



Multiscale evaluation of CMIP5 models using wavelet-based descriptive and diagnostic techniques

Marco Gallegati¹

Received: 14 May 2020 / Accepted: 10 November 2021
© The Author(s) 2022

Abstract

Time–frequency localization of model–data discrepancies may provide useful information for climate models inter-comparison, and especially for the goals of climate model refinement and improvement. CMIP5 models of the long-term historical (1850–2005) run experiment are compared using wavelet-based multiscale descriptive and diagnostic techniques with interesting results. Wavelet coherence maps can visualize the ability of alternative CMIP5 models to capture the observed climate variability at different time scales, while the performance of each CMIP5 model is assessed using goodness of fit relative measures on a scale-by-scale basis. Finally, the plots of wavelet decompositions of CMIP5 models and observed temperature series at different scales can detect and locate model/data disagreements across frequencies and over time, thus providing useful information to researchers for model diagnostic refinement and improvement.

Keywords CMIP5 model simulations · Wavelet transform · Time-frequency error · Multiscale evaluation

1 Introduction

The Coupled Model Intercomparison Project (CMIP) makes available to climate scientists a coordinated multi-model experiment that provides simulations of historical climate variability and change (Meehl and Coauthors 2007, Taylor et al., 2012). The standardized output from different models greatly facilitates model intercomparability through direct comparison between derived model simulations and the corresponding observed data. Figure 1 shows the dataset used in this study: the annual simulation of global near-surface temperature anomalies for each of the 44 CMIP5 models along with HadCRUT4 observations (thick solid line) superimposed. CMIP5 model simulations are able to reproduce the observed trend in the global-scale surface temperature. CMIP5 climate models are able to reproduce the observed warming observed over the last 150 years and the other recent climate changes. Nonetheless, large inter-model spreads relative to the model mean change and systematic errors are still evident, as well as differences between simulated and

✉ Marco Gallegati
marco.gallegati@univpm.it

¹ DISES, Polytechnic University of Marche, Ancona, Italy

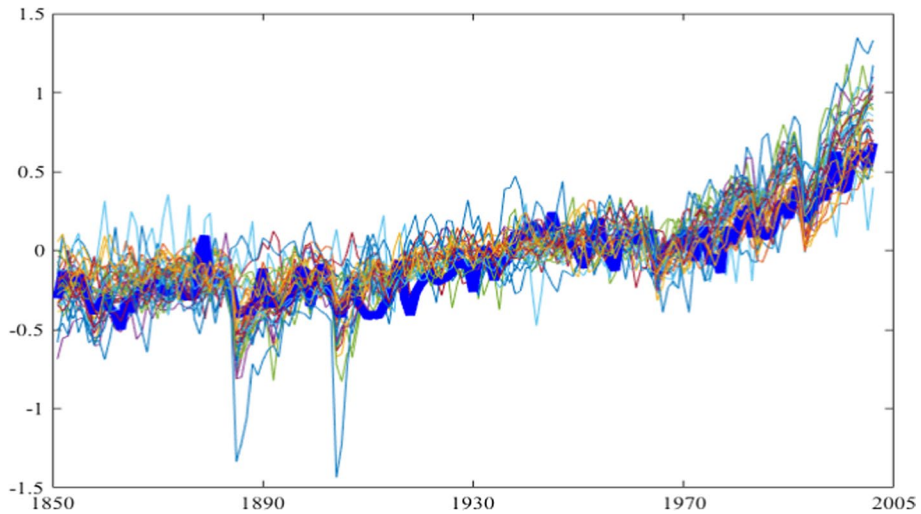


Fig. 1 Observed (thick line) and CMPI5 simulated (thin lines) anomalies in annual global mean surface temperature: historical run (1850–2005)

observed trends over periods as short as 10 to 15 years, e.g., 1870–1890, pre- and post-WWI, pre- and post-WWII, early 2000s.

Climate change studies, and simulation models, show that the climate system is an inherently multi-scale system, the complex dynamic pattern of which is due to the combination of anthropogenic and natural forces operating at different time scales (Mitchell, 1976). In this respect, climate models that are not able to combine shorter-term effects of natural climate variability within long-term anthropogenic-induced global warming are unlikely to capture the complex nature of global climate change and variability (e.g., Lin and Franzke, 2015, and Gallegati, 2018).

Model-data comparisons generally make use of quantitative statistical measures to analyze discrepancies between simulated and observed time series (Taylor, 2001). Traditional evaluation measures, such as the root mean square error (RMSE), only provide an overall idea about the model performance. As such they are extensively criticized for their inability to provide any insight into model/data disagreements and especially for not being effective enough in model refinement (e.g., Briggs and Levine, 1997; Gupta et al., 2008). Performance measures that neither distinguish between different types of error, nor take into account the scale and time dependency of errors are unlikely to provide a comprehensive assessment of global climate model performance (e.g., Reusser et al., 2008). More powerful forms of evaluation of model-data disagreement can be obtained by using methods that lead to enhanced comparisons by localizing the error at particular points in time and/or the associated scale or frequency at which these errors occur (e.g., Mahecha et al. 2010; Vargas et al., 2010, Vargas and Coauthors, 2013, Dietze and Coauthors, 2011).

The time–frequency localization property of the wavelet transform is particularly useful for identifying the scale and time dependency of errors in global climate models. Wavelet-based techniques have been widely applied in geophysical and climatological studies to investigate the relationships between temperature and dominant patterns of climate variability, such as the El Niño–Southern Oscillation (ENSO) and North Atlantic Oscillation (NAO) (e.g., Hudgins et al., 1993, Lau and Weng, 1995, Kumar

and Foufoula Georgiou 1997, Torrence and Compo, 1998, Torrence and Webster 1999, Park and Mann, 2000, Jevrejeva et al., 2003, Grinsted et al., 2004, among many others). Recently, Lin and Franzke (2015) and Gallegati (2018) have shown that wavelets are able to capture the multi-resolution temporal structure characterizing global average near-surface temperature anomalies.

