Available online at www.sciencedirect.com

SCIENCE DIRECT®

Global and Planetary Change 37 (2003) 19–32

GLOBAL AND PLANETARY CHANGE

www.elsevier.com/locate/gloplacha

Evaluation of Northern Hemisphere natural climate variability in multiple temperature reconstructions and global climate model simulations

J.L. Bell[a,*], L.C. Sloan[a], J. Revenaugh[a], P.B. Duffy[b]

[a] Department of Earth Sciences, University of California, Santa Cruz, CA 95064, USA

[b] Climate and Carbon Cycle Modeling Group, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

Received 1 April 2001; accepted 6 July 2002

Abstract

The detection of anthropogenic climate change in observations and the validation of climate models both rely on understanding natural climate variability. To evaluate internal climate variability, we apply spectral analysis to time series of surface air temperature (SAT) from nine coupled general circulation model (GCM) simulations, three recent global paleotemperature reconstructions, and Northern Hemisphere (NH) instrumental records. Our comparison is focused on the NH due to the greater spatial and temporal coverage and validation of the available NH temperature reconstructions. The paleotemperature reconstructions capture the general magnitude of NH climate variability, but not the precise variance and specific spatial, temporal, or periodic signals demonstrated in the instrumental record. The models achieved varying degrees of success for each measure of variability analyzed, with none of the models consistently capturing the appropriate variability. In general, the models performed best in the analysis of combined mean annual land and marine variability.

Keywords: climate; climate variability; climate model; spectral analysis

1. Introduction

Considerable effort has been made in recent years to compare general circulation model (GCM) results to instrumental observations and paleotemperature reconstructions (e.g., Barnett et al., 1996; Jones et al., 1998; Bell et al., 2000; etc.). The purposes of these studies have been to: (1) evaluate the performance of climate models, or (2) to determine whether or not humans have had a discernable influence on climate in the 20th century (Santer et al., 1995; Barnett et al., 2001; Levitus et al., 2001). The first goal is met by assessing whether or not the models capture significant features of the instrumental and proxy time series of temperature and by comparing the variability in the temperature records to the internal climate variability of the models.

Assessing how well a particular model simulates natural climate variability relies on comparison with instrumental and proxy records. Such comparisons are problematic for several reasons. Inherent in any

* Corresponding author. Tel.: +1-831-459-3504; fax: +1-831-459-3074.

E-mail address: jbell@es.ucsc.edu (J.L. Bell).

0921-8181/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0921-8181(02)00189-3

instrumental record is the possible anthropogenic influence on climate, which cannot be easily removed from the record. Also, instrumental data are practically nonexistent in some areas and spatial coverage decreases earlier in the record (Fig. 1). For example, high-latitude regions have very little instrumental coverage. Furthermore, instrumental records are limited to typically the last 150 years, rendering any longer-term variability analysis impossible. Proxy temperature data have the advantage of spanning much longer time periods, often several centuries. As a disadvantage, proxies of past temperatures are also limited in spatial extent and are not direct measurements of temperature, but are instead measurements of phenomena such as tree ring growth, variations in sediment layer thickness, or geochemical (typically δ^{18}O) changes that can be shown to vary as a function of climate. That these proxies all have systems that "filter" the original temperature signal introduces inherent uncertainties into the resulting temperature interpretations. Furthermore, determining low-frequency variability from proxy records can be difficult due to changes in the recorded climate signal with increasing length of the proxy record. For example, tree rings have age-related growth trends, and ice layers thin with depth. These trends must be removed from the proxy record and, in doing so, low-frequency climate data is inevitably removed as well. Despite these uncertainties, the length of the available proxy records can allow for more robust low-frequency natural variability estimates than is possible using instrumental data.

2. Proxy temperature data

Recent paleoclimate reconstructions typically include several different types of proxy indicators such as tree ring widths and/or densities, δ^{18}O measurements from ice cores or corals, and varved sediment thicknesses. Individual proxy temperature indicators represent particular temporal distributions and representative regions. Each proxy is calibrated (typically to observational data for an overlapping

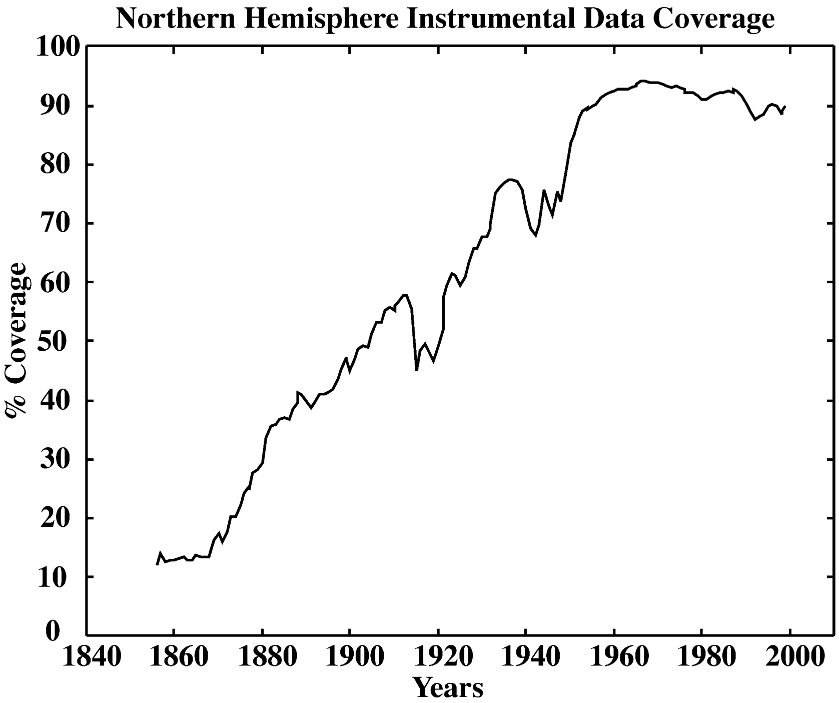

Fig. 1. Percent area of NH instrumental temperature records with respect to time (Jones, 1994; Parker et al., 1995; Jones et al., 1999).

period of time), and then combined with other proxies to create a composite time series. The proxy records used in this study are composite records, which represent greater spatial and temporal distribution than is possible with individual site proxies. It is important to note that many temperature reconstructions are composed of less than 30 individual records of varying length. For a summary of the spatial limitations and time scale dependence of common proxies, see Jones et al. (1998).

