PCMDI
Program for Climate Model Diagnosis and Intercomparison

UCRL-JC-131774

PCMDI Report No. 47

Determination of the Sampling Size of AGCM Ensemble Simulations

Michael F. Wehner

September 1998

Program for Climate Model Diagnosis and Intercomparison
University of California, Lawrence Livermore National Laboratory
Livermore CA 94550, USA

Table of Contents

Abstract

Historically, the high computational expense of performing lengthy climate simulations has limited the number of realizations in ensemble calculations. In this paper, we exploit the computational advantage of a recently developed parallel atmospheric general circulation model (AGCM) to determine the number of realizations required to calculate the model's statistics to a specified degree of certainty. Using standard statistical analysis techniques, this minimum ensemble size is found to be highly dependent on which output field is under examination. A strong dependence on location, season and averaging is also revealed.

1. Introduction

Multiple realizations of climate model simulations are increasingly being used to assess the variability and predictability of the real climate system (Barnett, et al 1997), (Zwiers, 1996), (Rowell, 1998). This is a fruitful approach, as the chaotic nature of the climate does not lend itself to deterministic simulation. Rather, by using an ensemble of simulations, the mean values and variances of selected quantities may be more realistically estimated than with a single calculation. However, the high computational burden associated with decadal integrations of general circulation models places pragmatic limits on the size of such ensembles. This limitation leads to uncertainties in model climate statistics that can mislead the analysis of a computational experiment. Hence, *a priori* estimation of the requisite number of realizations is of a practical use when considering the design of planned experiments. Clearly, such a minimum ensemble size depends strongly on what one expects to learn from the model.

In this paper, we present a methodology to assess requirements on ensemble size for a certain class of experiments. This technique relies upon estimations of the climate model's internal noise characteristics and the desired level of uncertainty in the model's output. In section 2, some current applications of ensemble climate simulations are discussed. The experimental design for this study is presented in section 3. Section 4 introduces the details of our technique to estimate minimum ensemble size. In section 5, results are obtained from the numerical experiment under a variety of different conditions for several variables. Some applications and limitations of this method are discussed in section 6. Finally, a discussion on the normality of climate model statistics is presented in an appendix.

2. Background

Much of the current interest in ensembles of simulations arises from recent attempts to perform seasonal forecasts. Brankovic and Palmer (1997) performed 9 sets of 120-day forecasts of all seasons over the years 1986-1990. By considering ensemble skill and consistency, they found that certain regions of the globe and seasons are good candidates for successful seasonal prediction. However, by extrapolating an analysis based on Student's t-statistic, they also find that the number of realizations required to confidently estimate the effect of SST anomalies is substantially greater than the number of simulations that they performed. More closely related to the present work in their experimental design are two related but separate studies of "potential predictability" using analysis of variance (ANOVA) techniques. Zwiers (1996) studied an ensemble of 6 calculations of the AMIP (defined below) period. Rowell (1998) studied an ensemble of 6 calculations of the 45-year period 1949-1993. Each of these authors measure potential predictability by partitioning the sources of model variability into an external component due solely to interannual changes in surface forcing and an internal component revealed by small changes in initial conditions. Both studies find significant latitudinal dependence of model potential predictability with greater values in the tropical regions. Significant longitudinal dependence is also found showing consistent influence of SST anomalies. Barnett, et al. (1997) examined the potential predictability of the Pacific-North American (PNA) pattern with ensembles of two models integrated over the period 1970 to 1993. They find that multiple realizations are required because the noise from model internal variability is about three times larger than the anticipated signal. They further conclude that a single AMIP realization is “not very useful” in quantifying the simulated interannual variability. Bengtsson, et al. (1996) analyzed atmospheric interannual variability with an ensemble of five AGCM integrations from 1979 to 1992. They concluded that larger ensembles may be necessary for ENSO prediction “in view of high internal atmospheric variability at high latitudes”. Kumar and Hoerling (1997) examined seven El Nino events with a thirteen member ensemble of AGCM integrations from 1950 to 1994. They found wide disparity between realizations in the correlation of individual El Nino events with the composite El Nino signal, further emphasizing the need for multiple realizations.

When atmospheric general circulation models are used as climate models, the surface forcing is generally specified in either of two manners. In the simpler, SST and sea ice extent are specified by climatological averages. If the seasonal cycle is to be represented, twelve monthly mean values are calculated and interpolated on a daily basis to provide a lower boundary condition. As this type of surface forcing is repetitive with the annual cycle, each year of a multi-year simulation is identical to every other except in the “initial” conditions at the start of year. Reasonable estimates of the model’s climate statistics may be estimated by approximating that each of these years is statistically independent. The internal variability of such experiments may then be represented by the interannual variability. Hence, in the absence of accumulated model bias, ensembles of multi-year simulations are not necessary.

In the second type of surface forcing, realistic SST and sea ice extent based on some sort of observations are used. In these kinds of simulations, (Gates, 1992, Folland and Rowell 1995) the surface forcing in each year is independent of the other. Part of the motivation in the design of this type of numerical experiment is to study the model’s response to an interannual variability in the surface forcing. However, in this case, the forcing in each of the simulated years are not identical and a single calculation through the specified time series is but one realization of a random variable (albeit of rather large dimension) determined by the initial conditions of the calculation. As a result, one cannot expect the mean values of these random variables to be uniquely defined by a single realization but rather possess some, presumably chaotic, dependence on initial conditions.

In fully coupled ocean-atmosphere general circulation models, the SST and sea ice extent are prognostic variables. If the makeup of the atmosphere is held constant, these calculations are similar to running a stand alone AGCM driven by climatological surface forcing. As the external forcing does not vary from year to year, a single stable long run may be used to estimate the statistical properties of the model’s output from the interannual variability.

Alternatively, coupled model simulations that include changes in atmospheric trace gas constituents or other forcings are similar to the AMIP-type experiments in that the external forcing is not held constant (Houghton, et al 1995, Meehl, et al 1997). As a result, the interannual variations of model output are time dependent functions of that forcing. Thus, ensembles of simulations are required to rigorously estimate the statistical properties of a model's output. Due the high computational cost of performing multi-century coupled model simulations, it is extremely important to estimate the necessary size of such ensembles prior to the experiment design.

3. Experiment Design

The Atmospheric Model Intercomparison Project (AMIP), an integration of the decade 1979 to 1988, provides a widely used standardized problem for the systematic intercomparison of AGCMs (Gates, 1992). In this configuration, the distribution of sea surface temperature (SST) and sea ice concentration is specified as a fixed lower boundary condition to the AGCM. To construct an ensemble of realizations, we performed twenty calculations of this period. All aspects of the model were maintained to be identical between simulations except for the initial conditions. These were generated in a prior twenty-day simulation by the daily output of a history restart file at a specified time of day. By starting the model with a different time slice of this restart history yet maintaining the model calendar, independent realizations of the AMIP period were obtained. Due to the highly chaotic nature of the atmospheric model, each realization is statistically independent after a short period of integration (Lorenz, 1968) (Barnett, 1995). Nonetheless, the first month of each realization is discarded in this analysis.

The model used in this study was the Lawrence Livermore National Laboratory's (LLNL) version of the UCLA AGCM. This model is a rewritten version of the model that UCLA used in their official submission to the PCMDI for AMIP. It differs from that model in that it is specifically targeted towards distributed memory massively parallel computers (Wehner, et al. 1995), (Wehner and Covey, 1995). The computational advantage of such parallel computing architectures (in this case, the Cray T3D) made this study feasible. In this study the horizontal resolution is 5^o in the longitudinal direction and 4^o in the latitudinal direction. There are fifteen vertical levels, one of which represents a well mixed planetary boundary layer. The model top is at 1.0 mb. Detailed documentation of the LLNL version is available on the AMIP World Wide Web pages (Phillips 1995).