Wavelet-based techniques can provide researchers with graphical diagnostic techniques and performance metrics at multiple time scales that can provide useful insights for model evaluation and refinement (development). Wavelet analysis, especially wavelet coherence, has been already used in comparing ecological models with the goal of model improvement (Williams et al., 2009; Wang et al., 2011). In this paper, similarly to Braverman et al. (2017) and Gong et al. (2018), we apply the wavelet-based decomposition methodology for the evaluation of CMIP5 climate models. However, while Braverman et al. (2017) and Chatterjee (2019) define climate as the coarsest level component from a wavelet decomposition of the observed data, Gong et al. (2018) and this paper focus on the oscillatory patterns at various wavelet scales so that it is more apparent which components of observed data are emulated by climate models. Moreover, differently from Braverman et al. (2017), Chatterjee (2019), and Gong et al. (2018), which use a hypothesis testing approach for evaluating whether model-simulated and observed climate-scale signals arise from the same population, we check the ability of the CMIP5 global climate models to reproduce the observed climate variability at all time scales using the continuous (CWT) and the discrete wavelet transform (DWT). Wavelet coherence plots, by visually comparing maps representing the (dis)agreement between simulations and observations in the time–frequency plane, may provide researchers with descriptive and diagnostic information about the different ability of CMIP5 models to capture the observed climate variability at different time scales. Wavelet decomposition allows an explicit quantitative assessment of model/data discrepancies, and the associated time scale over which these occur. The comparison of climate model simulations to observations using scale-based goodness of fit relative measures, such as the root mean square error (RMSE), allows to quantitatively evaluate the differences on a scale-by-scale basis. And the plots of wavelet decompositions of CMIP5 models and temperature series at different scales can provide useful information for model refinement and improvement.

Multiscale comparison of CMIP5 model results with observations suggests the presence of both similarities and differences in their performance. Wavelet coherence maps show that global climate models generally display a good performance in reproducing the observed climate variability at longer and multidecadal time scales, with this ability progressively extending to all time scales in recent decades. The analysis of model-data disagreements at different time scales shows models that provide “similar” performances according to global performance measures can differ considerably when analyzed at separate time scales. With some models performing better than others at certain time scales and/or periods, no individual model clearly emerges as “the best” overall in terms of model/data comparison. Interestingly, we show that model assessment can be further refined by explicitly locating model-data disagreements across frequency and over time through the visual comparison of the decomposed wavelet components for model results and observations. The localization of time-frequency model-data disagreements not only provides a useful perspective for climate model intercomparison, but may also be of particular interest for the goal of model improvement by pointing out model development needs.

2 Methodology

The approach used in this study for model intercomparison is based on a multiscale assessment of model performance. In particular, if the dynamics of the simulated and observed process are similar, decomposing simulated and observed time series into their time scale components allows localization of model-data disagreement in time as a function of frequency.

Wavelets are mathematical functions that transform a signal into a mathematically equivalent representation and cut up data into a set of time scale components, each associated with a specific frequency band and with a resolution matched to its scale into different frequency components, each with a resolution matched to its scale. In contrast to Fourier analysis, where the basis functions are defined globally over the whole computational domain, the wavelet transform uses a set of orthogonal basis functions, named wavelets, which are defined by a compactly supported, localized wavelet function that is dilated (or compressed) and shifted to provide a flexible time-scale window. Thus, unlike spectral methods,¹ wavelet analysis has the ability to handle a variety of nonstationary and complex signals, like those, and to attain an optimal trade-off between time and frequency resolution levels.

There are two types of wavelet transforms: the continuous wavelet transform (CWT), and the discrete wavelet transform (DWT). The CWT operates on smooth continuous functions and decompose signals on all scales. Therefore, the CWT is a highly redundant transform that produces information in a two-dimensional format where each wavelet coefficient is represented by a pair of data, designating time, or location, and scale (Gencay et al. 2003). The discrete wavelet transform (DWT), the discrete analogue of the CWT, produces information in the form of a time series (Percival, 2008). The key difference with the CWT is that the DWT uses only a limited discrete number of translated and dilated versions of the wavelet basis, with scales and locations normally based on a dyadic arrangement (i.e., integer powers of two). The wavelet method seems to be a very promising (model-free) approach to frequency extraction problems, as it allows simultaneous estimation of different unobserved components without making any explicit assumption about the characteristic of the data generating process.

The continuous wavelet transform (CWT) of a signal $x(t)$ with respect to the wavelet function ψ is a function $W_x(s, u)$.

$$W_x(\lambda, \tau) = \int_{-\infty}^{\infty} \psi_{\lambda, \tau}(t)x(t)dt$$

where the wavelet basis, called “mother wavelet”, defined as.

$$\psi_{\lambda, \tau}(t) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{t - \tau}{\lambda}\right)$$

is a function of two parameters s and u . The first is a scaling or dilation factor that controls the length of the wavelet; the latter is a location parameter that indicates where the wavelet is centered along the signal. The set of CWT wavelet coefficients, each representing the amplitude of the wavelet function at a particular position and for a particular wavelet scale, is obtained by projecting $x(t)$ onto the family of “wavelet daughters” $\psi_{(s,u)}$ obtained by scaling and translating the “mother wavelet” ψ by s and u , respectively.

¹ The assumption that the signal is homogeneous over time restricts the usefulness of Fourier methods to the analysis of stationary processes.

Let W_x and W_y be the continuous wavelet transform of the signals $x(\cdot)$ and $y(\cdot)$; their cross-wavelet power is given by $|W_{xy}|=|W_x W_y|$ and depicts the local covariance of two time series at each scale and frequency (see Hudgins et al. 1993). Being the product of two non-normalized wavelet spectra, the cross-wavelet can identify the significant cross-wavelet spectrum between two time series, although there is no significant correlation between them. The wavelet coherence is defined as the modulus of the wavelet cross spectrum normalized by the wavelet spectra of each signal,

$$R_{xy}^2 = \frac{|S(\lambda^{-1}W_{xy}(\lambda, \tau))|^2}{S(\lambda^{-1}|W_x(\lambda, \tau)|^2)S(\lambda^{-1}|W_y(\lambda, \tau)|^2)}$$

where S is a smoothing operator (see Torrence and Webster 1999). The squared wavelet coherence coefficient R_{xy}^2 can be considered a direct measure of the local correlation between two time series at each scale and it is analogous to the squared correlation coefficient in linear regression. Hence, it can be used to assess how the degree of association between two series changes across frequencies and over time. The outcome of the wavelet coherence analysis is in the form of a heatmap, which allows easily identifying low- and high-coherence power regions in the time–frequency plane, that is areas where the degree of association between two time series is weak or strong.