We analyze three proxy temperature records from Jones et al. (1998), Mann et al. (1999), and Briffa et al. (2001) (hereafter referred to as J98, M99, and B01, respectively). Each record has specific strengths and weaknesses. For example, J98 is a multiproxy record composed of a limited number of proxies (10), but all are of greater than 350 years in duration. J98 is composed of one NH instrumental and one historical temperature record and eight proxy temperature records derived from tree ring density and width, ice core melt layers, and oxygen isotopic values. This reconstruction was originally presented as a summer record (June–August), but it has been rescaled to represent NH growing season temperatures (April–September) from AD 1000 to the present. The inclusion of several different types of proxies should limit any weaknesses or biases associated with individual proxy types.

The B01 record consists of 387 tree ring density chronologies representing NH growing season temperatures. These tree ring chronologies span the last 600 years and are processed using the Age-Band Decomposition (ABD) technique. ABD is a relatively new and untested calibration technique designed to preserve long time scale variability, but it may not accurately represent high-frequency variability (Briffa et al., 2001). The large number of tree ring chronologies used in B01 should reduce spatial biases and growth-related trends, but the use of only one type of proxy makes the entire record susceptible to any and all biases associated with tree rings.

The M99 data is a multiproxy network of widely distributed instrumental, historical, and proxy (tree ring, coral and ice core δ^{18}O) records representing NH mean annual temperatures. Prior to AD 1400, the data are composed of 12 NH proxy indicators dating back to AD 1000. After AD 1400, the multiproxy network is composed of at least 22 indicators, increasing to 112 indicators closer to the present. The multiproxy approach should limit the effects of any weaknesses or biases of individual proxy types and locations (Mann et al., 1998).

3. Model simulations

We examine nine coupled GCM control runs that are intended to simulate preindustrial climate (Table 1). Each simulation employs constant external forcing (e.g., constant preindustrial greenhouse gas concentrations and present-day solar constant), with no solar variability or volcanic aerosol emissions included. We analyze only simulations of multi-century length in

Table 1
Model properties

Model	Length (years)	Ocean resolution	Atmosphere resolution	Land surface scheme	Flux correction
ECHAM1	960	(4.0 × 4.0) L11	T21 (5.6 × 5.6) L19	Modified bucket	a, b, c
ECHAM3	1000	(4.0 × 4.0) L11	T21 (5.6 × 5.6) L19	Modified bucket	a, b, c
ECHAM4	240	(2.8 × 2.8) L11	T42 (2.8 × 2.8) L19	Modified bucket	a, b
GFDL	1000	(4.5 × 3.7) L12	R15 (4.5 × 7.5) L9	Bucket	a, b
GFDL 14K	14000	(4.5 × 3.7) L12	R15 (4.5 × 7.5) L9	Bucket	a, b
GFDL R30	500	(1.875 × 2.25) L18	R30 (2.25 × 3.75) L14	Bucket	a, b
NCAR CSM	290	(2.0 × 2.4) L45	T42 (2.8 × 2.8) L18	Physical	None
DOE PCM	300	(0.67 × 0.67) L32	T42 (2.8 × 2.8) L18	Physical	None
HadCM2	1085	(2.5 × 3.75) L20	(2.5 × 3.75) L19	Modified bucket	a, b

Properties of the GCMs analyzed in this study. Ocean and atmosphere resolutions are given as grid cell size in degrees followed by the total number of vertical levels. The land surface schemes are described in Bell et al. (2000). The flux corrections are: a—heat, b—water or c—momentum. References available from http://www-pcmdi.llnl.gov/cmip/Table.htm.

order to address climate variability on interannual to centennial time scales. For two of the models (ECHAM3 and GFDL 14K), only annual mean results were provided; therefore, these models are omitted from some of our analyses.

4. Methods

4.1. Variability tests

We analyze variance, significant periodic energy and the shape of the power spectrum for temperatures in each proxy and model time series. Each analysis involves a two-step approach. First, we compare the proxy data with instrumental temperature records (Jones, 1994; Parker et al., 1995; Jones et al., 1999; hereafter referred to as observations) over a validation period of 1881–1980. Second, we compare the proxy and observational data to the model results.

The observations are in the form of a 5° latitude by 5° longitude gridded data set. The gridded data allows us to subject the observations to the same space and time domain analyses as the model results. The purpose of validating the proxy data with the observations is to evaluate how well the proxies capture the major features of the temperature record. Several factors may introduce error into the validation procedure. These factors include, but are not limited to, the incomplete spatial coverage of the observations and proxy data, anthropogenic influences on climate and the direct effect of changing $p\mathrm{CO}_2$ on tree growth (Briffa et al., 1998), different methods and periods of proxy calibration, and natural "filtering" of the climate signal recorded by the proxies. In addition, the relatively short validation period does not capture any of the effects of long-term climate variability that we want to assess in the proxy vs. model comparison.

The proxy data are compared to observations and model results on four different spatial and temporal scales: (1) mean annual land and marine; (2) mean annual land-only; (3) growing season land and marine; and (4) growing season land-only (hereafter referred to as MALM, MALO, GSLM, and GSLO, respectively). In the case of the two annually resolved models (ECHAM3 and GFDL 14K), only the two mean annual time series were created, and these models are used only for mean annual comparisons. As described in the Proxy Temperature Data section, the proxy data fall into different spatial and temporal categories. With the exception of J98, which has been rescaled to represent growing season temperatures, we present the proxy records as their authors originally did with respect to what time period and spatial distribution each record represents. Since there may be some disagreement within the proxy data community regarding whether these spatial and temporal distributions are correct, we include all three proxy records for equal comparison with the model simulations in each of the four domains above. The proxies are annually resolved and, therefore, are well suited for comparisons on annual to centennial time scales.

To avoid incorporating any anthropogenic influences on climate, the proxy data were truncated at AD 1850 for comparison with the model results. All model results, observations, and proxy data were detrended using a quadratic least squares fit. The quadratic detrend is more effective than the standard linear detrend at removing nonperiodic components of the time series, with periodicity defined as integer fractions of the total length of the time series.