4. Minimum ensemble size

To begin the statistical analysis of the AMIP ensemble, certain assumptions about the AGCM output data sets will be made. The first assumption is that each of the members of the ensemble are statistically independent. This assumption is justifiable by the well-known limits on the predictability of weather. In a chaotic system, any two realizations beginning arbitrarily close together will significantly depart from each other after a certain amount of time. In the atmosphere, under ideal circumstances, the limit for predictability from initial conditions is only a few weeks (Lorenz 1969, 1973). Practically speaking, at climate model resolutions, simulations significantly diverge from each other much more rapidly (Leith 1971). Nonetheless, we have discarded the first month of each simulation to ensure independence of the model's complete time sequence. Additionally, we have deliberately chosen initial conditions that are similar to each other. Although each initial condition differs by a day's worth of weather, any long-term biases generated by the model are not present in them. In contrast, if we had chosen the final state of one realization to serve as initial conditions for the next calculation, statistical independence might not be insured because of the possibility of an accumulation of model bias.

The second assumption is that the output from the AGCM realizations is Gaussian distributed. This assumption is made for statistical simplicity but may not be justified in all cases. For instance, distributions containing multiple mean climate states are precluded by this assumption. Likewise, single peaked distributions that are highly skewed by limits on the physical domain space or by other mechanisms are also disallowed. Both of these cases may indeed be possible. The former is represented by multiple climate equilibria (Manabe and Stouffer 1988) and the latter by state variables, such as snow depth or precipitation, which can only assume non-negative values. Quantifying the deviation from a normal distribution is possible given a large enough sample size. However, in the current study, a sample size of twenty does not appear to be sufficient to rule out non-Gaussian distributions with a high degree of statistical certainty. In the appendix, some results illustrating this conclusion are presented.

Nonetheless, the Gaussian assumption is a powerful one, allowing the usage of many standard statistical tools. In this study, we are interested in determining the minimum number of realizations required to calculate the mean climate state of the model. This may be restated in a more formal sense as the number of samples required to determine the mean value, m of a random variable to within a specified tolerance at a desired level of statistical confidence. For any normally distributed random variable, this number is,

$$n = \frac{Z_\alpha^2 \sigma^2}{E^2}, \tag{1}$$

where Z_α is a property of the normal distribution function, which depends on the desired statistical confidence, $1-\alpha$, σ^2 is an estimation of the population variance of the random variable and $m \pm E$ is the confidence interval of the population mean (Ott, 1993).

To successfully utilize this formula, a reasonable estimate of the population variance is required. Hence in this study, one additional assumption will be made. It is assumed that the sample variance, s calculated from twenty realizations sufficiently approximates the true population variance. From the chi-square relationship between these two quantities, a confidence interval for the population variance may be formed as,

$$\frac{(n-1)s^2}{\chi_U^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}, \tag{2}$$

where χ_L^2 and χ_U^2 are properties of the chi-square distribution with $n-1$ degrees of freedom and also are functions of the statistical confidence (Ott, 1993).

A statistical confidence of 95% is often chosen to constitute reasonable certainty. This translates into a one in twenty chance of not covering the population parameter. For the twenty samples in the current study, the 95% confidence values of χ_L^2 and χ_U^2 are 8.907 and 32.85 respectively. The 95% confidence interval for the population variance then becomes

$$0.58s^2 < \sigma^2 < 2.1s^2. \qquad (3)$$

We may then conclude that using the sample variance in equation 1 is a reasonable method to determine the minimum ensemble size to within approximately a factor of two with a high degree of statistical certainty. Finally, to continue to maintain this high level of certainty, we use the 95% confidence value, $Z_\alpha = 1.96$, in the analysis of equation 1.

5. Results

As might be expected, the amount of variability encountered in an AGCM simulation is a strong function of the field being analyzed. In general, longer term averages vary less between realizations than do shorter term averages. Additionally, some diagnostic fields exhibit greater variability than do others. The effect of spatial averaging also differs significantly from one diagnostic field to the next due to the differing amounts of spatial autocorrelation. To explore these issues and their consequences on estimating the minimum ensemble size, analyses are performed below on a number of different model output fields subject to these differing types of averaging.

a) Decadal mean seasonal averages

The long term simulated mean climatology is among the most important aspects of a climate model. Traditionally, three-month seasonal averages are often considered to be of a long enough time period to eliminate much of the noise due to weather processes yet frequent enough to resolve the seasonal cycle throughout the year. In this AMIP-based study, simulated data from ten individual seasons is used for the analysis of the (Northern Hemisphere) spring (MAM), summer (JJA) and fall (SON) averages. For the winter (DJF) average, only nine seasons are included because the integration period covered January 1979 to December 1988. In this subsection, we form the sample means and variances for the entire decade. Specifically, we first calculate the seasonal means over the entire decade for each realization followed by a calculation of the inter-realization variance from these means. In the subsection 5c, this order of calculation of variance and decadal averaging is reversed.

In figure 1, the number of realizations necessary to calculate the decadal mean seasonal surface temperatures to within ± 0.5° K at a statistical confidence of 95% is shown according to the formulae of the previous section. A cell-type graphic is shown to illustrate the point-by-point dependence. There are several interesting features to note. First, this figure serves as a limited test of the diagnostic software and the internal consistency of the ensemble by showing a result of zero over the open ocean. In these regions, the surface temperatures are constrained to be the specified SST values that are, of course, identical for each realization. Second, there is a strong seasonal and geographic distribution evident in figure 1. Outside of the polar regions, only the Northern Hemisphere in winter shows high variability and hence relatively large values for minimum ensemble size. These values are maximal over western central Eurasia at about twelve to fifteen simulations. In the polar regions, the variability is largest in regions where sea ice is undergoing formation. The minimum ensemble size is absurdly high in some of these regions and off the scale of the color bar (white). This is probably a result of the rather arbitrary changes in sea ice distribution dictated by AMIP that neglect the state of the atmosphere. In regions where sea ice is melting, variability is also reasonably high. Areas of permanent sea ice exhibit relatively low inter-realization variability. Finally, it should be recalled that the minimum ensemble size calculated in this manner is dependent on both the specified tolerance and statistical confidence according to equation 1. For example, reducing the tolerance to ±1.0°K reduces the values shown in figure 1 by a factor of four.

In figure 2, a similar result is shown for the decadal mean seasonally averaged precipitation. In this case, the specified tolerance is set to be 10% of the sample mean. Such a pointwise tolerance is more appropriate for this field due to its strong geographical dependence. In this case, a much larger pattern of variability in both space and time is seen. Tropical areas show significant inter-realization variability. Values of minimum ensemble size over tropical oceans vary in the ten to fifteen range. Note that there is a strong reduction of variability over tropical jungles. Finally note that in the very arid areas, such as deserts and off the coast of South America, the sample mean is so low as to produce very large values of minimum ensemble size. Clearly, these regions should be discounted when deciding how many simulations to make.

The local energy budget and hydrological cycle as analyzed in the preceding two figures appears to exhibit a high degree of inter-realization variability. It is fruitful to also examine the variability of more dynamical aspects of the simulated atmospheric circulation. In figure 3, the number of runs to calculate the equivalent sea level pressure to with 1.0mb at a 95% statistical confidence is shown. The most apparent feature is how much lower this number is than for the previous two fields despite a rather severe tolerance. Areas of modest variability are associated with the location of the Northern Hemisphere winter predominant low-pressure centers and also with polar winters in both hemispheres. Variability in the polar winters may be associated with sudden cross-polar flows in the middle atmosphere (Manney, et al 1994). These can have the effect of quickly changing the surface pressure by as much as 40 mb, especially in the south. Apparently, the frequency and severity of these events varies significantly between realizations on the seasonal scale.

Substantial inter-realization variability is not limited to model surface fields. Figure 4 shows the number of realizations required to calculate the decadal and zonal mean air temperature aloft (between 1.0 and 850 mb) to within 0.5^{o}K at a 95% statistical confidence. This result was obtained by first calculating the zonal and decadal mean values followed by a variance calculation. Variability associated with the polar nighttime winter jets exhibit strong hemispheric differences. Significantly more realizations are required to predict the middle atmospheric temperatures in the northern winter (DJF) than in the southern winter (JJA).

b) Spatial averaging

For some fields, the ensemble size dictated by examining each cell individually, as in figures 1 through 3, may be significantly larger than necessary. Averaging an output field over neighboring cells prior to calculating its variance partially reveals the nature of the spatial autocorrelation of the variance, if any, for that field. For fields that exhibit high positive spatial autocorrelation in their variances, we expect that the amount of variability will not decrease by much under modest spatial averaging. However, for those fields with variances that are not spatially autocorrelated or even exhibit spatial antiautocorrelation, we expect that the amount of variability will significantly decrease by spatial averaging. Figure 5 shows the minimum ensemble size for the surface temperature under the same conditions as in figure 1 except that an area weighted average of the original 4^ox5^o model output to an 8^ox10^o resolution is performed prior to the variance calculation. The high degree of spatial autocorrelation of surface temperature is evident in that the values shown in figure 5, although slightly reduced due to smoothing, are comparable to those shown in figure 1.

