Two Limits of Initial-value Decadal Predictability in a CGCM

Grant Branstator and Haiyan Teng

National Center for Atmospheric Research, Boulder, Colorado

Submitted to J. Climate

Revised, August 2010

Corresponding author address: Grant Branstator

National Center for Atmospheric Research, Boulder, CO, 80305.

E-mail: branst@ucar.edu

ABSTRACT

When the climate system undergoes changing external forcing (e.g. from increases in greenhouse gas and aerosol concentrations), there are two inherent limits on the gain in skill of decadal climate predictions that can be attained from initializing with the observed ocean state. One is the classical initial-value predictability limit that is a consequence of the system being chaotic, and the other corresponds to the forecast range at which information from the initial conditions is overcome by the forced response. These limits are not caused by model errors; they correspond to limits on the range of useful forecasts that would exist even if nature behaved exactly as the model behaves.

In this paper these two limits are quantified for Community Climate System Model, version 3 (CCSM3) with several 40-member climate change scenario experiments. Predictability of the upper-300m ocean temperature, on basin and global scales, is estimated by relative entropy from information theory. Despite some regional variations, overall, information from the ocean initial condition exceeds that from the forced response for about seven years. After about a decade the classical initial-value predictability limit is reached, at which point the initial condition has no remaining impact. Initial-value predictability receives a larger contribution from ensemble mean signals than from the distribution about the mean. Based on the two quantified limits we conclude that, to the extent that predictive skill relies solely on upper ocean heat content, in CCSM3 decadal prediction beyond a range of about 10 years is a boundary condition problem rather than an initial-value problem. Factors that our results are sensitive and insensitive to are also discussed.

1. Introduction

The scientific community is now taking on the challenge of using initialized models to produce time-evolving climate predictions for the next 10-30 years (Smith et al. 2007, Keenlyside et al. 2008, Pohlmann et al. 2009). Such predictions will be a key component of the next Intergovernmental Panel on Climate Change (IPCC) Assessment Report (Taylor et al. 2009). Compared with traditional climate change experiments, the fundamental difference in these forecasts is that the initial ocean state is determined from observations, and the hypothesis is that the resulting forecasts will substantially benefit from this added information. But the duration of the influence of the ocean initial condition remains unknown. Since the climate system is chaotic, inevitable errors in the initial condition grow with time causing the initial signals to fade (Lorenz 1963). Eventually the impact of the initial condition becomes undetectable, placing a fundamental limit on its influence. If one considers a situation where forcing of the climate system is changing, a second limit on initial condition influence should be introduced. For, if, as in the case with forcing by the ongoing changes in greenhouse gas (GHG) and aerosol concentrations, the system response increases with time, at some point the initial condition influence becomes of secondary importance compared to the forced response. In this paper, we quantify the forecast range at which these two limits are reached. Our results should help to determine the feasibility and value of decadal predictions (Meehl et al. 2009, Hurrell et al. 2009, Solomon et al. 2009).

Since the observational record is so short, there are no methods of measuring the predictability limits of the natural system. But one can estimate these limits for the numerical models that are used to simulate and predict nature. In general, predictability properties can be different for two dynamical systems that in many ways appear to be similar. Hence, studies, like ours, that quantify the initial value predictability of a particular model are not necessarily quantifying the predictability of nature. But just as the predictability limits of nature impose limits on the range at which initial conditions can influence forecasts, so do the predictability limits of the models used to predict nature. Hence the predictability of models needs to be quantified. And to the extent that a given model is a good surrogate for nature, its predictability limits give some indication of the predictability limits of nature.