After detrending each time series, the variance was calculated and the power spectrum was estimated using the multi-taper method (MTM; Thomson, 1982). MTM was used instead of the more commonly used Blackman–Tukey method (Blackman and Tukey, 1958) because of the reduction in estimator variance available with MTM (Thomson, 1982; Percival and Walden, 1993). (For a discussion of the advantages and disadvantages of various spectral techniques and applications, see Ghil et al., submitted for publication). We use spectral analysis to compare periodic energy between different time series and to estimate the shape of the spectrum. We are interested in periodic elements of the proxy and observational spectra because they are generally indicators of modes of climate variability (e.g., El Niño/Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO)).

Time series of climatic variables often have both memory and random processes associated with them. For example, the statistics of detrended temperature at discrete time intervals can be modeled as the product of the temperature at the previous interval and a random perturbation. This type of process, known as an autoregressive process, or AR(1) process, can be represented by the following equation (Chatfield, 1996):

$$X_t = \alpha X_{t-1} + Z_t$$

where αX_{t-1} represents the memory of the process, and the Z_t are independent, normally distributed random variables. The α term is known as the lag one autocorrelation coefficient; it determines the shape of the power spectrum. For $\alpha>0$, power is concentrated at low frequencies, producing a red spectrum; if $\alpha = 0$, the spectrum is white or uncorrelated. Hasselmann (1976) introduced the red noise process in terms of climate as a description of integration of short time scale forcing by a slow response system (Griffies and Bryan, 1997), e.g., the integration of atmospheric forcing by the oceans. Additional sources of low-frequency (long period) variance (e.g., the Pacific Decadal Oscillation (PDO) or the NAO; Appenzeller et al., 1998; Black et al., 1999) will be added to the red noise process. Thus, a time series not exhibiting energy above the median red noise at the zero frequency is thought to be underrepresenting the true climatic variance (Mann and Lees, 1996; Mann et al., 1998).

4.2. Sensitivity tests

In a study by Bell et al. (2000), model results were compared to observations. In that study, the authors accounted for the incomplete spatial coverage of the observations by removing model results, where there was no observational data. We have not done this in the present study. Instead, we have performed sensitivity tests of the effects of missing data and short record lengths on variance and on the shape of the power spectrum. We calculated the variance over all nonoverlapping periods of 100 years and found the mean 100-year variance (the length of the observational record) for each model simulation. We chose to use the HadCM2 model for the sensitivity tests because it displayed average 100-year variance with a relatively large sample size (10 nonoverlapping 100-year segments). The process was repeated after masking the model results with the missing observational data from 1881 to 1980, resulting in average 100-year variance values from identical time periods for the masked and unmasked model data. We used the *F*-test to determine that the masked variance is significantly greater than the unmasked variance, implying that by not masking the model results, in our current study, we are making conservative comparisons of simulated and observed variances. To test the effects of varying spatial coverage on the shape of the power spectrum, we calculated the mean lag one autocorrelation coefficient (α) in a similar manner to above. The mean lag one autocorrelation coefficient is greater when using the masked results. This implies that the observational α value will be artificially high, indicating inflated amounts of low-frequency energy and a more red spectrum than would be calculated if the observational data coverage were spatially and temporally complete.

5. Results

5.1. Variance

5.1.1. Validation

Detrended time series of each proxy and the observational record for each time–space domain (MALM, MALO, GSLM, and GSLO) are shown for the validation period, 1881–1980 (Fig. 2). Table 2 shows the proxy variance for three separate time periods, and Table 3 shows the variance of the models and observations for the different time–space domains. We use the *F*-test to test the null hypothesis that two population variances are equal. If the *F*-test result is $<5\%$, we reject the null hypothesis and accept the alternative hypothesis that the variances of the two populations are different. Table 2 also shows the *F*-test results for the comparisons of proxy and observational variances.

Based on the above standard, we reject the null hypothesis in the comparisons of B01 variance (Table 2) to observational GSLM, GSLO, and MALM variances (Table 3). The B01 variance is greater than the observational GSLM, GSLO, and MALM variances, but not significantly different from the observational MALO variance. For J98 and M99, we accept the null hypothesis in all comparisons with observations; the

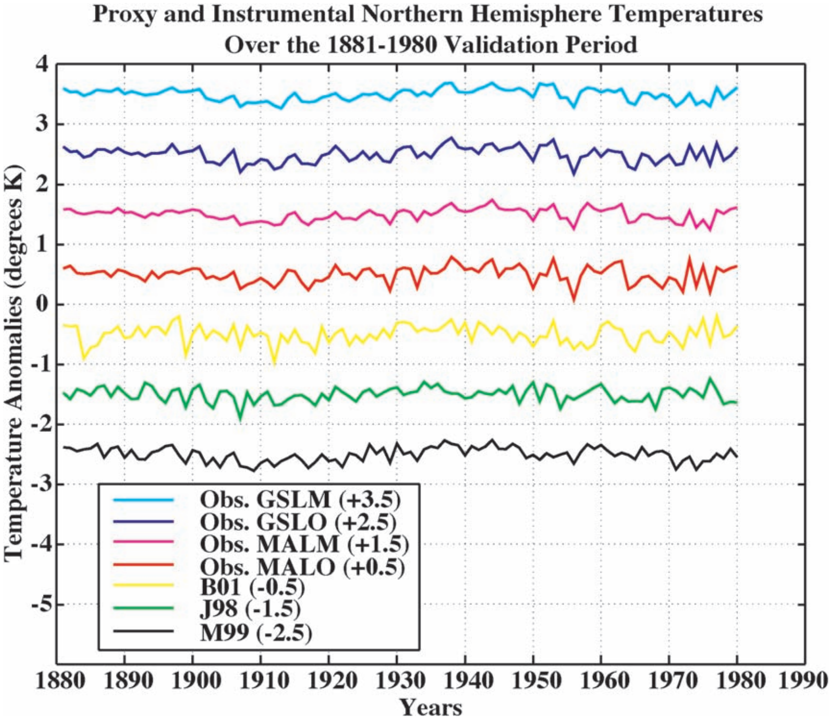

Fig. 2. Time series of the proxies and observational records (MALM, MALO, GSLM, and GSLO) are shown for the validation period, 1881–1980. Each time series has been detrended and is displayed in Fig. 3. The variances of the proxy time series appear in Table 2, and the variances of the observed time series appear in Table 3.

variances of the populations do not differ significantly.