The general formulation for the continuous wavelet transform can be restricted to the definition of the discrete wavelet transform (DWT) by discretizing the parameters s and u . The idea is to select τ and s so that the information contained in the signal can be summarized by a minimum number of wavelet coefficients. In order to obtain an orthonormal basis, a transform of the scaling parameter, $s=s_0^j$, and the Nyquist sampling rule, $u=ks_0^jT$, are used. When the computation is done octave by octave, i.e., $\lambda_0=2$, we get the following equation for the “mother wavelet”:

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(\frac{t-2^j k}{j}\right)$$

This function represents a sequence of rescaleable functions at a scale of $\lambda=2^j$, $j=1, 2, \dots, J$, and with time index k , $k=1, 2, 3, \dots, N/2^j$. The wavelet transform coefficient of the projection of the observed function $f(t)$ for $i=1, 2, 3, \dots, N$, $N=2^J$ on the wavelet $\psi_{j,k}(t)$ is given by:

$$d_{j,k}(t) \approx \int \psi_{j,k}(t)f(t)dt$$

For a complete reconstruction of a signal $f(t)$, one requires a scaling function, $\phi(\cdot)$, that represents the smoothest components of the signal. While the wavelet coefficients represent weighted “differences” at each scale, the scaling coefficients represent averaging at each scale. One defines the scaling function, also known as the “father wavelet”, by.

$$\phi_{j,k}(t) = 2^{-J/2}\phi\left(\frac{t-2^j k}{2^j}\right)$$

and the scaling function coefficients vector is given by.

$$s_{J,k}(t) \approx \int \phi_{J,k}(t)f(t)dt$$

By construction, we have an orthonormal set of basis functions, whose detailed properties depend on the choices made for the functions, $\varphi(\cdot)$ and $\psi(\cdot)$ (see for example the references cited above as well as Daubechies (1992)). At each scale, the entire real line is approximated by a sequence of “non-overlapping” wavelets. The deconstruction of the function $f(t)$ is, therefore:

$$x(t) \approx \sum_k s_{Jk} \phi_{Jk}(t) + \sum_k d_{Jk} \psi_{Jk}(t) + \dots + \sum_k d_{jk} \psi_{jk}(t) + \dots + \sum_k d_{1k} \psi_{1k}(t)$$

Further, the approximation can be re-written in terms of collections of coefficients at given scales as:

$$x(t) \approx S_J + D_J + \dots + D_j + \dots + D_1$$

S_J contains the “smooth component” of the signal, and the D_j , $j=1, 2, \dots, J$, the detail signal components at ever increasing levels of detail. The number of observations at each scale is given by $N/2^j$, $j=1, 2, \dots, J$. S_J provides the large scale road map; D_1 shows the pot holes. The previous equation indicates what is termed the multiresolution decomposition (MRD), where a signal is decomposed into several components, each associated with a different frequency band, with a resolution matched to its scale. Specifically, using multi-resolution technique by which different frequencies are analyzed with different resolutions, the wavelet transform gives good time resolution and poor frequency resolution at high frequencies, and good frequency resolution and poor time resolution at low frequencies.

3 Empirical results

CMIP5 experimental framework consists of a standard set (several types) of model simulations specifically designed for model evaluation (Taylor et al., 2012). One sub-category of such long-term experiments, the “historical” runs, includes historical simulations of climate change and variability over the period 1850–2005 with purpose of evaluating how realistic the models are in simulating past observed climate change. To allow for systematic model inter-comparison, the historical run is subject to changing conditions reflecting observed changes in anthropogenic (greenhouse-gas and aerosol emissions) and natural (solar activity and volcanic activity) external forcing.

In this section, we compare the simulations generated in the historical run experiment, from 44 different CMIP5 models against an observational benchmark (HadCRUT4 data set) using the annual time sequences of the variable *tas*, near-surface air temperature anomalies ($^{\circ}\text{C}$ from 1961–1990 mean), with the anomalies computed relative to the average of the period 1961–1990. Annual data are from the KNMI (Royal Netherlands Meteorological Institute) Climate Explorer website (<https://climexp.knmi.nl/>).² The ensemble mean is used for CMIP5 model with more than one individual ensemble member.³

² Annual data are preferred to monthly data as they allow us to explore the low-frequency relationships between modeled and observed time series.

³ Although detrending is a critical issue, as the results may be sensitive to the removal of trends, we decide to avoid detrending. The main reason is related to the long-term nature of the historical (1850–2005) run experiment which is to evaluate model performance against observed climate change at decadal and multidecadal time scales. In addition, the wavelet transform, because of its time-scale localization property and the frequency-dependent windowing of its time scale decomposition, has the ability to handle complex non-stationary processes with multiscale structures and scale-dependent relationships such as climatic variables (Lin and Franzke, 2015).

3.1 Wavelet coherence analysis of CMIP5 climate models

Wavelet coherence maps can provide descriptive and diagnostic information suitable for the assessment of model performance, as well as for model inter-comparison. When the two analyzed time series are the output of model simulation and their corresponding observations, the color map of the wavelet coherence plot provides researchers with an overall perception about the performance of the model that is based upon an objective, not subjective, data visualization method. The visual detection of low-coherence power regions in wavelet coherence maps allows to identify regions in the time–frequency plane where models and measurements are significantly different, i.e., model/data disagreements. In particular, as wavelet decomposition captures the high frequency components (short-term features) at the finer time scales and the low frequency components (long term features) at the coarser time scales, the time–frequency localization of errors explicitly identifies periods when model simulations diverge from observed data and the associated time scale over which this occurs.