Because initial-value predictability concerns how rapidly a cluster of similar initial states evolves to a distribution that is statistically indistinguishable from the system's climatological distribution, a common approach for quantifying initial-value predictability of a model is to perform ensemble experiments with perturbed initial conditions.[1] Most previous studies have focused on the North Atlantic, particularly the Atlantic meridional overturning circulation (AMOC, Griffies and Bryan 1997a, 1997b, Collins 2002, Collins and Sinha 2003, Pohlmann et al. 2004, Collins et al. 2006). Many of these investigations concur that the AMOC is potentially predictable a decade in advance, but the characteristics of the AMOC, including its predictability limits, vary from model to model (Collins et al. 2006, Latif et al. 2006, Hurrell et al. 2009). Like the North Atlantic, the North Pacific exhibits strong decadal variability (with the dominant mode called the Pacific Decadal Oscillation, PDO, Mantua et al. 1997). While many studies suggest that the North Pacific also has decadal predictability as a result of ocean Rossby wave propagation (Latif and Barnett 1994, Kwon and Deser 2006, Schneider et al. 2002, Sugiura et al. 2009), ocean advection processes (Saravanan and McWilliams 1998), or tropical-extratropical interaction (Gu and Philander 1997), only a few have attempted to quantified its predictability limit. Our recent ensemble experiments indicate that the PDO, which is EOF1 of both SST and the subsurface temperature intrinsic variability in the model we studied, is only predictable for less than 6 years (Teng and Branstator 2010), but EOF1 has a tendency to evolve to EOF2 through ocean advection. In combination these patterns form a mode that is predictable for more than a decade. The Southern Ocean is another region with possible decadal predictability (Boer 2000), but quantitative estimates are even scarcer. Though previous studies like these have provided valuable information about decadal predictability, they are inadequate in several respects. Our investigation is designed to ameliorate some of these shortcomings.

[1] This experimental approach makes it clear that studies about initial value predictability are not about model skill; indeed since it only involves comparisons of two model generated distributions, model errors are not being measured. Sometimes the term "perfect model assumption" is used to describe this approach.

Thus far most predictability studies have been carried out under equilibrium conditions in which the boundary conditions that control climate are held fixed. Hence only a few studies (Collins and Allen 2002, Boer 2009, 2010) have provided some indications of the second limit of initial-value predictability that interests us, namely the forecast range at which the influence of changes in forcing becomes larger than the influence of the initial conditions. When considering global mean temperature, Hawkins and Sutton (2009) suggest that the forced response gives more reliable information than do the initial conditions during the first predicted decade. On the other hand, Troccoli and Palmer (2009) and Latif et al. (2006) suggest that when regional or modal variables, respectively, are analyzed, predictability from the initial conditions may be more important than the forced response, for a decade or longer. These studies have only begun to evaluate the relative importance of ocean initial conditions and increasing greenhouse gases in decadal predictions. Here we have analyzed several Community Climate Model version 3 (CCSM3) ensemble experiments specifically designed to make it possible to address this issue. We use two different forcing, namely the SRES A1B and Commitment scenarios (Meehl et al. 2006). In addition to enabling us to quantify the two limits of initial-value predictability, these experiments allow us to assess whether the changing forcing impacts the duration of initial condition influence.

Another inadequacy in previous investigations concerns the measure used to quantify predictability. Many studies have investigated predictability by concentrating on the rate at which forecast distributions spread (e.g. Groetzner et al. 1999, Collins and Allen 2002, Pohlmann et al. 2006) while other studies have focused on the pace at which the ensemble mean signal weakens (e.g. Newman et al. 2007 and Alexander et al. 2008). Taken together the conclusion from these two types of investigations is that both factors make significant contributions to initial-value predictability. We take into account both factors (Section 3b) by using relative entropy (Kleeman 2002) to measure predictability. Furthermore, relative entropy has the advantage that it can measure predictability for multivariate states. As Teng and Branstator (2010) have pointed out, propagating phenomena contribute to decadal predictability and their predictability is difficult to assess with univariate measures.

A third inadequacy of many predictability investigations stems from decadal predictability limits being sensitive to the variables used to define the system state. SST is usually used as the indicator of the state of the ocean because it clearly has the potential to affect the atmosphere. But one might expect subsurface fields in the mixed layer, which are somewhat shielded from weather noise, to be more predictable and yet might still have the potential to affect the atmosphere on long time scales. Reported numerical experiments (e.g. Griffies and Bryan 1997a) support the contention that subsurface quantities are more predictable than SST. In Section 3a our study compares the predictability of SST with the predictability of layer mean temperature in the upper 300m and concludes that the latter is the superior field for isolating predictable fluctuations on decadal timescales. For this reason much of our study focuses on subsurface, depth-averaged temperature.