5.1.2. Growing season

For the NH GSLO, total variance is greatest in the HadCM2, GFDL, and GFDL R30 models, and smallest for the ECHAM1 model (Table 3). The same is true when the model results are sampled over combined land and marine domains (GSLM), but the overall amplitude of the variance is less (Fig. 3). This is consistent with earlier findings that both observed and modeled land-only variability is greater

Table 2
Proxy variance (°C^2) and F-test results (%)

Proxy	Proxy variance			F-test			
	1000–1850	1402–1850	1881–1980	GSLM (%)	GSLO (%)	MALM (%)	MALO (%)
B01	NA	0.037	0.025	0	2	0	19
J98	0.023	0.021	0.014	11	64	15	12
M99	0.012	0.011	0.014	8	73	12	15

Proxy variances for three different time intervals (1000–1850, 1402–1850, and 1881–1980) and results of the F-test are shown. The F-test null hypothesis (equal variances) is rejected for values < 5%. Observational variance values are in Table 3.

Table 3
Variance of models and observations (°C^2)

	GSLM	GSLO	GSLO/GSLM	MALM	MALO	MALO/MALM
Observations	0.010	0.015	1.500	0.010	0.019	1.900
ECHAM1	0.006	0.018	2.797	0.006	0.016	2.574
ECHAM3	NA	NA	NA	0.010	0.024	2.465
ECHAM4	0.010	0.020	2.009	0.012	0.028	2.347
GFDL	0.021	0.053	2.523	0.017	0.038	2.266
GFDL R30	0.032	0.060	1.875	0.032	0.049	1.550
GFDL 14K	NA	NA	NA	0.017	0.037	2.228
NCAR CSM	0.017	0.033	1.903	0.026	0.039	1.514
DOE PCM	0.012	0.021	1.702	0.018	0.030	1.692
HadCM2	0.020	0.042	2.136	0.018	0.035	1.918

Variance of models and observations for all space and time domains. Also included are the ratios of land-only to land and marine variances.

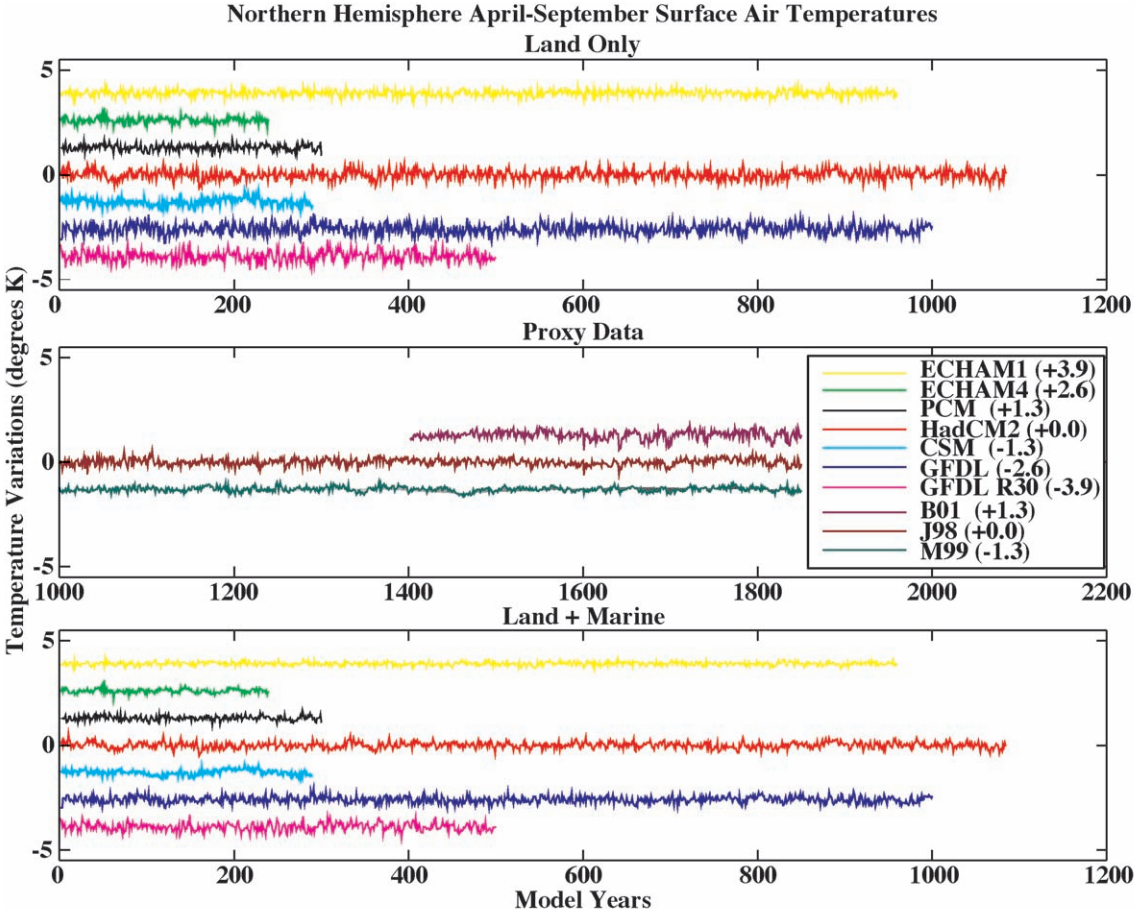

Fig. 3. Detrended model and proxy time series for NH growing season (April–September). For clarity, the results have been offset from the true values by the amount shown in parentheses in the legend. The upper and lower panels are the model results over land-only, and combined land and marine domains, respectively. The middle panel shows the detrended proxy data used in the analysis of growing season temperatures. The variances of each time series appear in Tables 3 and 4.

in magnitude than combined land and marine variability (Stouffer et al., 1994; Bell et al., 2000). Table 4 shows the *F*-test results of the comparisons of model variances to variances of the observations and the full length of each proxy. At most, the variance of an individual model simulation compares favorably (acceptance of the null hypothesis) with only one proxy in either domain (GSLM and GSLO). ECHAM4 and DOE PCM perform best in this analysis, with both models comparing favorably with both M99 and observations in the GSLM domain and with both J98 and observations in the GSLO domain (Table 4). In the GSLM domain, GFDL compares favorably with J98, and GFDL R30 compares favorably with B01. For the GSLO domain, ECHAM1 compares favorably with the observed variance, while NCAR CSM and HadCM2 both compare favorably with B01. Based on the *F*-test results, four of seven models overestimate both the GSLM and GSLO variances relative to the observed variances and ECHAM1 underestimates the observed GSLM variance.

