function. The black line shows the real part, the green line the imaginary one. (b) The Fourier transform of the Morlet wavelet (s = $10\delta t$).

3.3 Test of Significance

The power spectra computed so far comprise a large variety of subtle structures and not all of them encode significant changes. When analyzing, for example, a random signal, the resulting power spectrum will feature coherent structures as well. To distinguish between significant and random structures, a test of significance can be used [13]. If a value in a power spectrum is considered to be statistically significant according to this test, it is unlikely to have occurred by chance. The confidence is quantified by the test's confidence level.

Every test of significance begins with a null hypothesis H_0. The null hypothesis tells what is commonly expected and is used to identify extraordinary events. To establish such a hypothesis, some knowledge/assumption about the underlying process is required. Therefore, a background spectrum is chosen and it is assumed that different realizations of the process will be randomly distributed about the background spectrum. For many geophysical phenomena, an appropriate background spectrum is either white or red noise [31].

A simple model for red noise is the univariate lag-1 autoregressive or Markov process [31]:

$$x_n = \alpha x_{n-1} + z_n, \tag{8}$$

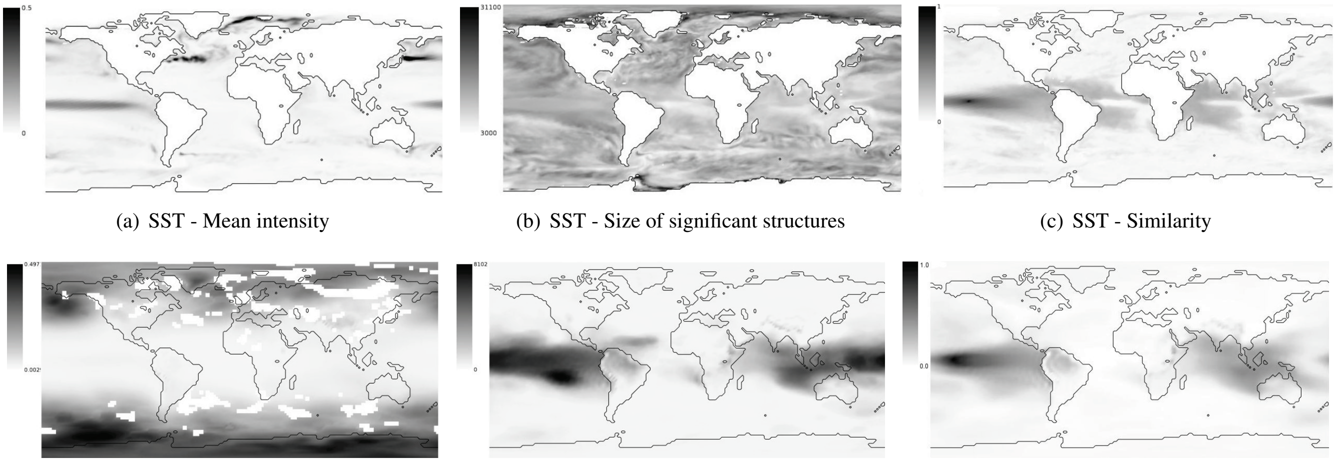
where α is the assumed lag-1 autocorrelation, $x_0 = 0$ and z_n are taken from Gaussian white noise. lag-1 autocorrelation measures the correlation between the current and the preceding time step. A red noise spectrum can be modeled using the discrete Fourier power spectrum of (8):

$$P_k = \frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos(2\pi k/N)}, \tag{9}$$

where $k = 0,\ldots,N/2$ is the frequency index. If $\alpha = 0$, the equation gives a white noise spectrum.

The null hypothesis for the wavelet power spectrum is that the time series x_n has a mean power spectrum given by (9). If a peak in the power spectrum of x_n is significantly larger than what would have been expected given (9), it is assumed to be a true feature with a certain confidence. The required confidence level is a parameter of the test. A common parameter is the 95% confidence level, which is equivalent to the 5% significance level. The significance level corresponds to the probability of observing an extreme value by chance. To choose the appropriate test statistic, the distribution of the time series has to be determined. Assuming that x_n is a normally distributed random variable, then both the real and imaginary part of the transformed signal $\hat{x}_k$ are normally distributed as well. Since the square of a normally distributed variable is χ^2 distributed with one degree of freedom (DOF), then $|\hat{x}_k|^2$ is χ^2 distributed with two DOFs, denoted by χ_2^2 [9]. Torrence and Compo [31] showed that if the original Fourier components are normally distributed, then the wavelet power spectrum $|W_n(s)|^2$ is χ_2^2 distributed:

$$\frac{|W_n(s)|^2}{\sigma^2} \Rightarrow \frac{1}{2}P_k\chi_2^2 \tag{10}$$



(a) SST - Mean intensity (b) SST - Size of significant structures (c) SST - Similarity

(d) MSLP - Mean intensity (e) MSLP - Size of significant structures (f) MSLP - Similarity

Fig. 4. Display of wavelet characteristics: (a,d) Mean intensity of the significant structures in each power spectrum. White areas indicate positions whose power spectra feature no significant structures. (b,e) Size of the significant structures given by number of included pixels. (c,f) Similarity fields for a position in the Nino3 region.

at time t and scale s, where ⇒ indicates "is distributed as" and k is the frequency index.