Through its experimental design, use of relative entropy and selection of state variable, our study contributes to the ongoing research into decadal prediction and specifically to a quantification of the two limits that influence the range of skillful prediction (Section 3). But for various reasons there are inherent uncertainties attached to the estimates of predictability that we find. In Section 4 and Appendix A we explore these uncertainties. Section 5 summarizes our results and discusses their implications including our finding that, in the model we have analyzed, for the world ocean initial value influence becomes undetectable after about a decade and becomes less important than the forced response at an even shorter range. This is true even though our approach maybe be biased toward giving optimistic limits of predictability.

2. Methods and measures

2a. Model and experiments

The model we have used for our study, CCSM3, is a fully coupled model that includes four components: atmosphere, ocean, land and sea ice (Collins et al. 2006). These components are linked via a flux coupler and no flux adjustments are employed. We use a version of CCSM3 that has a T42 atmospheric module and a nominal 1˚ ocean module. Though the climate of CCSM3 is similar to the climate of nature in many respects (Alexander et al. 2006), it is not a perfect match. Perhaps most pertinent for our study are differences between the structure and temporal characteristics of some of its prominent oceanic modes of variability and estimates of these quantities that have been derived from the short observational record (Danabasoglu 2008, Alexander et al., 2006). As explained in the introduction, this means that our study of CCSM3 predictability concerns limits on the influence of initial conditions in this model and not necessarily in nature.

Much of our analysis concerns two ensemble experiments that differ only in the anthropogenic forcing: one is forced by the SRES A1B scenario and the other has the forcing fixed at the year 2000 level (Meehl et al. 2006). Both ensembles are integrated for 62 years. We refer to them as the 'A1B' and the 'Commitment' ensemble, respectively. The A1B ensemble has 40 members whose initial ocean, land and sea ice conditions are identical and are equal to the January 1, 2000 state from a CCSM3 20th century historical experiment. The 40 atmospheric initial states come from different days in December 1999 and January 2000 from this same historical experiment. More details for the A1B ensemble are available in Teng and Branstator (2010). The Commitment ensemble also has 40 members and uses the same initial states as the A1B ensemble.

Clearly the duration of initial condition influence is different for initial conditions taken from different positions on the climate attractor, but taking into account this fact is difficult to do in a systematic fashion. In the current study we have taken the commonly applied approach to this issue (e.g. Collins and Sinha 2003) of considering several ensembles, each starting from initial conditions that are well separated from the ensemble of initial conditions used in the A1B and Commitment experiments. The choice of initial states is given in Section 4 while their structure and the method used to generate individual realizations have been described by Teng and Branstator (2010).

One further experiment that we have made use of is a 1000 year control integration (Bryan et al. 2006) of the same model used in the ensemble experiments but with forcing fixed at 1990 values. We have examined the last 700 years, after spinup has occurred, to determine the statistics of our system prior to initiation of the ensemble experiments.

2b. Initial-value predictability and forced predictability

When quantifying predictability for situations where the system climate is reacting to changing external conditions, it is helpful to think of two distinct time-evolving distributions. $P_e(t)$ is the distribution of predicted states resulting from marching a specific initial distribution of states, $P_e(0)$, forward in time. In Fig. 1a this distribution is depicted schematically by the green region. The second pertinent distribution, $P_c(t)$, represents the time-evolving reaction to the forcing and is independent of any particular initial state. One can estimate it from an ensemble of realizations, each beginning long before t=0 and each experiencing the same time dependent external forcing. The red region in Fig. 1a is such an ensemble. As Leith (1975) has explained, for situations where forcing is not fixed, the appropriate definition of climate corresponds to the statistics of the distribution $P_c(t)$ at any given time rather than the conventional view of climate being defined in terms of the statistics of a single realization during a long time interval. Eventually, assuming the system is transitive, as the influence of the particular initial condition is lost, $P_e(t)$ converges to $P_c(t)$. Measures of 'total predictability' deal with comparing $P_e(t)$ with $P_c(0)$, but it is also informative to consider two other comparisons that contribute to predictability. One is a comparison of $P_e(t)$ to $P_c(t)$; it represents 'initial-value predictability', which is the focus of our study. The second is a comparison of $P_c(t)$ to $P_c(0)$; it corresponds to 'forced predictability'. Note that depending on the exact measure used the two components may not add up to the 'total predictability', but these two components do provide a means of comparing the relative strength of the effects of the initial value and forcing.