The effect of spatial averaging on the inter-realization variability of precipitation is considerably different. Figure 6 shows the minimum ensemble size for precipitation under the same conditions as in figure 2 except for a spatial averaging. Large differences in the two sets of figures are apparent. The high variability associated with tropical convective processes over oceans is nearly eliminated. The remaining variability is associated with mid-latitude large-scale precipitation processes or with exceedingly dry regions. The pronounced effect of such minimal averaging reveals that convective precipitation in the model is spatially autocorrelated in a very different way than is temperature. There are a number of potential reasons for this including the possibility that a downwind drying effect causes patterns of correlated wet and dry columns. This and other related issues of spatial autocorrelation will be investigated in a later paper.

c) Single season averages

There is much current interest in seasonal forecasting (Barnett, 1995) (Brankovic and Palmer, 1997). In this case, predictions of an average climatology for a specific upcoming (single) season are the principal output of the model. Clearly, a model’s inter-realization variability increases as the period of temporal averaging decreases. In this section, we compute the number of realizations to calculate model seasonal output for each of the years in the AMIP period. We then average this quantity over the entire decade, effectively reversing the order of variance and decadal averaging procedures from the previous subsections. For ease of comparison with the preceding analysis, the tolerances and statistical certainties are kept the same for each field.

In figure 7a, the average number of realizations to calculate a single northern winter season (DJF) surface temperature to within $\pm 0.5^{\circ} K$ at a statistical confidence of 95% is shown. On first inspection, the *pattern* of variability is not so different for a single season than for the decadally averaged results of figure 1. This is true of other seasons as well. However, the color scale of figure 7a is a factor of five greater in range reflecting a larger magnitude of variance. The effect of temporal averaging on required number of realizations is more clearly shown in figure 7b, where we plot the ratio of the single season result to the decadal result. In this figure, a minimum value of unity is imposed on the calculated number of realizations for ease of interpretation. There is substantial seasonal and geographical structure to this ratio. However it is difficult to ascribe any particular significance to this structure other than to recognize the much larger effect of temporal averaging than spatial averaging for modeled temperature. In figure 7c, we show this same ratio as calculated for model precipitation averaged to the 8^{o}x10^{o} mesh. A similar increase in the magnitude of variability is calculated. Finally, in figure 7d, this ratio of single season to decadal ensemble size is plotted for equivalent sea level pressure. In this case, the increase in minimum ensemble size is considerably larger revealing that temporal averaging is more effective in reducing the internal variability of pressure than it is for temperature and precipitation.

6. Discussion

A methodology is presented to estimate *a priori* the required minimum size of ensemble climate simulations. In this paper, this minimum ensemble size is defined by the degree of certainty required for the analysis of the model's mean climate statistics and its internal variability. The technique, developed from standard statistical formulae, requires only a reasonable estimate of the model's inter-realization population variance. We have applied this methodology in the context of the AMIP experiment. However, generalization to other classes of climate simulations with temporal variations in forcing is appropriate.

We find that minimum ensemble size calculated in this manner is a highly subjective quantity, strongly dependent on the intended use of the model's output. As one might expect, a strong dependence on the chosen model output fields is exhibited. Seasonal and geographic considerations also play a significant role in deciding what the proper number of realizations should be. For instance, larger internal variability in the Northern Hemisphere wintertime thermal structure indicates that more realizations are required for studies involving northern wintertime temperatures than are those involving summertime temperatures. Spatial and/or temporal averaging has markedly different effects on different variables as seen in the previous section. The degree to which additional temporal averaging reduces internal variability varies greatly from one variable to another and is dependent on location. The effect of spatial averaging is dependent on an output field's spatial autocorrelation and can significantly change the character of variability.

A useful application of the approach presented here is to aid in determining the statistical significance of results due to changes to model parameterizations or forcing. In many cases, it is difficult to access whether such changes in model output are in fact a result of a change in algorithm or are simply part of the natural internal variability of the climate model. A carefully designed experiment to help answer this question is to perform pairs of ensemble simulations using differing model configurations. Under the assumption that the population variance is the same for each ensemble, the number of runs to determine the difference in a pair of ensemble means is exactly twice that given in equation 1. Here, the statistical confidence, α, and the tolerance, E now refers to the difference between the two ensemble means and n, the number of realizations, is that required for *each* ensemble. Hence, an *a priori* estimate of the magnitude of the anticipated changes to the model's mean statistics is needed. In many cases, this may be reasonably provided by considerations external to the climate model itself.

However, it should be noted that considerations other than the degree of accuracy of the ensemble mean may be relevant to determining the number of required realizations depending on the end uses of model output. For example, forecasters need to accurately understand their model's predictive skill. Optimization of ensemble size based on skill criteria may indeed be quite different (Barnett, et al. 1997).

It should be further noted that the results presented in figure 1-7 are applicable only to a specific atmospheric model running a specific experiment (AMIP). Other climate models certainly possess differing degrees of internal variability and thus will yield differing estimates of minimum ensemble size. However, it is likely that certain features of model internal variability may not be inconsistent with the internal variability of the real climate system. Planned model intercomparison projects are an ideal vehicle to add to our understanding of model variability. It is hoped that interested modeling groups will perform multiple realizations with their own models of AMIP2, the follow-on experiment to AMIP (Gates, et al. 1998). An experimental subproject has been charged with coordinating this activity (Zwiers, et al. 1998).

Finally, we consider that the techniques presented here are applicable to certain climate change scenario simulations using fully coupled ocean-atmosphere general circulation models. As is evident by the results of the previous sections, the required size of such ensembles may be formidable. It is hoped that high performance computing technologies will soon increase in power to the point where ensembles of multi-century climate simulations are feasible.

Appendix-Normality of AGCM output

Testing the normality assumption of a random distribution is possible given a sufficient sample size. The Lilliefors and Shapiro-Wilk algorithms are two commonly used tests for normality (Conover, 1980). Both of these "goodness of fit" tests posit a null hypothesis, H_0, that the distribution function is a Gaussian function with unspecified mean and variance. The statistical confidence for rejection of this null hypothesis is the test's principal result. One may interpret this confidence level as the probability that the distribution is non-Gaussian. Failure to reject the null hypothesis with high confidence does not necessarily mean that the actual distribution is Gaussian. Nonetheless, these statistical tests offer some guidance as to the reasonableness of the Gaussian approximation.

Figure A1 shows a map of where the Lilliefors test rejects the null hypothesis at the 90% statistical confidence level for decadal mean DJF precipitation, surface temperature and sea level pressure. Figure A2 is the same result from the Shapiro-Wilk test. Although the two sets of figures are not identical, rejection of the Gaussian distribution hypothesis is not widespread. The numerous isolated cells that do reject the null hypothesis are likely to be simply the result of chance as they comprise roughly 10% of the total number of cells. However, a few regions do seem to exhibit some spatial coherence. For instance, in section 5 we noted that the variance of the surface temperature over newly formed sea ice was quite large in this AMIP experiment. These same regions reject the null hypothesis at greater than a 99% statistical confidence. Precipitation over the Sahara is also likely to possess a non-Gaussian distribution function. Note however, that the Lilliefors test does not reject the null hypothesis in this region as robustly as the Shapiro-Wilk test.

As with similar statistical tests, rejection of the null hypothesis becomes highly likely for very large sample sizes. Care must be taken in such circumstances, as the criterion for rejection may be overly restrictive from a practical viewpoint. In simpler statistical tests, such as the f-test or student-t test, one can apply physical arguments to determine if statistically significant differences between the compared fields (i.e. the variances or means) are large enough to be important to an analysis. For “goodness of fit” tests, the value of the test statistic used to determine the statistical confidence for rejection of H_0 provides some absolute measure of the deviation of the sample distribution from the postulated normal distribution. However, such spurious rejection of the null hypothesis does not appear to be a significant factor for the modest sample size used in this study.