To test whether a pixel in the power spectrum belongs to a significant feature, it has to be compared to the critical value given by the test statistic (10). To obtain the critical value, the variance σ^2 of the time series is computed first. Afterwards, the background spectrum (9) is multiplied by the critical value of the χ_2^2 test and the variance and divided by two. If the value at the current pixel is greater than the critical value of the corresponding frequency, the pixel is considered to belong to a significant feature.

Examples of significant regions with different significance levels are shown in Figure 2.

4 Analysis of Climate Data

When applying the theory explained so far to real world data sets, different aspects in the implementation and visualization have to be considered. In the following, we will first give a few details on the implementation, continue with the post-processing of the wavelet plots to allow for an easier analysis of the data and conclude with the layout of the graphical user interface that is used for visualization and exploration.

4.1 Implementation Details

The first step in the wavelet analysis pipeline is the computation of the wavelet power spectra. The computation of a single line in a power spectrum consists of three steps: (1) Fourier transformation of the data and the wavelet. (2) Multiplication of the two signals in Fourier space. (3) Inverse transformation of the multiplied signal.

For the Fourier transform $\hat{x_n}$ of the input time series x_n, we use the FFTW library [5] which provides a fast implementation of the DFT. The DFT of the data is computed only once for the entire power spectrum. On each scale, however, the Fourier transform of the wavelet $\hat{\Phi}(s\omega_k)$ has to be recomputed. As the transformed wavelet can be given in analytic form, the Fourier transforms of $\Phi(s\omega_k)$ with required angular frequencies ω_k (see Eq. 4) can be directly evaluated using Equation 3. The result of the multiplication of corresponding data and wavelet frequencies is stored in a new data vector. The inverse DFT of the multiplied signal in the new vector is performed using the FFTW library. This procedure is conducted for all scales s_0 through s_J and stored in an image (matrix of size $N \times nScales$).

The computation of the power spectra for all time series at all positions in the data set is rather costly (about 1h for each variable). In the following, we will compute coherent structures in the data by comparing power spectra at different positions. To increase the performance in these steps, we store a discretized version of each power spectrum using 1 byte (range [0;255]) per pixel value. Hence, we need a quantization stepsize. Using the minimum and maximum of the entire data (all positions at all time steps), a theoretical global maximum of the wavelet transform can be computed analytically. This maximum would allow for an automated quantization of all power spectra (quantization: max/256). However, this theoretic maximum is hardly ever met and the actual maximum in the real data is much smaller and depends heavily on the structure of the data. Thus, we use an adaptive quantization that requires slightly more memory by storing the minimum and the maximum for each power spectrum and the individually quantized version of it (quantization: (max-min)/256).

For the test of significance, Equation (9) has to be evaluated which requires the lag-1 autocorrelation α. This parameter is estimated for each variable separately. The lag-k autocorrelation function of a time series x_n with N time steps is defined as [1]

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{11}$$

where $\bar{x}$ denotes the mean of the time series. The lag 1 coefficient is evaluated for each position and the global lag 1 autocorrelation parameter is the mean of these values.

4.2 Information-Assisted Variability Analysis

The visual representation of the power spectrum of a time series gives a simple and easy to analyze overview over significant frequencies and their distribution through time. However, the wavelet analysis of a standard climate data set with grid size 200×100 positions and 2000 time steps results in a set of 20,000 images (∼ 2000×50 pixels each). A simple investigation of all these images is very labor-intensive and cumbersome. Thus, methods are required that provide a more abstract representation of the wealth of information. In the following, we will present three strategies: The first method extracts scalar characteristics about the individual power spectra, which can be displayed using a colormap. The second one clusters regions with similar patterns in the spectra and allows for information-assisted interaction and the third technique allows for the identification of reoccuring patterns in different places of the data set using similarity fields.