For our investigation the CCSM3 ensembles described in the previous subsection give us an approximation to $P_e(t)$. Throughout our study we use annual mean quantities to define such distributions. For example the red dots in Fig. 1b show an approximation to $P_e(t)$ for annual mean depth-averaged upper-300m ocean temperature (which we denote 'T0-300') in a small box in the North Atlantic in the A1B experiment. On the other hand we do not have direct information about $P_c(t)$. Ideally $P_c(t)$ would be estimated from a large ensemble of 20th century climate simulations and their A1B extensions into the 21st century. In our case only a single realization of these computer intensive experiments is available. To deal with this problem we have used the following approach. As explained shortly, in our study we have assumed that distributions are well approximated by Gaussians. Furthermore, we have made two additional assumptions. First, the covariance structure of the system *climate* does not change as the forcing changes and is equal to the covariance structure in our control experiment. In the schematic this corresponds to only the mean of the red distribution changing with time. Second, the evolution of the *climate* mean of a given state variable can be well approximated by an analytical function of time whose parameters can be determined from forecast behavior after the effect of the initial condition is weak. Using this function and the assumption of unchanging covariances leads to an approximation of $P_c(t)$ at all forecast ranges.

The functions that we have assumed are good approximations to the time evolving climate mean (i.e., the means of $P_c(t)$) are the linear function

$$\overline{T}_{A1B}(t) = \overline{T}_{1999} + k(t - 1999) \qquad (1)$$

for the A1B experiment, and the exponential function

$$\overline{T}_{commit}(t) = \overline{T}_{1999} + A(1 - e^{-t/\tau}) \qquad (2)$$

for the Commitment experiment. In these expressions model year *t* varies from 2000 to 2061. We have estimated constant *k* by least squares fitting a line to A1B ensemble values during 2010-62. Evaluating that line for *t*=1999 gives $\overline{T}_{1999}$. Next we have inserted the resulting $\overline{T}_{1999}$ in (2) and calculated τ and A by least square fitting of Commitment values during 2010-2061.

Note that measures of initial-value predictability concern departures of the red and blue dots in Fig. 1b from the red and blue lines, respectively, which depict the climate means ($\overline{T}_{A1B}$ and $\overline{T}_{commit}$). We refer to these departures as the 'initial value components' of a forecast to distinguish them from the 'raw' forecast states. When initial-value predictability is lost the mean of the initial value components is zero and their distribution about their mean matches the distribution of $P_c(t)$ about its mean. In Fig. 1b early in the forecasts the distribution of the initial value component is clearly distinguishable from the climatological distribution both because of the narrowness of its spread and the separation of the ensemble mean from the climate mean. The forced predictability can be measured by the departures of $\overline{T}_{A1B}(t)$ and $\overline{T}_{commit}(t)$ from $\overline{T}_{1999}$

because the spread contribution to the distribution difference is negligible under our first assumption.

While Fig.1b suggests the coexistence of initial-value predictability and forced predictability for regional variables, the situation is different for global mean temperature (Fig.1c). The ensemble means are not very different from the climatological means, and the spreads of the dots do not increase dramatically in the first decades. Both suggest that in CCSM3 the initial conditions are less important for predicting global mean temperature, though the presence of year-to-year variability in estimates of observed global mean temperature (Brohan et al. 2006) suggests that internal processes do have some impact on this quantity in nature.

Some of the choices we have made when separating the initial value and forced components of our ensemble forecasts cannot be strictly justified. For example, to estimate the time evolving mean forced response we have assumed the influence of the initial conditions is relatively small after 2010. As Appendix A explains, our results are not completely insensitive to this decision, but it appears to be a reasonable compromise choice. A second choice that is not strictly valid is our decision to assume that covariances are not affected by forcing changes. Meehl et al. (2006) point out that ENSO variability weakens in CCSM3 in reaction to increased GHG concentrations, and in results not shown here we have found that even larger changes in variability in our A1B experiment occur in the North Atlantic near the end of the integrations. But as we will present in the next section, our major interests, the two limits of initial-value predictability, both occur in the first 1-2 decades, and even in the North Atlantic variability reactions to GHG changes are so weak at this range as to have no discernable effect on our results.