Acknowledgements

The author expresses his gratitude to Dr. Ben Santer for initially suggesting this course of study. This work was performed under the auspices of the US Department of Energy Office of Biological and Environmental Research by the Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

References

Barnett TP (1995) Monte Carlo climate forecasting. J Clim 8:1005-1022

Barnett TP, Arpe K, Bengtsson L, Ji M, Kumar A (1997) Potential predictability and AMIP implications of midlatitude climate variability in two general circulation models. J Clim 10:2321-2329

Bengtsson L, Arpe K, Roeckner E, Schulzweida U (1996) Climate predictability experiments with a general circulation model. Clim Dyn 12:261-278

Brankovic C and Palmer TN (1997) Atmospheric seasonal predictability and estimates of ensemble size. Mon Weather Rev 125: 859-874.

Conover WJ (1980) Practical nonparametric statistics, 2nd ed. Wiley

Folland CK and Rowell DE, eds. (1995) Workshop on simulations of the climate of the twentieth century using GISST. Hadley Centre Technical Note (CRTN) 56

Gates WL (1992) The Atmospheric Model Intercomparison Project, Bull Am Meterol Soc 73:1962-1970

Gates WL, Boyle J, Covey C, Dease C, Doutriaux, C, Drach R, Fiorino M, Gleckler P, Hnilo J, Marlais S, Philips T, Potter G, Santer BD, Sperper KR, Taylor K and Williams D (1998) An overview of the results of the Atmospheric Model Intercomparison Project (AMIP), to appear in Bull Am Meterol Soc

Houghton, JT, Meira Filho LG, Callander BA, Harris N, Kattenberg A, Maskell K, eds. (1995) Climate Change 1995: The Science of Climate Change, Intergovernmental Panel on Climate Change, Cambridge University Press.

Kumar A and Hoerling MP (1997) Interpretation and implications of the observed Inter-El Nino variability. J Clim 10:83-91

Leith CE (1971) Atmospheric Predictability and two-dimensional turbulence. J Atmos Sci 28:145-161

Lorenz EN (1968) Climatic determinism. Meteor Monogr 8:1-3

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26:636-646

Lorenz EN (1973) On the existence of extended range predictability. J App Meteorol 12:543-546

Manabe S and Stouffer RJ (1988) Two stable equilibria of a coupled ocean-atmosphere model. J Clim 9:841-866

Manney GL, Farrara JD and Mechoso CR (1994) Simulations of the February 1979 sudden warming: Model comparisons and three-dimensional evolution. Mon Weath Rev 122:1115-1140.

Meehl GA, Boer GJ, Covey C, Latif M, Stouffer RJ (1997) Intercomparison makes for a better climate model, EOS Trans 78:445-451

Ott RL (1993) An introduction to statistical methods and data analysis, 4th ed. Duxbury Press, Belmont, CA. (Or many other introductory statistics books.)

Phillips T (1995), AMIP Model Documentation, http://www-pcmdi.llnl.gov/modeldoc/amip/23llnl_ToC.html

Rowell DP (1998) Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. J Clim 11:109-120

Wehner MF and Covey C (1995) Description and validation of the LLNL/UCLA parallel atmospheric GCM. UCRL-ID-123223, Lawrence Livermore National Laboratory Internal Report

Wehner MF, Mirin AA, Eltgroth PG, Dannevik WP, Mechoso CR, Farrara J, Spahr J (1995) Performance of a distributed memory finite difference atmospheric general circulation model. Parallel Computing 21:1655-1675

Zwiers F (1996) Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. Clim Dyn 12:825-847

Zwiers F, et al (1998) AMIP experimental subproject No. 1: Multiple realizations. http://www-pcmdi.llnl.gov/amip/EXPSUBS/epno1.html

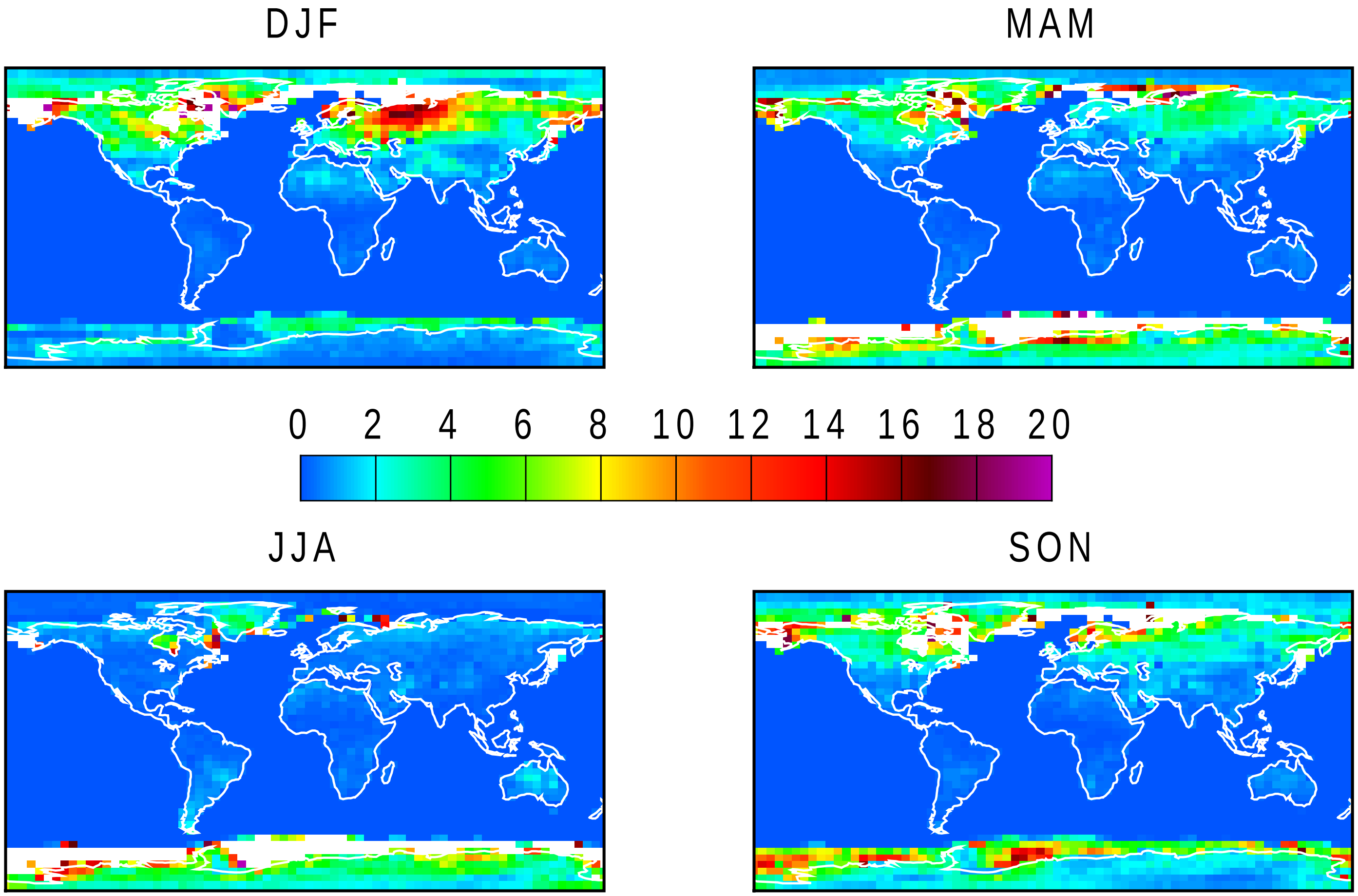

Figure 1: The number of realizations required to estimate the decadally averaged seasonal mean surface temperatures to within 0.5^{o}K with 95% statistical confidence.