4.2.1 Definition of Characteristics - How to Find Power Spectra with Special Properties

Special characteristics of power spectra can give an easy overview over the structure in the entire data set. Figures 4(a) and 4(d), for example, highlight areas where the power spectra feature structures with very high values and hence, high variability in the data. One such simplifying statistic is to assign each position in the data set the mean value

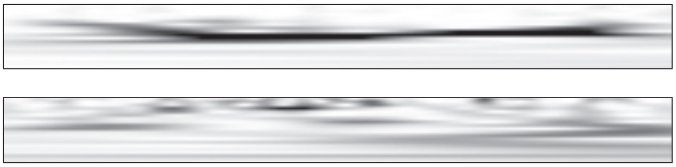


Fig. 5. Wavelet power spectra featuring different structures.

of the power spectrum. However, this simple statistic often hides relevant structures, as can be seen when comparing the two power spectra in Figure 5 which have both the same mean (average over all pixel values). While the first image features one distinct dark structure on a light background, the second one lacks such a clear formation. To allow for a better characterization of significant structures, we first have to partition the image into relevant structures and background. The test of significance described in Section 3.3 gives an isovalue (critical value) for each scale. Pixels holding a value above this threshold belong with high probability to a significant feature. Using these isovalues, the power spectra can easily be segmented into relevant signal and noise (cf. Fig. 2).

After the segmentation of the power spectrum, different scalar quantities can be derived. Figure 4 depicts (a,d) the mean intensity of the structures in each power spectrum and (b,e) the size in pixels of significant structures for different variables. If the user is interested in signals with a large amplitude, the mean intensity is the better indicator. Large structures that often extend over a longer time period can be found using the second statistic. A combination of both characteristics helps finding large structures with large values. Furthermore, we found the following statistics helpful: the number of contours, the length of the longest contour and whether the major contour exhibits an upwards or downwards trend, revealing phenomena that change their frequency over time.

4.2.2 Clustering using Mutual Information - How to Find Coherent Structures

While the first technique concentrated on the aggregation of individual power spectra to a single scalar value, we will now concentrate on the identification of regions with similar patterns. Thus, the physical domain is divided into different subregions each holding a characteristic pattern. Picking few large subregions and displaying their power spectra simultaneously gives a more detailed overview over the entire data set and its subdivision into different climatic behavior.

From the large variety of clustering algorithms, we chose the region growing paradigm as we are looking for contiguous spatial subsets. The region growing requires three additional pieces of input: Positions where to start the region growing, a distance measure and a stopping criterion which tells when to abort the growing process.

As starting positions we chose regions with structures holding high values. Therefore, each power spectrum in the data set is segmented using the thresholding technique from the previous section and the mean of the values above the threshold is computed. The vector of mean intensities is sorted in decreasing order. The region growing starts at the position with the power spectrum featuring the structure with highest values. The distance measure is the mutual information between both power spectra. Mutual information is a well known similarity measure widely used in image registration. To define mutual information, we need a few basic quantities from information theory [2].

Entropy is a measure of the uncertainty in the random variable X and gives the number of bits that are required on average to describe this random variable. Let X be a discrete random variable with alphabet $\mathscr{X}$ and probability mass function $p(x) = Pr(X = x)$, $x \in \mathscr{X}$. The entropy $H(X)$ of a discrete random variable X is defined by [2]

$$H(X) = -\sum_{x\in\mathscr{X}} p(x)\log_2 p(x). \tag{12}$$

The joint entropy measures the uncertainty in a pair of random variables. The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as [2]

$$H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log_2 p(x,y) \tag{13}$$

The mutual information is a measure of the amount of information that one random variable contains about another. Alternatively, it can be thought of as the reduction in the uncertainty of one random variable due to the knowledge of the other one. Consider two random variables X and Y with joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X,Y)$ is then defined by [2]

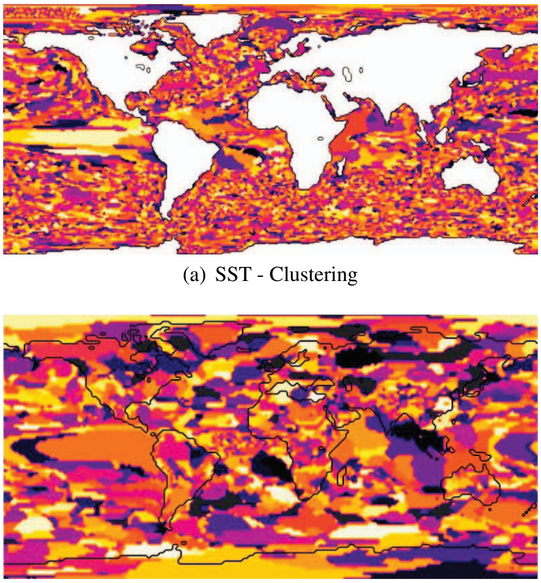
$$\begin{aligned} I(X,Y) &= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\ &= H(X)+H(Y)-H(X,Y). \end{aligned} \tag{14}$$

Mutual information is symmetric, i.e., $I(X,Y) = I(Y,X)$. It can be thought of as the dependence between the two random variables and is equal to zero if and only if X and Y are independent.