2c. Relative entropy

As described by Kleeman (2002) and Majda et al. (2005), relative entropy is a means of comparing a distribution to a baseline distribution and thus can be employed in predictability studies where one wants to determine whether and by how much a forecast distribution differs from a climatological distribution. In full generality, the relative entropy of the distribution P_x relative to the baseline distribution P_b is

$$R = \int_{\mathbb{S}} P_x(s)\log_2\left(\frac{P_x(s)}{P_b(s)}\right)ds \qquad (3),$$

where s is the state and $\mathbb{S}$ represents the system state space. In addition to the advantages mentioned in the introduction, relative entropy has a well defined interpretation. It represents the information, in binary bits, in P_x (say, forecast $P_e(t)$) over and above the information one would have if one knew no more than a state was a member of P_b (say, climatology $P_c(0)$). A common application of relative information concerns the average number of bits that are required to represent the current state of a system. If the system has distribution P_x, this number will be larger if one uses a system of representing (in information parlance 'coding') states that was developed under the assumption states are drawn from distribution P_b than if one uses a system based on knowledge that the actual distribution is P_x. Relative entropy is the number of extra bits needed when knowledge of P_x (or in our case knowledge of a perfect forecast ensemble) is not taken into account. As an indication of what one bit of information represents, consider a system with a finite number of states, and imagine that they are represented in terms of a binary basis with equal probability of each bit being 0 or 1. If a forecast specifies the actual value of one of these bits, then that forecast has relative entropy of 1. Note that it corresponds to reducing by a factor of 2 the possible configurations of the system. Similarly a forecast that specifies *m* bits has a relative entropy of *m* and corresponds to reducing the number of possible states by a factor of 2^m. In the more general application we make of relative entropy, it need not take on integer values, but it continues to be a measure of how much more precisely we know the state of the system as a result of having a (perfect) forecast ensemble.

In our study we represent the system state by a vector of finite length, *n.* Since extremely large samples are required to estimate general multivariate distributions of even modest dimension, we have approximated our distributions by Gaussians. In this case, (3) becomes

$$R = \frac{1}{2}\log_2(e)\left[\ln\left(\frac{\det(\sigma_b^2)}{\det(\sigma_x^2)}\right) + trace(\sigma_x^2 / \sigma_b^2) + (\mu_x - \mu_b)^T(\sigma_b^2)^{-1}(\mu_x - \mu_b) - n\right] \qquad (4),$$

where μ_b and μ_x stand for the mean state vectors in distributions P_b and P_x, respectively, while σ_b^2 and σ_x^2 corresponds to covariance matrices representing relationships between elements of state vectors in these same distributions. One feature of this approximation is that it can be decomposed into contributions from the mean, namely the third term in the brackets, and from the covariances, which are the rest of the terms. Commonly these are referred to as the *signal* and the *dispersion* components, respectively (Kleeman 2002). Note that while Kleeman (2002) and Teng and Branstator (2010) chose to express values of relative entropy in base *e*, here we use base 2. We do this because we are more accustomed to mentally raising 2 to a power than *e* to a power. This choice leads to the factor of $\log_2(e)$ in (4) that does not appear in expressions in those papers.

2d. Basins and bases

As we have explained, when considering predictability it is best to be able to represent propagating phenomena. This implies a field representation of variables should be used. On the other hand we wish to be able to distinguish predictability characteristics in different geographical locations. Together these factors suggest it is desirable to measure predictability in different basins of the ocean.