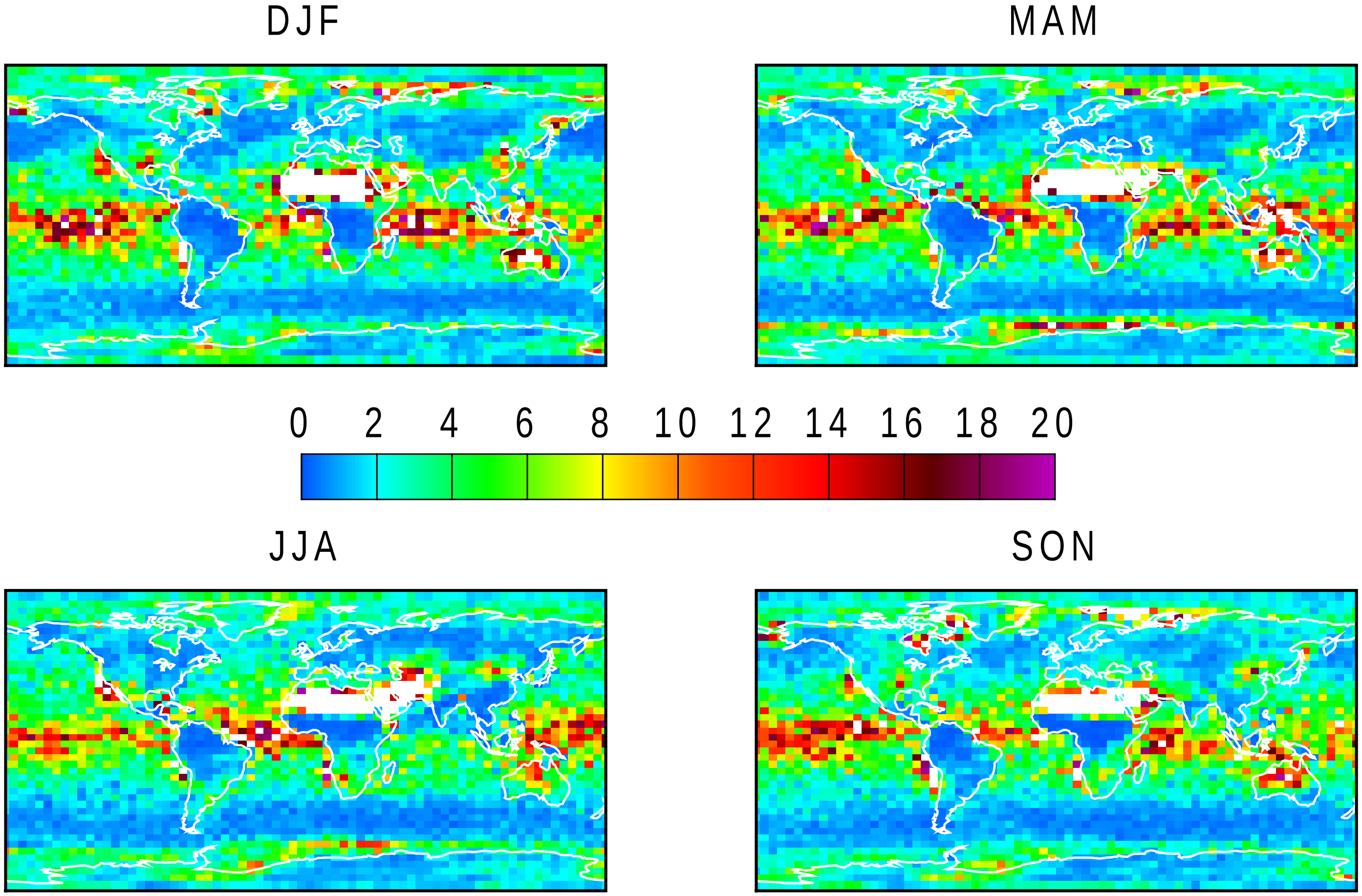

Figure 2: The number of realizations required to estimate the decadally averaged seasonal mean precipitation to within 10% of the calculated mean value with 95% statistical confidence.

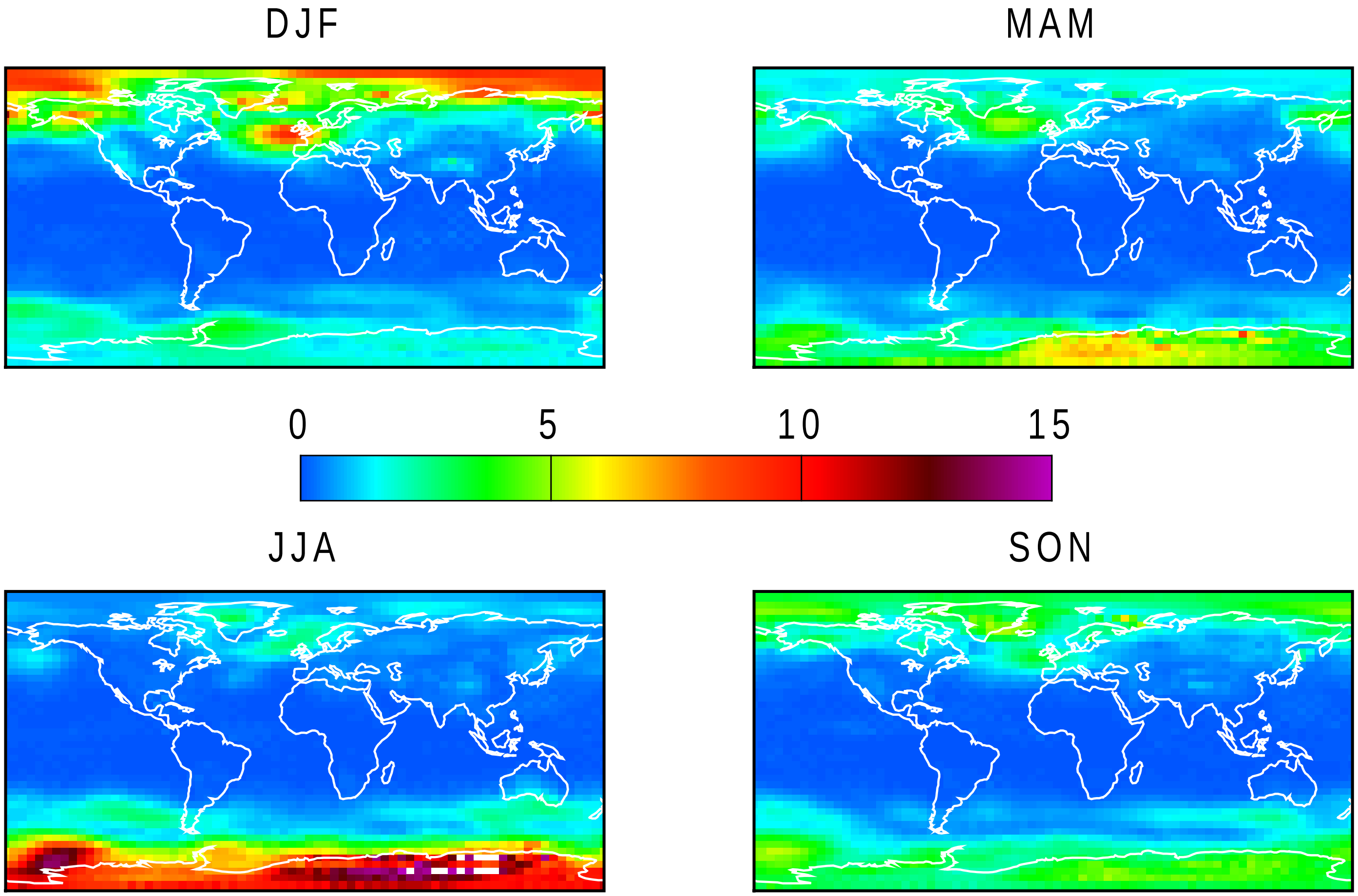

Figure 3: The number of realizations required to estimate the decadally averaged seasonal mean sea level pressure to within 1.0mb with 95% statistical confidence