In the present case, the two random variables X and Y represent the pixel values in the two power spectra to be compared. The probability mass function $p(X = x)$ gives the probability that the power spectrum contains a pixel of value x. Therefore, a distribution containing 256 bins ranging from 0 to 255 is constructed and for each pixel in the power spectrum, the bin with the corresponding value is incremented. The probability $p(X = x)$ is given by getting the value of bin x and dividing it by the number of pixels in the power spectrum. Analogous to the individual distributions the joint distribution $p(x,y)$ is constructed by using a matrix of size 256×256 and counting the cooccurrence of value x in the first and value y in the second power spectrum at the same pixel coordinate. Again, probabilities are calculated by dividing the matrix entry by the number of pixels in the power spectrum. Afterwards, the mutual information, i.e., the distance between two power spectra, can be directly evaluated using Equation (14).

As mutual information as defined in Equation (14) has no upper bound, the choice of an appropriate stopping criterion would be difficult. Hence, normalized mutual information is used to obtain similarity values in the range [0;1]. Normalized mutual information is defined by [29]

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}. \tag{15}$$



(a) SST - Clustering

(b) MSLP - Clustering

Fig. 6. Clustering of power spectra using mutual information.

Several other definitions of normalized mutual information exist (e.g. [22]), which achieve similar results. We chose the normalization presented in equation (15), as it ranges precisely from 0 to 1 ($NMI(X,X) = 1$) and as it is analog to a normalized inner product in Hilbert space.

The third input parameter is the stopping criterion, which specifies how much information the current power spectrum has to contain about the central one (the starting point of the growing process) in order to belong to the same cluster. The stopping criterion is commonly a scalar parameter specified by the user telling how much difference between power spectra is allowed. An automatic suggestion would be helpful, but is difficult to provide, as it highly depends on the structure of the data. Finding an appropriate threshold is similar to finding a good isovalue when displaying isocontours, which is not solved either. In our application we commonly found a threshold for minimal normalized mutual information of 0.5 appropriate. Further studies might give better insight into a good automatic choice.

Now that we have the three input parameters, the clustering can be computed. Starting at the power spectrum containing structures with the highest mean, we iteratively add pixels in the 4-neighborhood of the current cluster whose normalized mutual information to the power spectrum of the starting point is greater than the threshold. A status vector is updated synchronously. Value -2 indicates that the pixel has not yet been assigned to a cluster. Value -1 indicates that the pixel is already in the queue of pixels to be checked for similarity. Positive values give the ID of the cluster the pixel has been assigned to. The algorithm stops when all pixels have been assigned to a cluster, i.e., the power spectrum with lowest structural mean is reached.

4.2.3 Similarity Fields - How to Find Locations With A Similar Evolution

In climate research the statistical characteristics of several phenomena and their location are well known. In the coupled climate system, however, events at distinct locations might be linked to each other over large distances or at different phases. Understanding such teleconnections, which, for example, are known for the ENSO phenomenon [15], helps to understand the physical mechanisms of these climate anomalies. To provide assistance in this direction, we found similarity fields helpful. Similarity fields show for a defined position or cluster the normalized mutual information between each position in the data set and the reference point or the mean of the reference subset. Figures 4(c) and 4(f) illustrate this concept for a position in the Nino3 region. The SST anomaly shows high similarity values for most of the tropical oceans, implying that teleconnections are likely to occur in these regions. The similarity field for the MSLP shows a large and strong maximum in the central tropical Pacific as well as a large maximum in southeast Asia - both regions which are known to be teleconnected within the El Niño Phenomenon. This example nicely illustrates the ability of the similarity fields to highlight teleconnections.

4.3 GUI and Interaction

The GUI of the wavelet analysis program consists of several windows. The main window as illustrated in Figure 7 visualizes the data set to be analyzed using a colormap. The user can interactively switch between the following representations: single time step of the original data, clustered regions, and property fields (i.e., mean intensity or size of relevant structures). By clicking with the right mouse button into the field, a pop-up menu is started providing access to additional information. The user can display the time series data and/or the wavelet power spectrum at the selected position. Moreover, we integrated an automatic selection of regions that are relevant in climate change research such as the Nino3 and Nino3.4 regions [32] and the positions used for the definition of the NAO and SOI index and corresponding differences. When selecting either of the options an additional window opens to display the required information.

5 Results

In order to verify the robustness of the framework described above, we have analyzed the IPCC A1B simulation data with respect to the suspected variability changes of the El Niño phenomenon. We always repeated the same analysis steps with a subset of the Control simulation (CTL), where the CO_2 concentration was kept frozen at the preindustrial level of 1860. Changes of the frequency and the magnitude of the fluctuations caused by climate change should supposedly not be detectable in the CTL run.