In what follows we use wavelet coherence analysis to compare the performances of different competing models in an explicit manner.⁴ Indeed, wavelet coherence plots provide useful insights for model intercomparison with respect to traditional measures which use global performance estimates to quantify (dis)agreements between simulated and observed time series. Wavelet coherence plots presented in Fig. 2 are calculated using the Morlet wavelet, a complex wavelet that produces complex transforms and thus can provide us with information on both amplitude and phase. We use the Morlet wavelet with $\omega_0=6$ (where ω_0 is dimensionless frequency) since this particular choice provides a good balance between time and frequency localization and also simplifies the interpretation of the wavelet analysis because the wavelet scale is inversely related to the frequency, frequency $\approx 1/\text{scale}$ (Grinsted et al. 2004). In each panel of Fig. 2, time is recorded on the horizontal axis and the vertical axis gives us the periods (and the corresponding scales of the wavelet transform). Since the CWT need to be computed discretely, the scale axis is partitioned into a set of scales, the largest value corresponding to 52 years.⁵ Reading across the graph at a given value for the wavelet scaling one sees how the power of the projection varies across the time domain at a given scale, while reading down the graph at a given point in time one sees how the power varies with the scaling of the wavelet (see Ramsey and Zhang 1995).

Since the magnitude of each squared coherence coefficient is indicated by the color scale, with the color code for power ranging from dark blue (low coherence) to yellow (high coherence), the warmer the color, the higher the coherence power between the two series at that location in the time–frequency plane. The statistical significance of all wavelet coherence spectra was evaluated using the cumulative area-wise test (Schulte 2016, Schulte 2019) to account for the simultaneous testing of multiple hypotheses (Maraun and Kurths, 2004; Maraun et al., 2007).⁶ Contours enclose regions of 5% cumulative areawise

⁴ The CWT analysis was performed using the Wavelet Matlab software package available at justinschulte.com.

⁵ Following Torrence and Compo (1998), the maximum scale value J is obtained as $J = \lfloor \log_2(N/2) \rfloor$ where ≤ 12 defines the number of intermediate scale values computed between two consecutive powers of two.

⁶ The arc-wise test should be preferred to the cumulative areawise test when testing for the presence of periodicities embedded in time series (Schulte, 2019). The cumulative areawise test developed by Schulte (2016) is well suited for the purpose of identifying model/data disagreement in the time–frequency plane through statistically significant coherence regions in local wavelet coherence plots.

Fig. 2 Wavelet coherence plots between several CMIP5 climate simulation models and HadCRUT4 (historical run over the period 1850–2005)

significance against the null hypothesis of an autoregressive process of first order (AR1).⁷ A black thin line marks the cone of influence which is the region where edge effects become significant at different scales. Areas of high coherence occurring outside the cone of influence should be interpreted with caution, as they result from a significant contribution of zero padding at the beginning and the end of the time series.

Notwithstanding the analysis has been performed using the ensemble mean of all 44 CMIP5 models, the wavelet coherence presented in Fig. 2 refers to 10 models only for reasons of space. Models are selected in order to put in evidence typical similarities and differences among models.⁸ The color maps of the wavelet coherence plots in Fig. 2 reveal that systematically recurrent areas of model-data disagreements are clearly evident for CMIP simulation models. From the distribution of the blue and yellow areas in the time–frequency plane, a consistent pattern emerges in all wavelet coherence plots. Almost all models are able to capture the variability of observed data at multidecadal time scales, between 16 and 32 years, until early 1900s. The longer-term variability of observed data is captured by most of the models for the whole sample, with this ability progressively increasing towards higher frequencies from the post-WWII period onwards. By contrast, short-term anomalies, allocated in high-frequency components of wavelet coherence spectra, are rarely captured by the simulations. This is not surprising, given the long-term nature of the historical run in the CMIP experiment.⁹

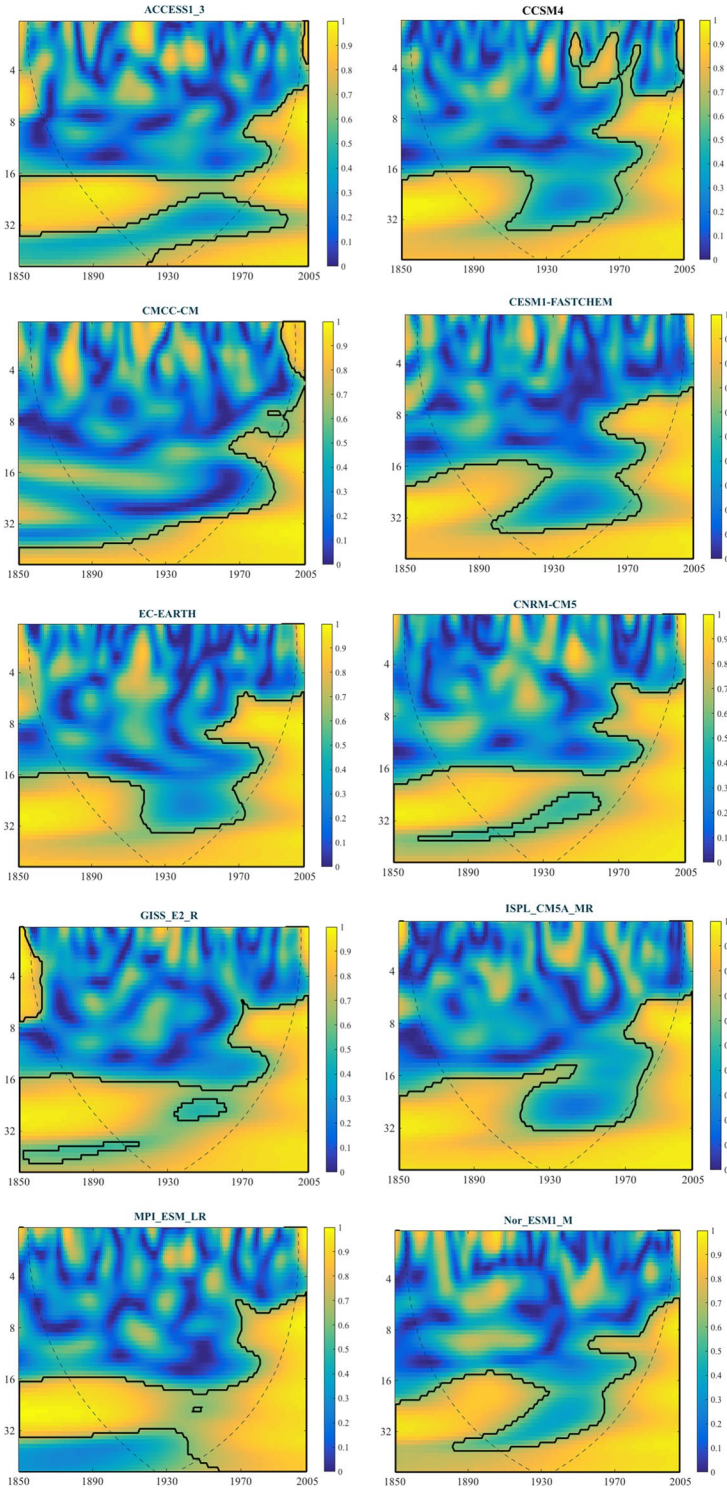
Beyond time scales corresponding to higher frequency components (periods smaller than 16 years), there is also evidence of model/data disagreements at longer time scales (periods greater than 32 years) until early 1900s, and from the early 1900s to late 1960s at scales between 16 and 32 years. By contrast, it is evident a good performance of all simulation models in capturing the anthropogenic-induced global warming trend from the 2nd half of the twentieth century. Indeed, model performance estimates are likely to be strongly dominated by the trending evolution of external forcings, especially GHG concentrations, as the main long-term driver of temperature. An interesting feature common to all simulation models is that their good performance, generally limited to longer frequencies, tends to spread towards higher frequencies from 1960 onwards, as evidenced by the high coherence region in the south-east corner of each panel. However, since the scale-elongated regions of statistical significance are mostly concentrated in the cone of influence region, we are cautious in making conjectures on the increasing ability of CMIP5 simulation models to reproduce the climate variability observed in recent decades on all time scales.