To decide on the boundaries of the basins we have been guided by local timescales of intrinsic variability in the CCSM3 control experiment. As the four examples at the bottom of Fig. 2 demonstrate, when we have examined variance spectra of T0-300 at various locations we have found a broad range of behavior. For example, variance spectra in the North Pacific and Tropical Atlantic locations shown in that figure are essentially red but with different decay rates, the North Atlantic point also has pronounced low-frequency variability but with two distinct peaks, and the Tropical Pacific point has a pronounced peak with a period of two years (Collins et al. 2006). As a means of visualizing the geographical dependence of the spectra we have summarized each one in terms of a single characteristic time scale. If V(*f*,*x*) is variance per unit frequency for frequency *f* at location *x* (the quantity plotted in the lower panels for Fig. 2), then we assign variability at *x* the time scale

$$\tau_x = \sum_k V(f_k, x) / \sum_k f_k V(f_k, x),$$

where *k* points to each of the frequencies we can resolve in the control experiment. These time scales are simply variance-weighted mean frequencies (expressed as cycles per year) transformed to a period. They are plotted in the top panel of Fig. 2. They make clear the much shorter timescale behavior in the tropics compared to the extratropics, a contrast that may be enhanced by the underrepresentation of tropical/extratropical connections on decadal timescales that has been found in CCSM3 (Alexander et al. 2006). Reasoning that prominent propagating features are unlikely to cross between regions with very different timescales, we have felt justified in choosing basin perimeters based on continental boundaries and the timescale separation of tropical and extratropical regions. The eight basins determined in this way that we use to partition the world ocean are outlined in this figure. In Sections 3 and 4 we describe predictability in the North and Tropical Pacific and Atlantic basins while in Section 5 we summarize our results in terms of global statistics that take into account all eight regions.

When calculating relative entropy, we have chosen to represent fields in terms of EOF bases. These bases are calculated for each basin from control run variability. One constraint on the appropriate EOF truncation is that for the covariance matrices in (4) to be nonsingular, state vectors can be no longer than the sample size minus 1. Hence our state vector of principal components cannot be larger than 39. And to guard against including variability that is too weak to be estimated with adequate accuracy we have chosen to make *n* even smaller. Indeed, in the following section our calculations of relative entropy employ fields that are represented in terms of the leading 15 EOFs in each region. (For T0-300 these correspond to 73-89% of the variance in the control run depending on the region.) In section 4 we report on the sensitivity of our results to this choice of truncation.

3. Results

3a. Spread and mean

Before using relative entropy, we examine two, more familiar, gridpoint based indicators of initial-value predictability. The first of these is RMSD, the square root of the regionally averaged squared difference between all combinations of realization pairs within an experiment. RMSD has the same value for the raw states and for the initial-value components. Using this quantity we see for how long the spread in the A1B and Commitment ensembles remains statistically distinct from random states in the control.

Figure 3 shows RMSD for T0-300 for A1B (blue solid) and Commitment (blue dashed) in each of four basins. It also shows the spread of SST in red. As expected, in all cases the spread initially increases and eventually converges to control values, but the forecast range at which convergence happens is in general very different for SST and T0-300. In all but the Tropical Pacific basin, the effects of the initial distribution are detectable much longer for forecasts of T0-300 than for SST. (Calculations not shown indicate the similarity of convergence rate for these two variables in the Tropical Pacific is a reflection of the strong covariability of these fields there.) A second interesting characteristic in these plots is the large variations among the basins. The longer intrinsic time scales of the midlatitude basins seen in Fig. 2 appear to go hand in hand with longer saturation times. Based on the 95% significance threshold indicated by dashed horizontal lines in Fig. 3, in the two midlatitude regions the ensembles remain distinguishable from random states for 7 to 11 years, while for the tropical regions predictability lasts 2 years for the Tropical Pacific and 6 years for the Tropical Atlantic. The somewhat longer duration in the North Atlantic compared to the North Pacific is a common feature in predictability studies though often (e.g., Collins 2002) the contrast is more pronounced than in Fig.3.

Turning to the second contributor to initial-value predictability, namely the ensemble mean of the initial-value components, we calculate its RMS amplitude in each basin at various forecast ranges in the A1B and Commitment ensembles (Fig.4). The evolution of RMS amplitude has qualitative similarities to the evolution of RMSD in Fig. 3. It converges more or less monotonically to the mean value for averages of 40 random states from the control run. The range at which convergence occurs varies from basin to basin though not as dramatically as RMSD. And SST loses its influence somewhat sooner than T0-300 because of SST's more pronounced intrinsic variability as reflected in contrasting significance thresholds. One pronounced distinction from RMSD (Fig.4) is that in every basin convergence occurs noticeably later for this measure.