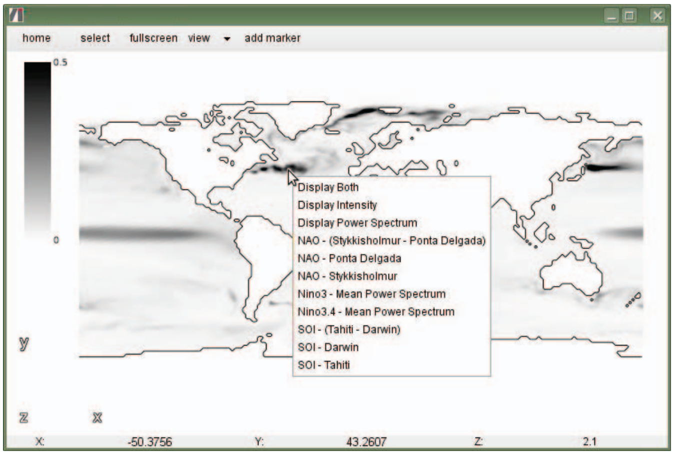


Fig. 7. Main window of the wavelet analysis application.

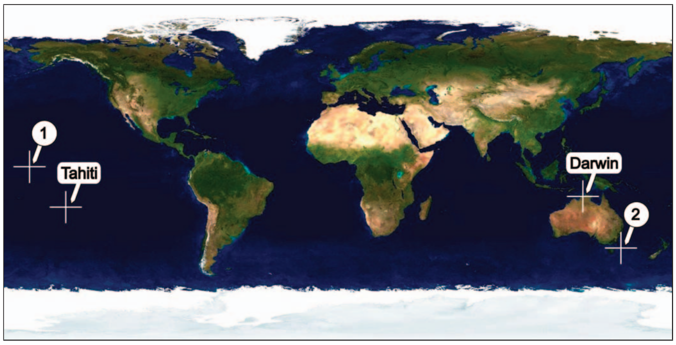
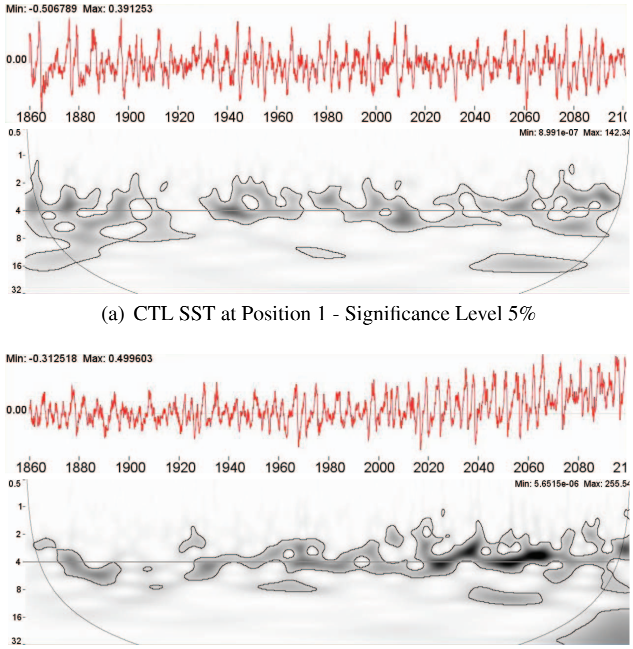


Fig. 8. Analyzed positions: The difference between the mean sealevel pressure of Tahiti and Darwin is used to define the southern oscillation index (SOI). Position 1 is located inside the Nino3 region. Extraordinary wavelet patterns are observed at position 2. (Image courtesy of the blue marble project [28].)

Figure 4(a) shows the intensity map of the SST variability for the entire time series. A pronounced maximum spans the tropical Pacific in east-west direction. In the northern mid-latitudes, two other even stronger maxima are visible. These are frontal regions, where the subpolar and subtropical gyre waters meet. In the Atlantic, variations in the position of the strong meridional temperature gradient between northward advected warm Gulf Stream water masses and southward advected cold water masses of the Labrador Current can be observed. Variations in the positions of the front cause the regional maximum of temperature variations. Another strong maximum is located in the north Atlantic at the edge of the Arctic sea ice. The maximum reflects the retreat of the sea ice due to anthropogenic warming.

The map with the size of the significant regions (Fig. 4(b)) does not show outstanding features for the SST, except for some maxima along the sea ice edges. The clustering (Fig. 6), though, shows some larger and therefore interesting structures in the Tropics. In the mid-latitudes, the clustering yielded mainly smaller regions, which can be explained by a spatially more heterogeneous variability.