An interesting feature of wavelet coherence maps is that they allow to identify both similarities and differences among simulations from different climate models. The estimated wavelet coherence plots presented in Fig. 3 refer to two CMIP5 global climate models, *MIROC-ESM-CHEM* (left) and *inmcm4* (right), exhibiting a common pattern that is distinct from that of the other models. Both wavelet coherence plots show that the long-term

⁷ A red noise AR(1) process can be considered an acceptable noise background model for the global surface temperature (e.g., Grinsted et al. 2004, Schulte et al. 2018), as many climatic variables have higher power at longer periods.

⁸ The wavelet coherence plots for all 44 models are available on request by the author.

⁹ Since CMIP5 climate model experiment is designed to explore the ability of simulation models to predict temperature changed on decadal and multidecadal time scales, global climate models are not expected to emulate the short-term pattern of observed data (see Gong et al. 2018).



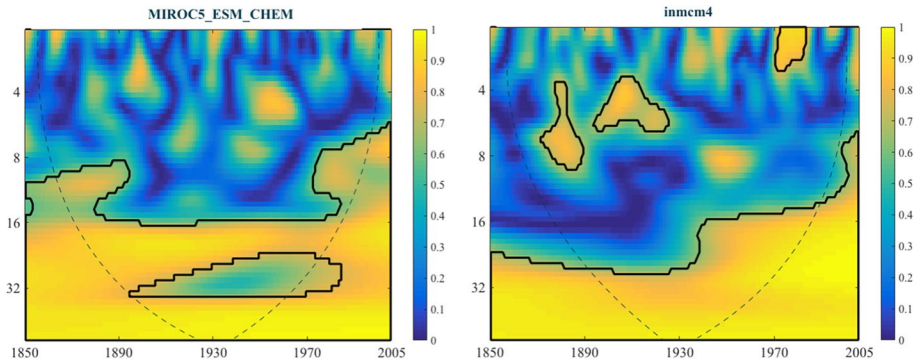


Fig. 3 Wavelet coherence plots between MIROC-ESM-CHEM (left) and inmcm4 (right) simulations and HadCRUT4 (historical run over the period 1850–2005)

multidecadal pattern in the global temperature series is picked up well by the simulations. This frequency-dependent relationships is captured by the *MIROC-ESM-CHEM* simulation model at scales greater than 16 years throughout the sample. Otherwise, for the *inmcm4* simulation model, the long-term pattern is captured at scales greater than 32 years in the pre-WWII period, and at scales greater than 16 years in the post-WWII period.

In sum, we show that wavelet coherence plots may provide clear visual evidence on the similarities and differences among simulations of the CMPI5 climate models. In conformity with the characteristic of CMPI5 project to predict climate over time scales of decades, the multiscale analysis indicates that all global climate simulation models perform poorly at finer time scales and better at coarser time scales, that is at scales corresponding to periods greater than 16 years. Although visualization gives us useful information about the model performance at the user level, these results are essentially descriptive in nature and do not provide any quantitative information about the overall model performance. In order to get quantitative measures of performance assessment, we need to move from the CWT to the DWT.

3.2 Wavelet decomposed components for climate model inter-comparison

CWT tools like the wavelet coherence provide information that allow a systematic model inter-comparison, but that are essentially descriptive and exploratory in nature. By contrast, the application of the DWT provides a quantitative measures of model performance at each time scale, thus making possible a direct comparison with standard methods. After decomposing simulated and observed time series into their time scale components, the analysis of model simulation results and observations may be carried out at any time scales separately.

The main advantage of the DWT for the proposed approach is that, differently from the CWT, wavelet-based multiscale performance measures may be obtained by decomposing simulated and observed time series into subsignals. This allows us to quantify model performance as a function of frequency and to localize model-data disagreement in time, since the different time scale components obtained by the DWT decompositions are expected to display similar patterns.

The DWT, using only a limited number of translated and dilated versions of the mother wavelet to decompose the original signal, provides a more parsimonious representation

than the CWT, but has several drawbacks. Therefore, in practical applications a compromise between the CWT, with continuous variations in scale, and the DWT is generally used: the maximal overlap DWT (MODWT). The MODWT is a non-orthogonal variant of the classical discrete wavelet transform that, unlike the DWT, is (i) translation invariant, as shifts in the signal do not change the pattern of coefficients, (ii) can be applied to data sets of length not divisible by 2^J , and (iii) returns at each scale a number of coefficients equal to the length of the original series.