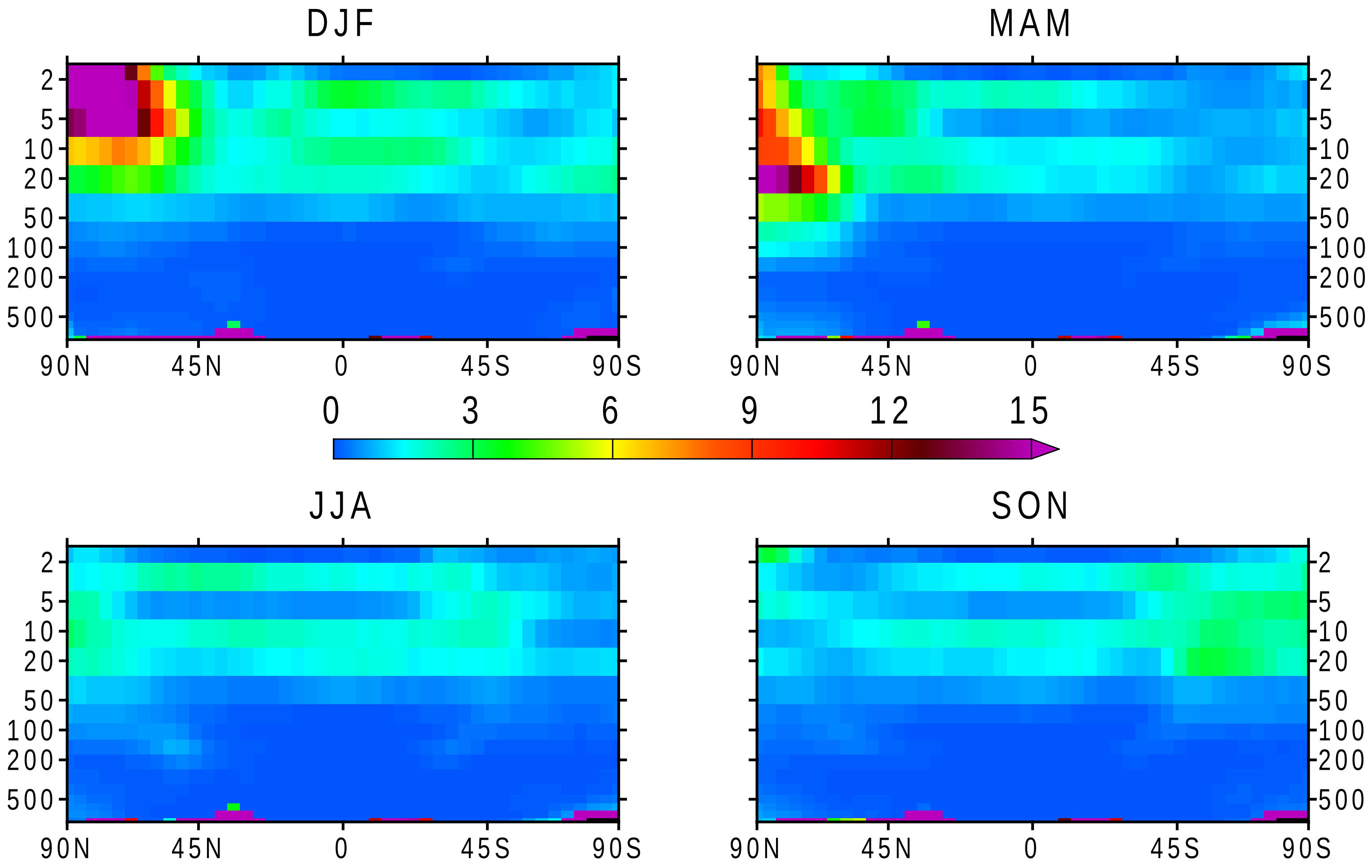

Figure 4: The number of realizations required to estimate the decadally averaged zonal mean seasonal air temperatures to within 0.5ºK with 95% statistical confidence. (The units of the vertical axis are pressure in hPa.)

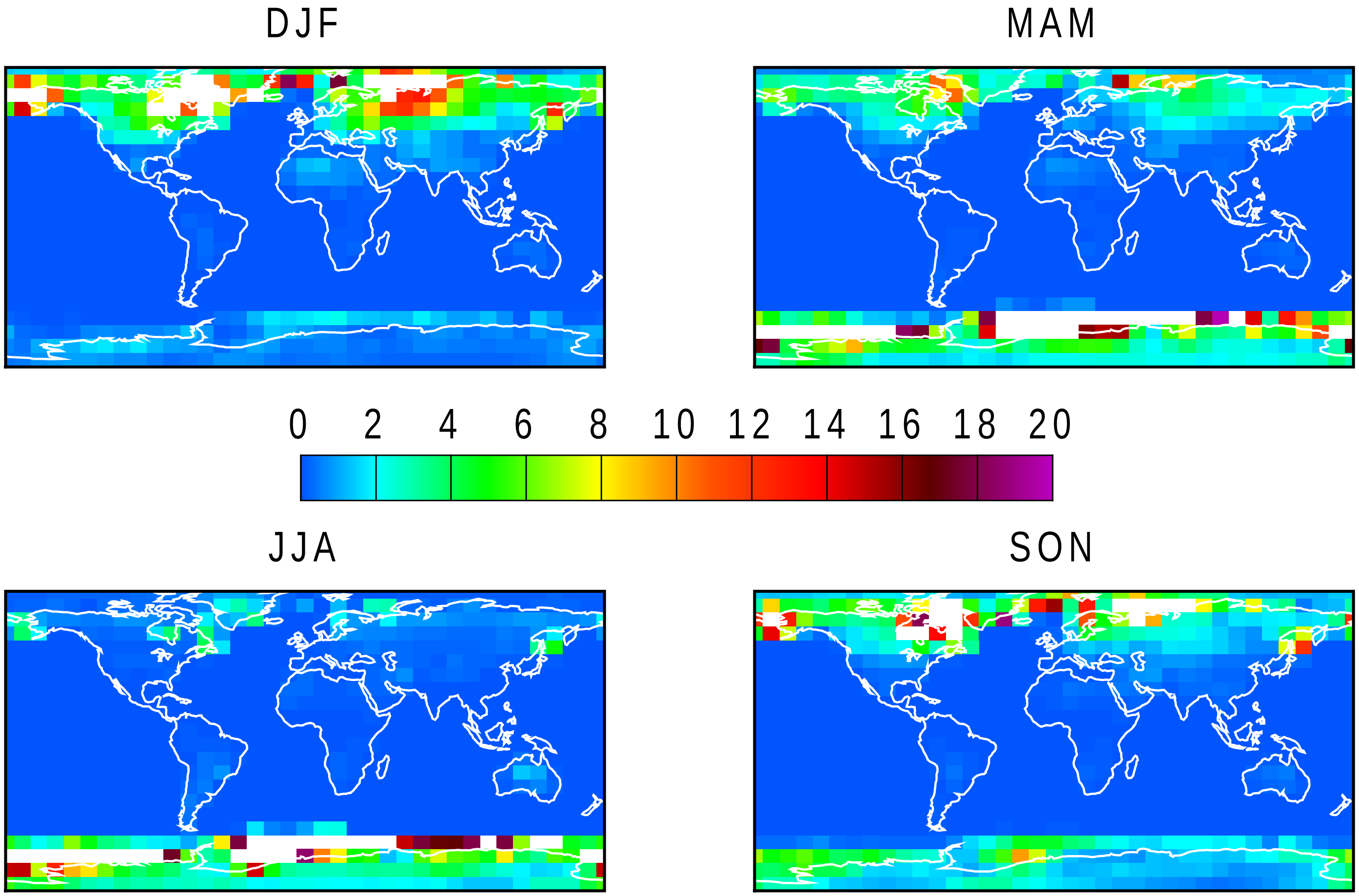

Figure 5: The number of realizations required to estimate the decadally averaged seasonal mean surface temperatures to within 0.5^{o}K with 95% statistical confidence after a spatial averaging of the original field to 8^{o}X10^o.