3b. Relative entropy

The results based on spread and ensemble mean amplitude do not measure their combined effect, their relative importance, or the amount of information in a forecast before its predictability disappears. To address these issues we use relative entropy. Given the relatively short predictability of SST found in the previous section, we only consider T0-300.

When we plot relative entropy for the raw states in the A1B (black solid line in Fig. 5) and Commitment (black dashed line) ensembles as a function of forecast range, based on comparing $P_e(t)$ for each experiment to the year 1999 climatology ($P_c(0)$), we find they have a distinctive U-shape. This shape occurs in all basins. Presumably this is a reflection of a) an initial period during which predictability from the initial state, with its gradual loss of information through increasing spread (Fig. 3) and decreasing amplitude of ensemble mean anomalies (Fig. 4), dominates and b) a succeeding period during which an ever increasing forced response dominates.

In order to quantitatively compare the relative contributions of predictability from the initial state and from the forced response, we use the approach described in Section 2b which entails calculating relative entropy of $P_e(t)$ relative to $P_c(t)$ and of $P_c(t)$ relative to $P_c(0)$, respectively. As depicted by the blue lines in Fig. 5, initial-value predictability is virtually the same in the two experiments. Indeed, in both experiments it loses its significance at the same range. We designate the range at which this happens, indicated by the blue lines crossing the red dashed 95% significance line in Fig. 5, the 'saturation' range for predictability.[2] This is the first of the two limiting times on initial value influence mentioned in the introduction.

By comparison, information resulting from the forced response is by definition very small at the beginning of the experiments (green lines in Fig. 5) and becomes significant after a few years. We call the range at which it is significantly different from

[2] The value of relative entropy does not asymptote to zero in these calculations because we are using finite ensembles. Even when the forecast represents a distribution that is identical to the background distribution, estimates of covariances and means do not match the background values exactly, and because relative entropy is positive definite, this leads to positive values of relative entropy. We calculate a distribution of relative entropy values for randomly drawn ensembles of a given size from the control and use its 95th percentile as the 95% significance level.

control random states the range of ‘emergence’. Eventually its relative entropy surpasses that from the initial condition. This happens 4 to 9 years into the forecasts, depending on basin, at a range we designate as the ‘crossover’ range. This range is the second limiting time referred to in the introduction. Although after about a decade the forced response of the A1B ensemble has somewhat higher relative entropy than the Commitment ensemble, the ‘crossover’ range is about the same in the two experiments. Late in the forecasts, beyond approximately year 15, the difference in forced predictability resulting from GHGs increasing in the A1B experiment and not in the Commitment experiment becomes very apparent, but this is beyond the limits of initial-value predictability we are investigating.

Figure 5 shows marked distinctions in the predictability properties of the various basins with the contrast in saturation times between the North Atlantic and Tropical Pacific being particularly strong. Previous studies have speculated that variability induced by the initial state is potentially more predictable at higher latitudes than in the Tropics based on decadal variability being more prominent in the extratropics (Boer 2000). Our results provide a quantitative estimate of this contrast, with saturation for the North Atlantic and Tropical Pacific differing by about five years. Another distinction is that the forced predictability ‘emerges’ within about two years in the tropical basins compared to 4-5 years in the extratropical basins. This agrees with Boer’s (2010) conclusion that the forced response to GHG increases is more predictable in the Tropics than at higher latitudes. The faster emergence probably results from there being smaller intrinsic interannual variability in the Tropics, as one can see from the geographical distribution of T0-300 variance in our control experiment (not shown). Similarly, the tropical crossover range is about 3 years shorter than the extratropical crossover range. It should be noted that the emergence and crossover ranges that we estimate here are for basin scales, and it may take decades for the forced predictability to emerge at smaller scales (Karoly and Wu 2005, Knutson et al. 1999).