The principle of the model-data comparison on multiple time scales implies that both observed and modeled time series are first decomposed into subsignals corresponding to different frequency ranges. After that qualitative and/or quantitative model-data comparisons can be carried out on the corresponding pairs of subsignals. In this way, the degree of model/data disagreement is still estimated in the time domain, although measured within a well defined frequency range.

The simulated and observed time series are first decomposed into $J=4$ decomposition levels using the MODWT by applying the Daubechies least asymmetric (LA) wavelet filter of length $L=8$ based on eight nonzero coefficients (Daubechies, 1992) with reflecting boundary conditions. Four wavelet details vectors D_4 , D_3 , D_2 , and D_1 and one wavelet smooth vector, S_4 , each associated with a specific time period 2^{j-1} , are obtained. In particular, since we use annual data, the first detail level D_1 captures oscillations between 2 and 4 years, while details D_2 , D_3 , and D_4 capture oscillations with a period of 4–8, 8–16, and 16–32 years, respectively. Finally, the smooth component S_4 captures the low-frequency oscillations with a period longer than 32 years.

Table 1 reports the values of a standard performance assessment measure, the root mean square error (RMSE), for the CMPI5 models that display a statistics lower than 0.60 calculated on raw data. Two findings are noteworthy. First, similar aggregate performance may be the result of quite different behavior at different time scales, e.g., *GISS-E2-R* and *GISS-E2-R-CC*. For instance, when we consider climate variability on longer time scales only, i.e., S_4 and D_4 , the *GISS-E2-R-CC* model provides better performance at the longest time scale S_4 , the other, the *GISS-E2-R* model, at the scale D_4 . Interestingly, when the longer time scales S_4 and D_4 are considered jointly, the performance of the *MIROC-ESM-CHEM* model is superior to that of all other models.

Performance measures like scale-based RMSEs lead to model-data comparison for each defined frequency range. However, as shown by wavelet coherence maps in Fig. 3, model/data disagreements are both frequency- and time-dependent. By plotting pair of time scale oscillatory components of observed and simulated series, we can check whether the model is deemed to be able to capture all the features of the observed data at each frequency range and over time.

We examine model/data disagreements at different frequency ranges for several models well performing according to wavelet coherence maps and scale-based RMS errors criteria, that is CNRM-CM5, MIROC-ESM-CHEM and inmcm4. In Fig. 4 we plot the observed (blue line) and the simulated (red line) series at longer time scales. Two panels are displayed for each model: the left panel contains the long-term smooth components S_4 , the right panel the D_4 wavelet component, that is the highest level deviations from their corresponding smooth component.

The historical long-term evolution of the global surface temperature reveals a complex nonlinear pattern. Following the trendless pattern of the global surface temperature in the 1850–1910 period, the warming trend has proceeded in a stepwise fashion, with the general increase in global average temperature interrupted by a short “cooling” period, as it is generally identified in the climate change literature.

Table 1 Scale-based RMSEs for CMPI5 simulation models with values <0.6 at the raw series level

Model	Institution	Raw series	S ₄ > 32 yrs	D ₄ 16–32 yrs	D ₃ 8–16 yrs	D ₂ 4–8 yrs	Average (S ₄ , D ₄)
CESM1-CAM5	Community Earth System Model Contributors, US	0.5335	0.4070	0.7286	0.9736	0.9863	0.5678
CNRM-CM5	Meteo-France/Cntr Natl Recher Meteor, France	0.4903	0.3107	0.7475	1.0671	1.0491	0.5291
CSIRO-MK3.0	Commonwealth Sci& IndustRes Org, Australia	0.5783	0.4746	0.7380	0.9746	0.9755	0.6063
EC-EARTH	European Earth System Model, Ireland	0.5429	0.4175	0.7230	1.0578	1.0155	0.5702
F-GOALS	Institute of Atmospheric Physics and Tsinghua University, China	0.5641	0.4237	0.8038	1.0938	1.1259	0.6137
GISS-E2-H	NASA/Goddard Institute for Space Studies, USA	0.5325	0.4025	0.7192	1.0616	1.0068	0.5608
GISS-E2-R	NASA/Goddard Institute for Space Studies, USA	0.5062	0.3520	0.7069	1.0913	1.0322	0.5294
GISS-E2-R-CC	NASA Goddard Institute for Space Studies, USA	0.5095	0.2895	0.9897	1.1538	1.1026	0.6396
inmcm4	Institute for Numerical Mathematics, Russia	0.5573	0.3455	0.7774	1.1259	1.2715	0.5614
MIROC5	JAMSTEC, AORI and NIES, Japan	0.5367	0.3437	0.9428	1.2276	1.1328	0.6432
MIROC-ESM	JAMSTEC, AORI and NIES, Japan	0.5369	0.3561	0.6814	1.1807	1.0732	0.5187
MIROC-ESM-CHEM	JAMSTEC, AORI and NIES, Japan	0.5505	0.3281	0.6766	1.3195	1.2328	0.5023
MPI-ESM-MR	Max Planck Institute for Meteorology, Germany	0.5897	0.4786	0.7837	1.1445	0.9514	0.6311
Nor-ESM1-M	Norwegian Climate Center, Norway	0.5564	0.3892	0.9164	1.0168	1.0711	0.6528

Note: The three lowest values of the RMSEs at aggregate and time scale level are reported in bold.