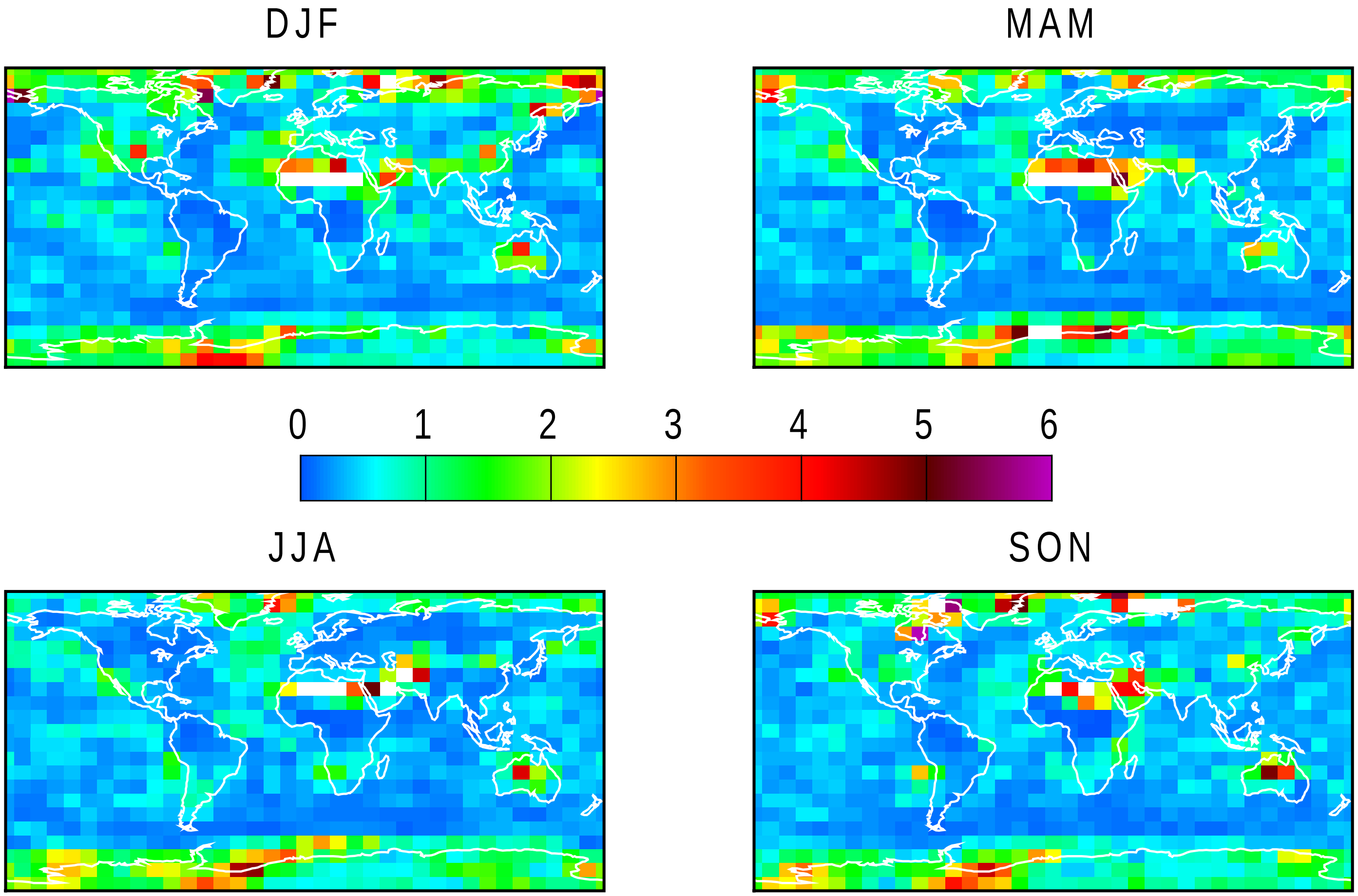

Figure 6: The number of realizations required to estimate the decadally averaged seasonal mean precipitation to within 10% of the calculated mean value with 95% statistical confidence after a spatial averaging of the original field to $8^{o}X10^{o}$.

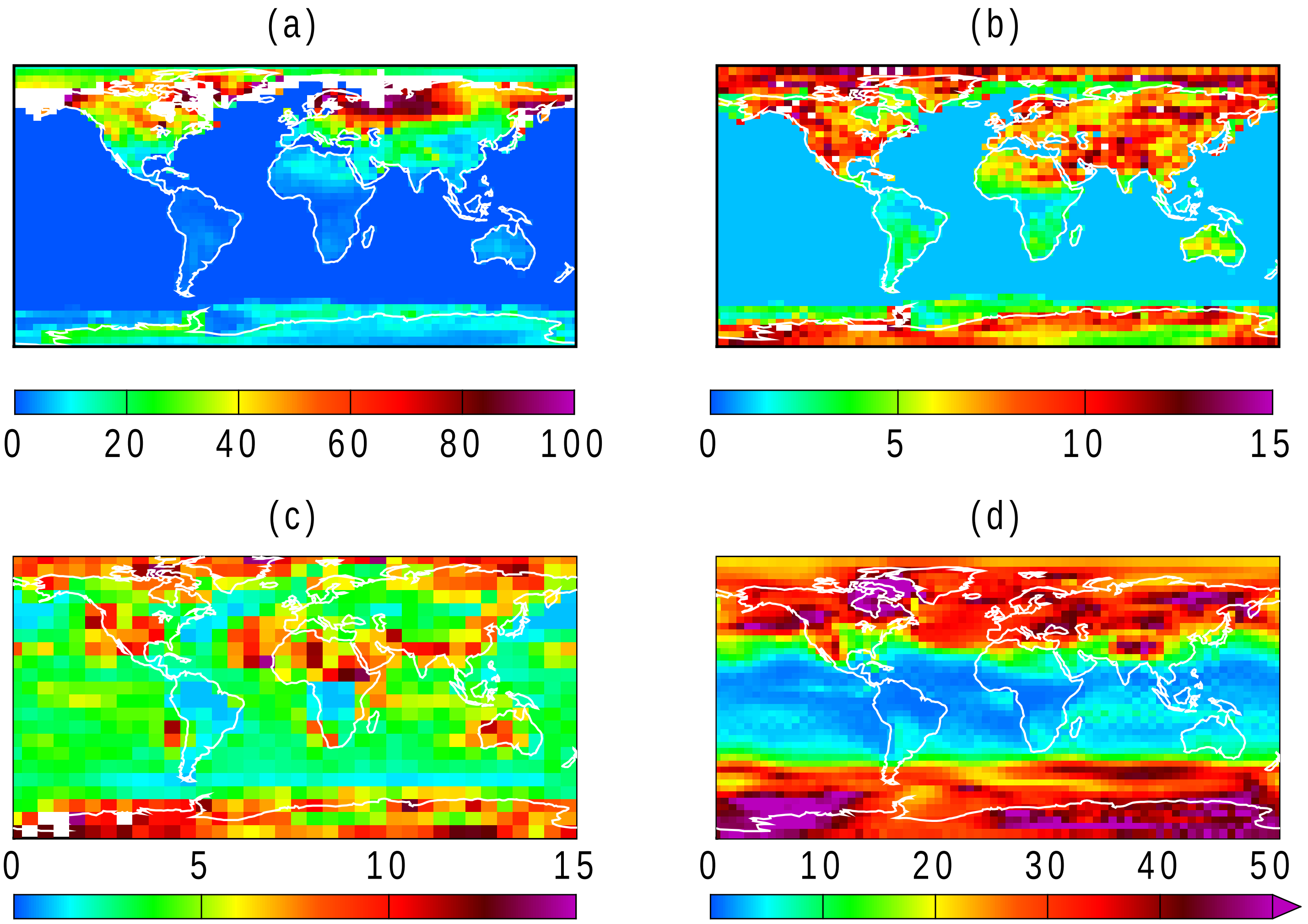

Figure 7 a) The average number of realizations required to estimate a single DJF mean surface temperatures to within 0.5°K with 95% statistical confidence. b,c,d) The ratio of the required number of single DJF realizations to the required number of decadal DJF realizations. b) Surface temperature (TG) to within 0.5 °K at 95% statistical confidence. c) Precipitation (PR) averaged to 8°X10° to within 10% of the calculated mean value with 95% statistical confidence. d) Sea level pressure (PSL) to within 1.0mb with 95% statistical confidence.