5.1.3. Mean annual

For NH mean annual temperatures (MALM and MALO) (Fig. 4), temperature variance is more

Table 4
F-test results (%)

Model	B01	J98	M99	Observations	Model	B01	J98	M99	Observations
GSLM (%)					*MALM (%)*				
ECHAM1	0	0	0	0	ECHAM1	0	0	0	0
ECHAM3	NA	NA	NA	NA	ECHAM3	0	0	2	76
ECHAM4	0	0	9	99	ECHAM4	0	0	97	38
GFDL	0	19	0	0	GFDL	0	0	0	0
GFDL R30	14	0	0	0	GFDL R30	11	0	0	0
GFDL 14K	NA	NA	NA	NA	GFDL 14K	0	0	0	0
NCAR CSM	0	0	0	0	NCAR CSM	0	16	0	0
DOE PCM	0	0	75	20	DOE PCM	0	2	0	0
HadCM2	0	2	0	0	HadCM2	0	0	0	0
GSLO (%)					*MALO (%)*				
ECHAM1	0	0	0	32	ECHAM1	0	0	0	29
ECHAM3	NA	NA	NA	NA	ECHAM3	0	41	0	12
ECHAM4	0	26	0	10	ECHAM4	3	3	0	2
GFDL	0	0	0	0	GFDL	78	0	0	0
GFDL R30	0	0	0	0	GFDL R30	0	0	0	0
GFDL 14K	NA	NA	NA	NA	GFDL 14K	95	0	0	0
NCAR CSM	28	0	0	0	NCAR CSM	51	0	0	0
DOE PCM	0	44	0	5	DOE PCM	8	0	0	1
HadCM2	12	0	0	0	HadCM2	55	0	0	0

F-test results of comparisons between model variances and full-length proxy variances or observational variances for the validation period. A 5% significance level is used for evaluation of the null hypothesis (equal variances).

similar from model to model than for the growing season results (not shown), with the exception of the ECHAM1 model, which has slightly reduced variance relative to the other eight GCMs (Table 3). Again, the amplitude of the variance is greater in the land-only case for every model and for observations. ECHAM4 variance compares favorably with both M99 and observed variances in the MALM domain. Other favorable comparisons in the MALM domain include ECHAM3 with observations, NCAR CSM with J98, and GFDL R30 with B01. In the MALO domain, only ECHAM3 variance compares favorably with both proxy (J98) and observed variances (Table 4). Five of the nine model variances compare favorably with B01 in the MALO domain, but for each of the five models, the null hypothesis is rejected when comparing the model variances to the observations (Table 4). Based on the *F*-test results, seven of nine models overestimate the observed MALO variance, six of nine models overestimate the observed MALM variance, and ECHAM1 underestimates the observed MALM variance.

5.2. Periodic energy

We compared peaks in the power spectra that were above the median red noise background with greater than 90% confidence (power spectra not shown). For the validation period (1881–1980), the observations show significant quasi-periodic energy in the range of 2 to 5 years in all four time–space domains (not shown). The proxies also capture the quasi-periodic, 2–5 years energy over both the validation period and in the full record length. The proxies do not exhibit any significant energy for periods greater than 5 years, with the exception of B01, which displays additional energy at ~ 24 years (for the full record length). All of the models demonstrate significant energy at periods of 2–5 years in all four cases. Six of nine models also display significant energy between 6 and 10 years in the MALM domain (not shown), which is not found in any of the proxy or observational time series. In all four time–space domains, many of the models exhibit energy at periods greater than 5 years (not shown), but no