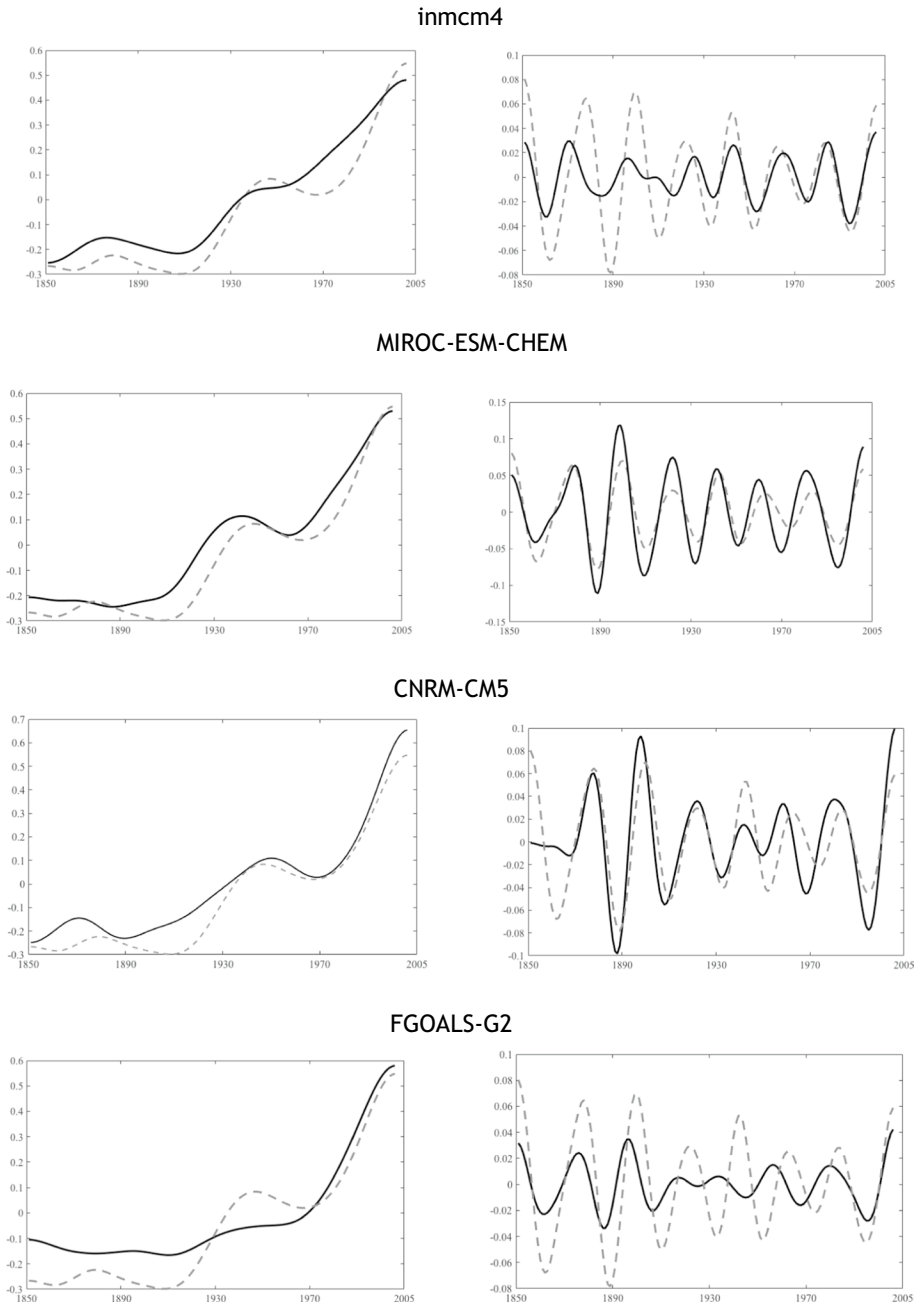


Fig. 4 Plots of wavelet decompositions of inmcm4, MIROC-ESM-CHEM, CNRM-CM5, and FGOALS-G2 model simulations (black solid lines) against HadCRUT4 global temperature (gray dotted lines) at scale levels S_4 (left column) and D_4 (right column) for the period 1850–2005

Within the general ability of all models to reproduce the trend segmented pattern of the process, some disagreements between model simulations and observations are anyway evident. Indeed, while the observed oscillatory trendless pattern is well captured by *inmcm4*, the nonlinear upward trending pattern is well captured by *CNRM-CM5*, especially from 1960 onwards, and, to a lesser extent, also by *MIROC-ESM-CHEM*. By contrast, *inmcm4* experiences some difficulties in reproducing the nonlinearity of the long-term trending pattern, as it fails to capture the slightly decreasing warming pattern in the 1945–1970 “cooling” period. At multidecadal time scales, D_4 , *MIROC-ESM-CHEM* simulation output is very close to the corresponding oscillatory pattern of the observed temperature series until late 1950s, whereas *inmcm4* is the best performing model from 1960 onwards.

In sum, the very good performance of all models in capturing the nonlinear pattern of global warming from 1970 onwards suggests that current climate models are able to simulate the dominant features at longer term scales, the most important being the response to external forcing, particularly anthropogenic forcing such as GHG. Nonetheless, we show that the multiscale decomposition of observed and simulated series may be very useful for the comparison of models which show similar aggregate performances. When these performances differ considerably at different decomposition levels, model assessment can be further refined by the explicit location of model-data disagreements across frequency and over time.

4 Conclusions

This paper evaluates the performance of CMIP5 climate model simulations using (continuous) wavelet coherence maps and scale-based (discrete) goodness of fit relative measures, such as the root mean square error (RMSE). Exploratory diagnostic graphical analysis suggests the existence of recurrent model-data (dis)agreements at different scales of variability, particularly high coherence at lower frequencies (and vice versa). This is consistent with the aim of the CMIP5 model experiment which is designed to explore the ability of models to predict climate on decadal and multidecadal time scales. Moreover, the RMSE comparison at different time scales allows to quantitatively evaluate the performance of alternative models, whereas the plots of wavelet decompositions of CMIP5 climate model simulations and temperature series, by explicitly locating model-data disagreements in frequency and time, can provide descriptive and diagnostic information for model inter-comparison and the refinement of model development.

In terms of model/data comparison, we find that some models perform better than others at certain time scales and/or periods, but no individual model clearly emerges as “the best” overall model. Moreover, models displaying “similar” aggregate performance are shown to display quite different performances when analyzed at different time scales. In particular, we show that wavelet-based methods can increase the resolution of the assessment of the performance of competing models with respect to conventional evaluation strategies because of their ability to localize the timing of model–data disagreement on a scale-by-scale basis.