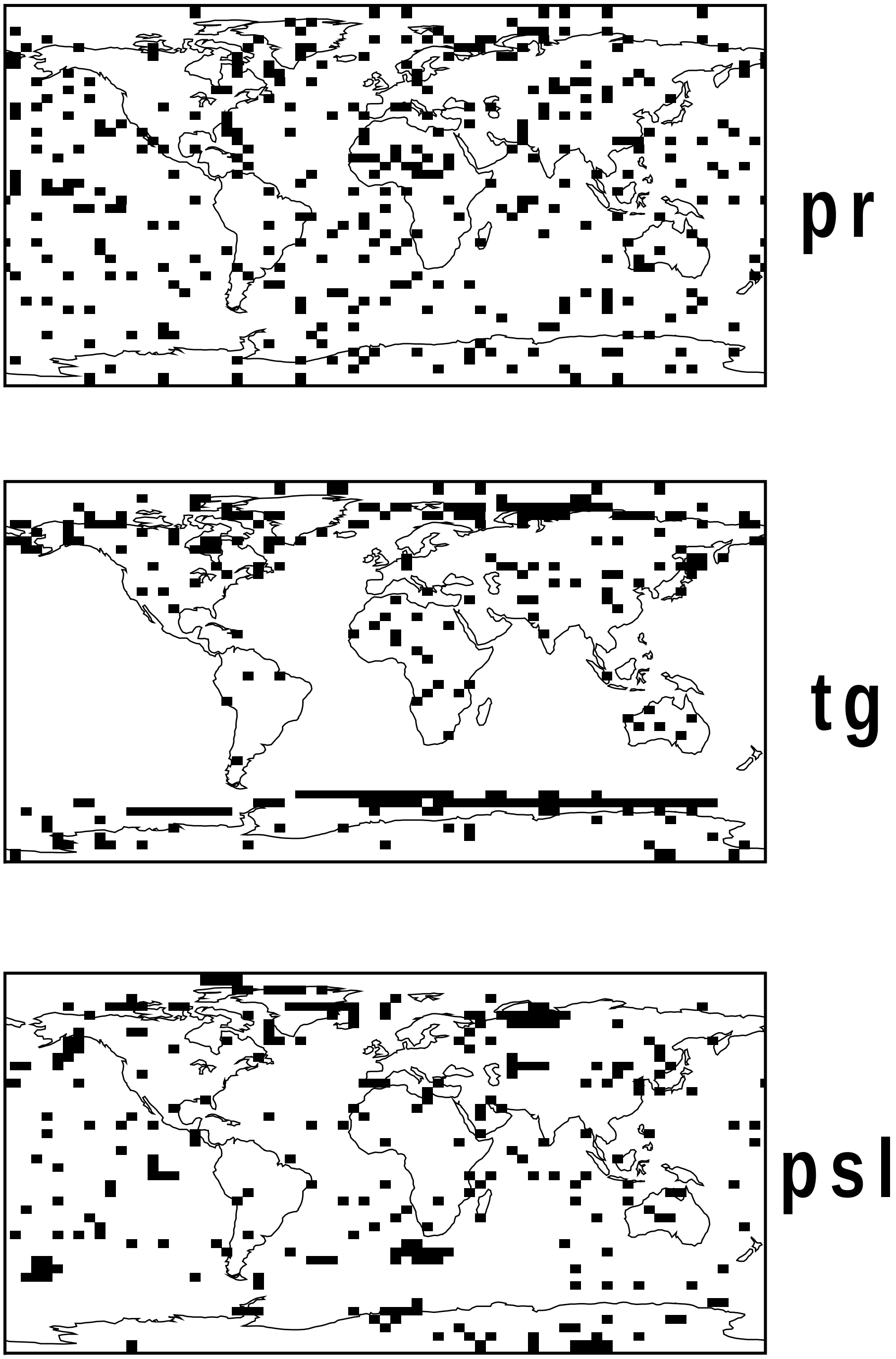

Figure A1: Areas of the globe where the statistical confidence for rejection of the null hypothesis exceeds 90% as calculated by the Lilliefors test. TG: surface temperature. PR: precipitation. PSL: sea level pressure.

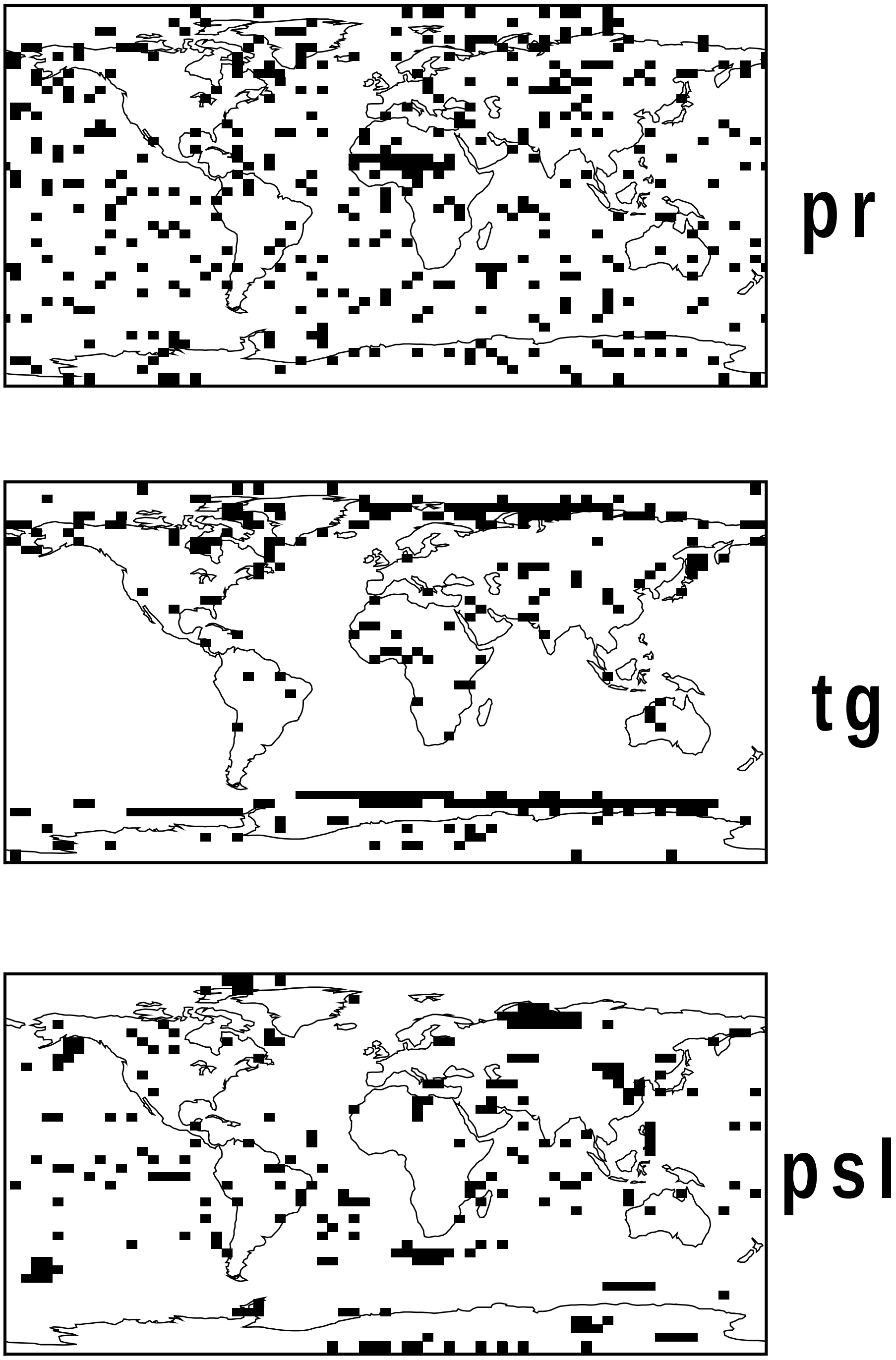

Figure A2: Areas of the globe where the statistical confidence for rejection of the null hypothesis exceeds 90% as calculated by the Shapiro-Wilk test. TG: surface temperature. PR: precipitation. PSL: sea level pressure.