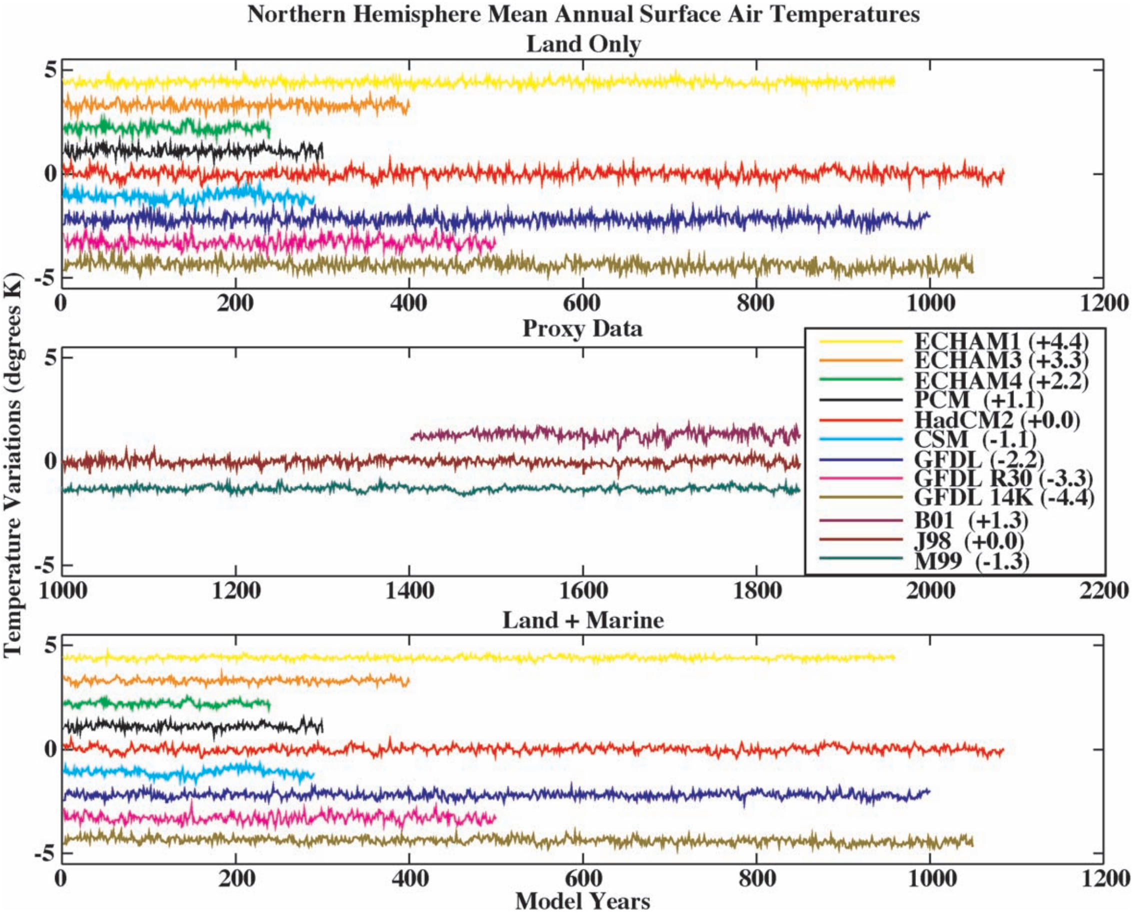

Fig. 4. Detrended NH mean annual temperatures. Only the first 1050 years of GFDL 14K are shown (14 000 years total). The proxy data displayed in the middle panel is the same data displayed in Fig. 3. Other details are also shown in Fig. 3.

consistent pattern is evident. Given the large number of bandwidths analyzed for each time series and the 10% test significance level, we would expect roughly this many false identifications of periodic energy.

5.3. Spectral shape

The lag one autocorrelation coefficient (α) is a measure of the shape of the power spectrum. An α value approaching zero indicates a white noise spectrum where the signal is entirely uncorrelated. An α value of one indicates a red noise spectrum where the signal is highly correlated. (A red spectrum (larger α) has greater energy in the low frequencies and less energy in the high frequencies.) Fig. 5 demonstrates the shape of power spectra with different values of α. A climate signal should exhibit some degree of redness in the spectrum due to the influence of slow response systems such as the ocean, which may vary with the length of the record.

Of the three proxies, only M99 does relatively well in the validation of α values (Table 5). The M99 α value for the validation period falls in between the α values of the observational MALM and MALO records (Table 6). Both J98 and B01 have much smaller α values during the validation period than any of the α values for the observational records. The observations show higher growing season α values than mean annual values. In an intermediate

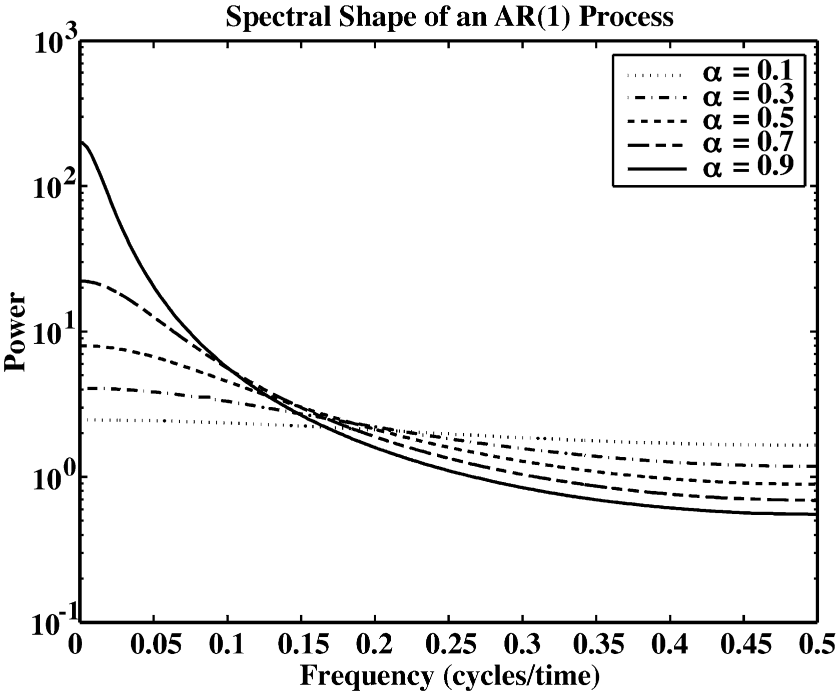

Fig. 5. Changes in the shape of an AR(1) process power spectrum due to changes in the value of the lag one autocorrelation coefficient (α).

comparison over a period when all three proxies overlap (1402–1850), the α values for J98 and B01 are almost identical and much closer to that of M99, although they are still less than the M99 value. Both M99 and J98 α values decrease when calculated over the entire period of 1000–1850. Table 6 shows the α values for the model results for the full length of each model simulation and for the observations over the validation period. The models generally compare favorably with the observations and proxies in both of the mean annual domains (MALM and MALO), slightly worse in the GSLM domain, and poorly in the GSLO domain.

Table 5
Lag one autocorrelation coefficients (α) of the proxies and observations

Proxies				Observations			
Period	B01	J98	M99	GSLM	GSLO	MALM	MALO
1881–1980	0.10	0.11	0.38	0.58	0.43	0.52	0.27
1402–1850	0.47	0.46	0.60				
1000–1850	NA	0.38	0.50				

Lag one autocorrelation coefficients (α) of the observations and of the proxies for three separate time intervals (1000–1850, 1402–1850, and 1881–1980) are shown.

6. Discussion and implications

6.1. Variance

The variance of each time series has been examined as a first-order analysis of climate variability. Observational data of the validation period (1881–1980) indicate relatively low variance particularly for the combined land and marine cases. The low variance may be due, at least in part, to the relatively short length of the observational record. Two of the three proxy records (J98 and B01) have significantly greater variance over the entire length of the record compared to the relatively short validation period (Table 2), which may be due to low-frequency variability not evident over the short validation period or to the spatial limitations associated with the proxy data. However, our sensitivity test of the effects of incomplete, nonrandom changes in spatial coverage

Table 6
Lag one autocorrelation coefficients (α) of the models and observations

	GSLM	GSLO	GSLM/GSLO	MALM	MALO	MALM/MALO	GSLM/MALM	GSLO/MALO
Observations	0.58	0.43	1.33	0.52	0.27	1.92	1.11	1.60
ECHAM1	0.25	0.16	1.52	0.37	0.26	1.39	0.68	0.62
ECHAM3	NA	NA	NA	0.41	0.23	1.77	NA	NA
ECHAM4	0.33	0.14	2.38	0.51	0.36	1.41	0.65	0.38
GFDL	0.32	0.19	1.64	0.44	0.26	1.70	0.72	0.75
GFDL 14K	NA	NA	NA	0.41	0.24	1.69	NA	NA
GFDL R30	0.48	0.28	1.72	0.61	0.41	1.47	0.79	0.67
NCAR CSM	0.22	0.07	3.36	0.40	0.23	1.72	0.55	0.28
DOE PCM	0.23	0.23	1.01	0.37	0.29	1.31	0.61	0.80
HadCM2	0.45	0.23	1.94	0.58	0.37	1.55	0.77	0.62

Model and observation lag one autocorrelation coefficients (α) are shown for all space–time domains (MALM, MALO, GSLM, and GSLO). Also shown are the ratios of land and marine to land-only α values and growing season to mean annual α values.

through time indicate that variance would be significantly lower (based on the *F*-test results) if spatially and temporally complete data were available.