Since global climate simulation models encode the current understanding of the climate system with its interactions between processes (physical, biological, and chemical) and between components (atmosphere, ocean, sea ice, etc.), the information regarding the scale and time at which model simulations do not match observations can help in identifying the essential problems of climate models. For instance, time–frequency localized model/data

discrepancies can also be useful for model development by including or excluding certain processes or components of the climate system, the relative importance of which varies with the time scale of interest.

Acknowledgements We thank three anonymous referees for their careful evaluation of the paper and their extremely useful suggestion. We also want to thank the participants at the Econometric models of climate change conference (EMCC IV) held at the University of Milano Bicocca, 29-30 August 2019, for their suggestions and comments.

Funding Partial financial support was received from Polytechnic University of Marche.

Data availability Data are available on request by the author.

Declarations

Conflict of interest The author declares no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Braverman A, Chatterjee S, Heyman M, Cressie N (2017) Probabilistic evaluation of competing climate models. *Advances in Statistical Climatology, Meteorology and Oceanography* 3:93–105
- Briggs WM, Levine RA (1997) Wavelets and field forecast verification. *Mon Weather Rev* 125(6):1329–1341
- Chatterjee S (2019) The scale enhanced wild bootstrap method for evaluating climate models using wavelets. *Statist Probab Lett* 144:69–73
- Daubechies I. (1992) Ten lectures on wavelets. CBSMNSF regional conference series in applied mathematics. SIAM, Philadelphia.
- Dietze M.C. and Coauthors (2011), Characterizing the performance of ecosystem models across time scales: a spectral analysis of the North American Carbon Program site-level synthesis, *Journal of Geophysical Research: Biogeosciences*, 116(G4) <https://doi.org/10.1029/2011jg001661>
- Gallegati M (2018) A systematic wavelet based exploratory analysis of climatic variables. *Clim Change* 148(1):325–338
- Gencay R, Selcuk F, Whitcher B (2003) Systemic risk and timescales. *Quantitative Finance* 3(2):108–116
- Grinsted A, Moore JC, Jevrejeva S (2004) Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process Geophys* 11:561–566
- Gong K, Braverman A, Chatterjee S (2018) On a technique for evaluating the quality of earth system models. In *Proceedings of the 8th International Workshop on Climate Informatics: CI 2018*, 93–96
- Gupta HV, Wagener T, Liu YQ (2008) Reconciling theory with observations: towards a diagnostic approach to model evaluation. *Hydrol Processes* 22(18):3802–3813
- Hudgins L, Friehe CA, Mayer ME (1993) Wavelet transforms and atmospheric turbulence. *Phys Rev Lett* 71(20):3279–3282
- Jevrejeva J, Moore C, Grinsted A (2003) Influence of the Arctic Oscillation and El Nino-Southern Oscillation (ENSO) on ice conditions in the Baltic Sea: the wavelet approach. *J Geophys Res* 108(D21):46–77
- Kumar P, Foufoula Georgiou E (1997) Wavelet analysis for geophysical applications. *Rev Geophys*:385–412
- Lau KM, Weng H (1995) Climate signal detection using wavelet transform: how to make a time series sing. *Bull Am Meteorol Soc*:2391–2402

- Lin Y, Franzke CLE (2015) Scale dependency of the global mean surface temperature trend and its implication for the recent hiatus of global warming. *Sci Rep* 5 <https://doi.org/10.1038/srep12971>
- Mahecha et al. (2010) Comparing observations and process-based simulation of biosphere-atmosphere exchanges on multiple timescales. *J Geophys Res* 115
- Maraun D, Kurths J (2004) Cross wavelet analysis: significance testing and pitfalls. *Nonlinear Process Geophys* 11:505–514
- Maraun D, Kurths J, Holschneider M (2007) Nonstationary Gaussian processes in wavelet domain: synthesis, estimation, and significance testing. *Phys Rev E* 75
- Meehl G.A. and Coauthors (2007) Global climate projections. *Climate Change 2007: The Physical Science Basis*, Cambridge University Press
- Meehl GA, Hu A, Tebaldi C (2010) Decadal prediction in the Pacific region. *Am Meteorol Soc* 2959–2973
- Mitchell JM Jr (1976) An overview of climatic variability and its causal mechanism. *Quat Res* 6(4):481–493
- Park J, Mann ME (2000) Interannual temperature events and shifts in global temperature: a multiwavelet correlation approach. *Earth Interact* 4(1)
- Percival D.B. (2008) Analysis of geophysical time series using discrete wavelet transforms: an overview. *Nonlinear Time Series Analysis in the Geosciences* 61–79
- Ramsey JB, Zhang Z (1995) The analysis of foreign exchange data using waveform dictionaries. *J Empir Financ* 4:341–372
- Reusser D, Loukopoulos P, Stauffacher M, Scholz R (2008) Classifying railway stations for sustainable transitions – balancing node and place functions. *J Transp Geogr* 16(3):191–202
- Schulte JA (2016) Cumulative areawise testing in wavelet analysis: theoretical developments and application to Indian rainfall. *Nonlin. Processes Geophys* 26:91–108
- Schulte JA, Georgas N, Saba V, Howell P (2018) North pacific influences on long island sound temperature variability. *J Clim* 31(7):2745–2769
- Schulte JA (2019) Statistical hypothesis testing in wavelet analysis and its application to geophysical time series. *Nonlin Processes Geophys* 26:91–108
- Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 79(1):61–67
- Torrence C, Webster PJ (1999) The annual cycle of persistence in the El Niño–Southern Oscillation. *Q J R Meteorol Soc* 124(550):1985–2004
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Biogeosci* 106(D7):7183–7192
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Am Meteorol Soc* 485–98
- Vargas R, Detto M, Baldocchi DD, Allen MF (2010) Multi-scale analysis of temporal variability of soil CO₂ production as influenced by weather and vegetation. *Glob Change Biol* 16:1589–1605
- Vargas R, Coauthors (2013) Drought influences the accuracy of simulated ecosystem fluxes: a model–data meta-analysis for Mediterranean oak woodlands. *Ecosystems* 16:749–764
- Williams M et al (2009) Improving land surface models with FLUXNET data. *Biogeosciences* 6:1341–1359
- Wang W, Hu S, Li Y (2011) Wavelet transform method for synthetic generation of daily streamflow. *Water Resour Management* 25(1):41–57

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.