The intensity map for the MSLP (Fig. 4(d)) shows quite strong maxima in the mid latitudes and over the poles, while the variability in the Tropics is relatively weak. The analysis of the size of the significant structures (Fig. 4(e)), though, results in strong maxima in the tropics. The statistical properties of the cyclone activity in the mid latitudes causes strong variability, while the less noisy MSLP patterns in



(a) CTL SST at Position 1 - Significance Level 5%

(b) A1B SST at Position 1 - Significance Level 5%

Fig. 9. Wavelet power spectrum and original time series of (a) CTL SST and (b) SST for a position in the Nino3 region (position 1 in Fig. 8).

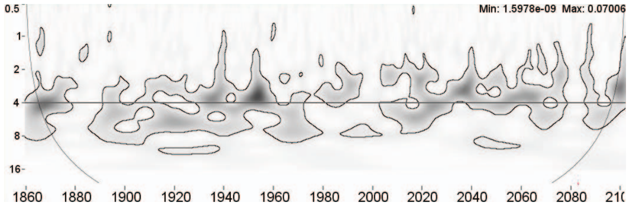


Fig. 10. Southern Oscillation Index (SOI): Wavelet power spectrum for the difference time series between the time series of Tahiti and Darwin as used for the definition of the SOI (cf. positions in Fig. 8).

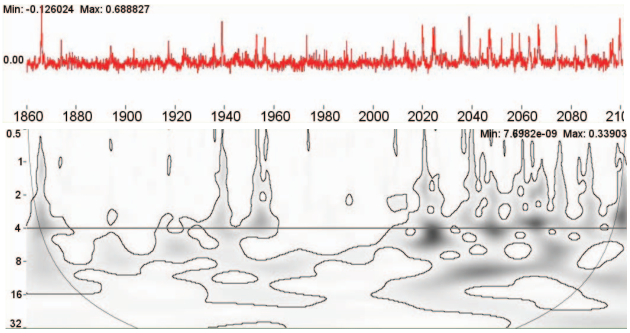


Fig. 11. Time series and wavelet power spectrum of the west-east component of the 10m wind (U10M) west of the Nino3 region.

the Tropics allow to identify larger similar regions, finally resulting in larger coherent structures through the clustering process (Fig. 6).

In order to find and analyze teleconnections, it proved to be helpful to display regions with similar variability properties for a given region or location. In Figure 4(c) and 4(f), the similarity of the SST and MSLP variability is shown for a selected location within the Nino3 region.

Figure 9 shows, for the Nino3 region, the time series of the SST (red curves) and the according temporal development of the power spectrum for the CTL experiment (a) and for the IPCC A1B simulation (b). While the CTL simulation (a) does not show any trends in the time series and in the power spectrum, the expected global warming trend and a shift towards enhanced amplitudes can be observed in the time series of the A1B run (b). The according power spectrum shows a shift towards shorter periods: throughout the 21th century, an increasing number of events can be observed in the range between 2 and 4 years.

The variability of the SOI shows a similar change towards shorter time scales as the SST in the Nino3 region (reduction of energy in time band between 4 and 8 years and enhanced variability below 2 years). However, the enhancement of the amplitude of variability is not as obvious (cw. Fig. 10). The zonal wind on the equator west of the Nino3 region shows a behavior rather similar to the SST (Fig. 11): enhanced variance and more variability during the 21st century with time scales below 4 years. But this quantity also shows a tendency towards more decadal variability.

The time series of SST east of Tasmania shows a quite interesting behavior (see position 2 in Fig. 8 and wavelet plot Fig. 12). The strong amplitude in the 1-year band reflects a change of the seasonal cycle of the SST in a warming climate. The temperature maximum shifts from February to March and the coldest month from September to October (Fig. 13). Besides, the warming in summer is stronger than in winter, thus leading to a larger amplitude of the seasonal cycle. For the last 50 years of the experiment, IPCC-A1B shows characteristics of jumps between two separate states. This would explain the strong increase of the decadal variability towards the end of the experiment. Figure 14 shows the mean zonal wind (U10M) for the period 1961-1990 encoded as height field. The height field is colorized by the projected change of U10M for 2071-2100 relative to 1961-1990. The long and high ridge north of Antarctica show the westerlies, a very strong wind band around Antarctica. The red southern side and blue northern side of the ridge visually show the simulated annual mean poleward shift of the westerlies due to climate change. The associated changes in wind stress curl cause a poleward shift of the gyre systems in the South Pacific. As a consequence warm subtropical waters penetrate further poleward east of Australia/Tasmania. Whereas in the first part of the simulation (and in the control simulation) the influence of subtropical water masses at this location is rather small, it increases in the course of the simulation. Decadal variability and the associated movement of a strong water mass front lead to the apparently bipolar distribution of SSTs at the end of the simulation.