Over the validation period, the proxy variances generally compare favorably with the observational variances (acceptance of the null hypothesis of equal variances based on the *F*-test results). The magnitude of the M99 variance falls in between the variances of the MALM and MALO observations, while the J98 and B01 variances (Table 2) tend to be closer to the observed GSLO and MALO variances (Table 3). This may be due in part to the "filtering" of the temperature signal by the various proxy indicators and/or to the types of proxy indicators used for these reconstructions.

The majority of the models overestimate the observed variances, and our sensitivity tests demonstrate that this overestimation may be even greater than the results indicate. Furthermore, both the proxies and observations are affected by solar variability and volcanic eruptions that are not present in the model simulations, which should increase variance values. Another approach in evaluating the model variances is to compare the observed ratio of land-only vs. land and marine variances to the model simulated ratios (Table 3). The observational variance for GSLO is ~ 50% greater than GSLM, while MALO variance is ~ 100% greater than MALM (Table 3). If we assume these relative variances are correct, then we can evaluate the model variances by evaluating the ratio of land-only to combined land and marine variances for both the seasonal and annual means. When comparing the seasonal means, all of the models overestimate the observational ratio of GSLO to GSLM variance. In the annual mean case, most models overestimate the MALO to MALM variance as well, although the HadCM2 MALO to MALM ratio is approximately equal to the observational MALO to MALM ratio (Table 3).

6.2. Periodic energy

Analysis of periodic energy over the validation period reveals a strong 2–5 years signal, which is evident in the mean seasonal and annual NH observations. The 2–5-year period is a well-known period and is easily attributed to ENSO. Somewhat surprising is the lack of a periodic signal greater than 5 years in the observations. Well-known climatic events, such as the Pacific Decadal and North Atlantic Oscillations (Appenzeller et al., 1998; Black et al., 1999), have return periods of greater than 5 years and presumably would appear in the observations. The lack of additional periodic energy may be due to the coverage affects discussed above, the relatively short length of the validation period, or to the nature of these climate processes, which may not manifest consistent (periodic or quasi-periodic) SAT changes over hemispheric seasonal and annual mean scales. For the full length of the proxies, only B01 displays significant energy with a period greater than 5 years (~ 24 years). This may indicate that the ABD calibration truly is preserving greater low-frequency variability as suggested (Briffa et al., 2001). Another possibility is the existence of a quasi-periodic climate signal that is most coherent during the time span of the B01 record. To test this, we repeated the spectral analysis of J98 and M99 over the B01 time interval (1402–1850). The new M99 spectrum does not display any additional periodic energy (outside of the 2–5-year period), but the J98 spectrum now exhibits a significant peak at ∼ 32 years, supporting the possibility of a quasi-periodic signal over the period of 1402–1850. A third possibility is that the peak in B01 at 24 years (and at 32 years in time-limited J98) is the result of simple chance or false peak identification.

All of the model spectra capture high-frequency energy (2–5 years periodicity) similar to the proxy and observational spectra. Many of the model spectra also exhibit significant energy at periods greater than 5 years. Specific attribution of the sources of this low-frequency variability exhibited by the models is beyond the scope of this study. Furthermore, some of these peaks may be due to statistical error as discussed above. While all of the models have significant energy at periods of 2–5 years, it is questionable if this energy can actually be attributed to a realistic simulation of ENSO. Bell et al. (2000) found that many of the models used in this study underestimate the tropical variability often associated with the ENSO signature. Spatial analysis on a model by model basis is necessary to attribute model variability to simulation of realistic climate processes.

6.3. Spectral shape

The lag one autocorrelation coefficient (α) of the power spectra is an important quantitative estimate of the distribution of internally generated model variability. Allen and Smith (1994) calculate $\alpha = 0.8$ for observed global mean annual temperatures over the period 1861–1990. They also found that introducing land and ice data and/or taking a global average temperature increases the level of autocorrelation, although our analysis shows lower autocorrelation in the land-only cases. Since land has less thermal inertia than oceans or ice and, therefore, less memory, it follows that the land-only cases should have decreased levels of autocorrelation, in contrast to the results of Allen and Smith (1994). Therefore, our estimates of lower α values may be acceptable for observations that are not global and, in general, do not include ice data or the associated land areas where ice is found (due to the observation coverage). The proxy autocorrelation coefficients are generally lower than the observational autocorrelation coefficients over the validation interval. Our sensitivity test of missing observational data indicates that the observational α values are probably too high (more red) and would in reality be lower (more white) and, possibly, more similar to the proxy values. For longer intervals where low-frequency energy should be more evident and lead to increased α values, the proxy α values are greater.

The models do an acceptable job of matching the proxies in the combined land and marine cases, but they perform poorly in the land-only cases. The models also compare more favorably with the proxies mean annually than seasonally. In comparison with observations, the model α values are fairly similar in the mean annual cases but always too low in the growing season cases. Since our sensitivity tests indicate that the observational α values may be too high, we compare the observational ratios of the combined land and marine cases to the land-only cases (Table 6). The mean annual observed land and marine α are almost twice that of the MALO case, while the two cases are more similar during the growing season. None of the models match the observed ratio in the mean annual case, although ECHAM3, GFDL and NCAR CSM get fairly close. These models get ratios of ∼ 1.73 by simulating MALO values which are similar to the observed MALO values, but with reduced MALM values. The observed growing season ratio is 1.33, meaning more closely similar α values in the land-only and land and marine cases. The models generally get a much higher ratio (greater discrepancies between the two cases) with the exception of DOE PCM which actually has a ratio of ∼ 1.0 (equal values in both cases); however, both values are half of the observed values, indicating insufficient low-frequency energy.