6 Discussion and Conclusion

In this paper we described a framework based on wavelet analysis to investigate variability climate change data. To make the wavelet analysis applicable to an entire 2D multivariate data set, three different approaches were proposed: First, scalar characteristics allow for an easy overview over the entire data by compressing the data to a single scalar field. Second, the clustering based on mutual information partitions the domain into regions of similar variability patterns and can be used to either investigate the spatial dynamics of climate change or to compare larger regions featuring similar patterns. The third technique, similarity fields, supports the expert user in identifying regions that feature similar temporal patterns as familiar ones. Additionally, areas and locations used by climate researchers to track climate variability have been integrated into the framework. Direct access to the according visualizations for the predefined markers allows for the comparison of the data with references found in the literature.

Using the current implementation of the wavelet transform, the computation of all power spectra for one data set takes approximately one hour on a standard PC. After this pre-processing step the analysis of the loaded data set is possible with interactive response time. To decrease preprocessing time, the power spectra can be computed in parallel. As each computation requires only a single time series of approximately 3000 double values, memory is not an issue and computing time can be reduced tremendously. A case study evaluating the use of the different scalar characteristics would be beneficial to automatically suggest appropriate techniques for different fields.

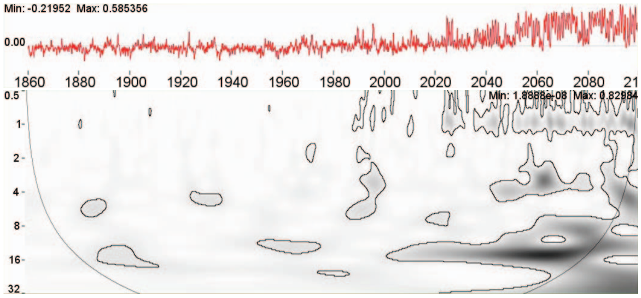


Fig. 12. Power spectrum and time series of the SST at position 2 - significance level 5%

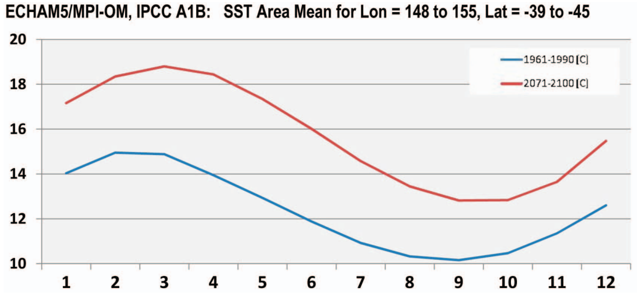


Fig. 13. Mean annual cycle of the SST for an area surrounding position 2: (red) 2071 - 2100 and (blue) 1961 - 1990 .

Extending the results of [15] we were able to visually analyze the transient change of El Niño frequency due to climate change. Moreover, we could identify further interesting warming induced variability changes east of Tasmania, which has not been described in literature before.

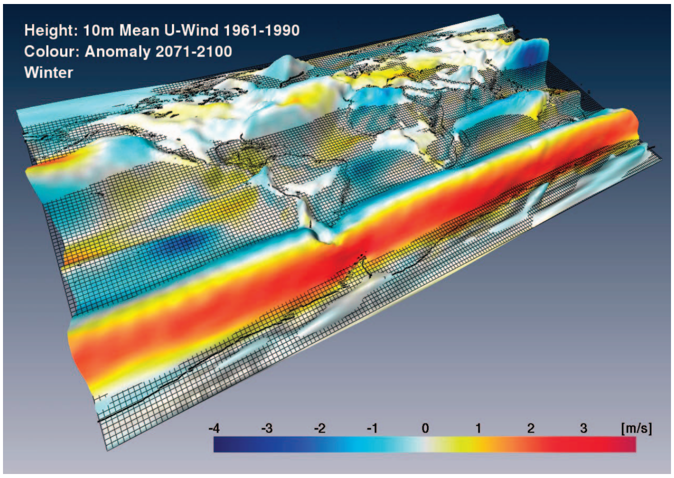


Fig. 14. Visualization of global changes for U10M (DJF): height encodes the mean U10M of the years 1961 – 1990, color encodes the projected change for the U10M wind for 2071 – 2100 relative to the same period.

References

[1] G. E. P. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[3] H. Doleisch, P. Muigg, and H. Hauser. Interactive visual analysis of hurricane isabel with simvis. In *IEEE Visualization (Vis'04) Contest*, 2004.

[4] M. Farge. Wavelet transforms and their applications to turbulence. *Annual Review of Fluid Mechanics*, 24:295–457, 1992.

[5] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005.