We also compared the α value ratios of the growing season vs. the mean annual cases (Table 6). For the observations, the combined land and marine cases are very similar and both are greater than the land-only cases, which is to be expected due to the influence of the oceans. Both observed α ratios of growing season to mean annual were greater than 1.0, indicating larger growing season α values. None of the models were able to capture the observed distribution of growing season vs. mean annual energy, resulting in α value ratios of growing season vs. mean annual that were less than 1.0.

7. Conclusions

We have applied quantitative methods to evaluate the internal climate variability simulated by several coupled general circulation models and paleotemperature reconstructions. In the process of this evaluation, we have demonstrated that both the proxy records and the model results achieve moderate success in precisely portraying statistics of climate variability. Although the proxies do not seem to capture the exact fingerprint of climate variability in terms of the precise variance, spectral shape, or periodic energy, they do appear to adequately capture the general magnitude of Northern Hemisphere climate variability. For each measure of variability analyzed, the models achieved varying degrees of success, with none of the models consistently capturing the appropriate variability. Most of the models overestimate variance in all four domains. In the analysis of spectral shape and periodic energy, the models broadly match the proxies and observations, but probably not entirely for the right reasons. The models have been shown to inaccurately portray ENSO; yet, they demonstrate periodic energy similar to what ENSO would produce. In terms of spectral shape, the models performed best in the analyses of mean annual cases, but failed to capture the distribution of energy between the growing season and mean annual cases. In general, the models performed worse in all analyses of the growing season cases and in the land-only cases. Conversely, the models performed best in the analysis of combined mean annual land and marine variability.

Acknowledgements

Funding for LCS and JLB from the David and Lucile Packard Foundation is gratefully acknowledged. JLB would like to thank Curt Covey, Ron Stouffer, Michael Mann, and Keith Briffa for useful discussion of data and methods. The authors also thank everyone who contributed proxy data and model results to this study as well as the two anonymous reviewers for insightful suggestions.

References

Allen, M.R., Smith, L.A., 1994. Investigating the origins and significance of low-frequency modes of climate variability. Geophys. Res. Lett. 21 (10), 883–886.

Appenzeller, C., Stocker, T.F., Anklin, M., 1998. North Atlantic Oscillation dynamics recorded in Greenland ice cores. Science 282, 446–449.

Barnett, T.P., Santer, B.D., Jones, P.D., Bradley, R.S., Briffa, K.R., 1996. Estimates of low frequency natural variability in near-surface air temperature. Holocene 6, 255–263.

Barnett, T.P., Pierce, D.W., Schnur, R., 2001. Detection of anthropogenic climate change in the world's oceans. Science 292, 270–274.

Bell, J., Duffy, P., Covey, C., Sloan, L., 2000. Comparison of temperature variability in observations and sixteen climate model simulations. Geophys. Res. Lett. 27 (2), 763–766.

Black, D.E., Peterson, L.C., Overpeck, J.T., Kaplan, A., Evans, M.N., Kashgarian, M., 1999. Eight centuries of North Atlantic Ocean atmosphere variability. Science 286, 1709–1713.

Blackman, R.B., Tukey, J.W., 1958. The Measurement of Power Spectra from the Point of View of Communication Engineering Dover, New York, NY.

Briffa, K.R., Schweingruber, F.H., Jones, P.D., Osborn, T.J., Shiyatov, S.G., Vaganov, E.A., 1998. Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. Nature 391, 678–682.

Briffa, K.R., Osborn, T.J., Schweingruber, F.H., Harris, I.C., Jones, P.D., Shiyatov, S.G., Vaganov, E.A., 2001. Low-frequency temperature variations from a northern tree-ring-density network. J. Geophys. Res. 106, 2929–2941.

Chatfield, C., 1996. The Analysis of Time Series, an Introduction. Chapman & Hall/CRC Press, Boca Raton, FL.

Ghil, M., Allen, M.R., Dettinger, M.D., Ide, K., Kondrashov, D., Mann, M.E., Robertson, A.W., Saunders, A., Tian, Y., Varadi, F., Yiou, P., submitted for publication. Advanced spectral methods for climatic time series. Rev. Geophys.

Griffies, S.M., Bryan, K., 1997. A predictability study of simulated North Atlantic multidecadal variability. Clim. Dyn. 13, 459–487.

Hasselmann, K., 1976. Stochastic climate models: Part I. Theory. Tellus 28, 473–485.

Jones, P.D., 1994. Hemispheric surface air temperature variations: a reanalysis and an update to 1993. J. Climate 7, 1794–1802.

Jones, P.D., Briffa, K.R., Barnett, T.P., Tett, S.F.B., 1998. High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with general circulation model control-run temperatures. Holocene 8 (4), 455–471.

Jones, P.D., New, M., Parker, D.E., Martin, S., Rigor, I.G., 1999. Surface air temperature and its changes over the past 150 years. Rev. Geophys. 37, 173–199.

Levitus, S., Antonov, J.I., Wang, J., Delworth, T.L., Dixon, K.W., Broccoli, A.J., 2001. Anthropogenic warming of earth's climate system. Science 292, 267–270.

Mann, M.E., Lees, J.M., 1996. Robust estimation of background noise and signal detection in climatic time series. Clim. Change 33, 409–445.

Mann, M.E., Bradley, R.S., Hughes, M.K., 1998. Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392, 779–787.

Mann, M.E., Bradley, R.S., Hughes, M.K., 1999. Northern Hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. Geophys. Res. Lett. 26 (6), 762–795.

Parker, D.E., Folland, C.K., Jackson, M., 1995. Marine surface temperature: observed variations and data requirements. Clim. Change 31, 559–600.

Percival, D.B., Walden, A.T., 1993. Spectral Analysis for Physical Applications. Cambridge Univ. Press, Cambridge, UK.

Santer, B.D., Wigley, T.M.L., Barnett, T.P., Anyamba, E., 1995. In: Houghton, J.T. et al. (Eds.), Detection of Climate Change and Attribution of Causes, in Climate Change 1995: The Science of Climate Change. Cambridge University Press, Cambridge, UK.

Stouffer, R.J., Manabe, S., Vinnikov, K.Y., 1994. Model assessment of the role of natural variability in recent global warming. Nature 367, 634–636.

Thomson, D.J., 1982. Spectrum estimation and harmonic analysis. Proc. IEEE 70 (9), 1055–1096.