[6] M. Gross, O. Staadt, and R. Gatti. Efficient triangular surface approximations using wavelets and quadtree data structures. *Visualization and Computer Graphics, IEEE Transactions on*, 2(2):130–143, June 1996.

[7] B. Jähne. *Digital Image Processing - Concepts, Algorithmus and Scientific Applications*. Springer-Verlag, Heidelberg, 1991.

[8] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1459–1466, 2008.

[9] G. M. Jenkins and D. G. Watts. *Spectral Analysis and Its Applications*. Holden-Day, 1968.

[10] K. Lau and H. Weng. Climate signal detection using wavelet transform: How to make a time series sing. *Bull. Am. Met. S.*, 76:2391–2402, 1995.

[11] K.-L. Ma and H.-W. Shen. *Vis. Handbook*, chapter Visualization Techniques for Time-Varying Volume Data. Academic Press, Inc., 2004.

[12] D. Middleton, T. Scheitlin, and B. Wilhelmson. *The Vis. Handbook*, chapter Visualization in Weather and Climate Research. Elsevier, 2005.

[13] L. B. Mohr. *Understanding significance testing*. Sage Pub., Inc., 2003.

[14] C. M. Moy, G. O. Seltzer, D. T. Rodbell, and D. M. Anderson. Variability of el nino/southern oscillation activity at millennial timescales during the holocene epoch. *Nature*, 420:162–165, 2002.

[15] W. A. Müller and E. Roeckner. ENSO teleconnections in projections of future climate in ECHAM5/MPI-OM. *Clim. Dyn.*, 31:533–549, 2008.

[16] G. M. Nielson, I.-H. Jung, and J. Sung. Haar wavelets over triangular domains with applications to multiresolution models for flow over a sphere. In *IEEE Visualization*, pages 143–150, 1997.

[17] T. Nocke, S. Schlechtweg, and H. Schumann. Icon-based visualization using mosaic metaphors. In *Ninth International Conference on Information Visualisation (IV'05)*, pages 103–109, 2005.

[18] T. Nocke, T. Sterzel, M. Böttinger, and M. Wrobel. Visualization of climate and climate change data: An overview. In *Ehlers et al. (Eds.) Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research (ISDE'08), Wichmann, Heidelberg*, pages 226–232, 2008.

[19] D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000.

[20] P. Piscaronoft, J. Kalvová, and R. Brázdil. Cycles and trends in the czech temperature series using wavelet transforms. *Intern. Journal of Climatology*, 24(13):1661–1670, 2004.

[21] R. W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography (Develop. in Atmos. Sci.)*, volume 17. Elsevier, 1988.

[22] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[23] W. Ribarsky, N. Faust, Z. Wartell, C. Shaw, and J. Jang. *R. Ladner, K. Shaw, and Mahdi Abdelguerfi, Editors: Mining Spatio-Temporal Information Systems*, chapter Visual Query of Time-Dependent 3D Weather in a Global Geospatial Environment. Kluwer, 2002.

[24] K. Riley, D. Ebert, C. Hansen, and J. Levit. Visually accurate multi-field weather visualization. In *14th IEEE Visualization 2003 (VIS 2003)*, 2003.

[25] E. Roeckner et al. *The atmospheric general circulation model ECHAM 5. PART I: Model description*. Tech. Rep. 349, Max Planck Inst. for Meteorol., Hamburg, Germany, 2003.

[26] A. Simmons, J. Wallace, and G. Branstator. Barotropic wave propagation and instability, and atmospheric teleconnection patterns. *J. Atmos. Sci.*, 40:1363–1392, 1983.

[27] D. M. Sonechkin and N. M. Datsenko. Wavelet Analysis of Nonstationary and Chaotic Time Series with an Application to the Climate Change Problem. *Pure and Applied Geophysics*, 157:653–677, 2000.

[28] R. Stöckli, E. Vermote, N. Saleos, R. Simmon, and D. Herring. The blue marble next generation - a true color earth dataset including seasonal dynamics from modis. *Sumitted to EOS (AGU)*, 2005.

[29] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining partitionings. In *Proc. Conference on Artificial Intelligence (AAAI 2002), Edmonton*, pages 93–98. AAAI/MIT Press, 2002.

[30] A. Timmermann, J. Oberhuber, A. Bacher, M. Esch, M. Latif, and E. Roeckner. Increased el niño frequency in a climate model forced by future greenhouse warming. *Nature*, 398:694–697, 1999.

[31] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78, 1998.

[32] K. E. Trenberth. El Niño Definition. *Exchanges, Newsletter of the Clim. Variability and Predictability Programme (CLIVAR)*, 1(3):6–8, 1996.

[33] Working Group I contribution to the Fourth Assessment Report of the IPCC. IPCC AR4: Climate Change 2007 - The Physical Science Basis. Technical report, IPCC, 2007.