Contrasts between basins within similar latitudinal bands are less pronounced but not trivial. One way to measure this is to take advantage of relative entropy’s ability to quantify the information resulting from the initial state at each stage of an ensemble forecast. The horizontal dotted lines in Fig 5 indicate the value of relative entropy that corresponds to three more bits of information than exist at the saturation threshold. (Recall in the simple situation of a system with a finite number of equally probable binary states, this would mean a reduction by a factor of 8 in the number of forecast states compared to the number of climatological states.) By comparing the blue curves in Fig. 5 to the dotted lines, we see that in midlatitudes and in the Tropics at least 3 bits of information remain in the two Atlantic basins two to three years longer than in the Pacific basins.

3c. Structure of mean anomalies

Because the practical value of forecasts could depend on what aspects of the forecast distribution contain information, we have further decomposed the relative entropy for initial-value predictability into its dispersion and signal components (Fig. 6). In all basins except the North Pacific, we find that the signal component saturates after the dispersion component. We saw this same behavior in Section 3a when considering RMSD and RMS amplitude as indicators of predictability. Here we use the identical measure for both aspects of initial-value predictability so that we can not only compare the saturation times of the two components, but we can also quantitatively compare their information content at a given range. For example the greater predictability in the mean anomalies compared to the spread in the North Atlantic can be measured by the fact that at year 2005 the signal has about 3 bits of significant information while the dispersion component has only about 2.

Since we find ensemble mean anomalies contain at least as much information as ensemble spread it is of interest to examine the structure of these mean anomalies. Ensemble averages of the initial-value components of the A1B (right) and Commitment (left) ensembles at five year intervals are shown in Fig. 7. At a range of 5 years, there are T0-300 anomalies with amplitudes between 0.2°C and 0.3°C in all regions. Many of these features are statistically significant when compared to averages of 40 random anomalies from the control experiment (stippling). The similarity between anomalies in the two experiments is striking, another indication of statistical significance. At year 5 the fields in Fig. 7ab have a pattern correlation of 0.88. This similarity also shows that the insensitivity to forcing scenario that we noted earlier in basin scale measures (Figures

3,4,5,6) is actually a reflection of insensitivity on a much finer scale. As expected from the basin scale saturation times, the situation is very different at a forecast range of 10 years (Fig. 7c,d). Mean anomalies have weakened so much that it would be difficult to establish field significance though in a few regions, particularly the North Atlantic and Southern Ocean, similarities between the two experiments remain. By year 15 only the North Atlantic retains strong features. When intermediate years are examined, these anomalies correspond to a counterclockwise rotating combination of anomalies between 40N and 60N whose statistical significance is bolstered by its presence in both ensembles.

Next we examine the time evolving forced response relative to the year 1999 climate state at a range of 5 and 10 years (Fig.8). There is very little difference in the amplitude or structure of the forced response between the two experiments. For our investigation, what is more important is that the forced anomalies are strongest in extratropical regions that also have strong unforced features. These regions include the North Pacific north of 40N, and the North Atlantic east of Newfoundland and in the GIN Sea. It is only at a range of about 10 years (Fig. 8c,d) that the mean initial value component (Fig. 7c,d) has become weak enough for the forced response to become dominant. But even at this range there are some extratropical initial-value produced features that are of approximately equal amplitude to the forced response. Hence the basin measures of crossover appear to underestimate the importance of initial-value predictability at some locations on smaller spatial scales.

4. Sensitivities and robustness

In arriving at the estimates of predictability and its limits, we have made a number of choices that could affect our results. To determine how robust our findings are to these choices we have carried out a number of tests.

One source of uncertainty in our results is the composition of the state vector. We have seen very large sensitivity to our results depending on whether we use T0-300 or SST as the state variable. Even if we settle on layer mean subsurface temperature there is a question as to what is the appropriate layer to use. We have reasoned that a layer that largely represents the mixed layer is a reasonable compromise choice. But with annual mean mixed layer depths varying between tropical values of less than 100m and extratropical values of 500m or more, it is possible that in some regions our results would be strongly affected if another layer thickness were used. When, however, we have employed layers ranging from 0-100m to 0-500m we have found little sensitivity in the extratropical basins with saturation times and relative entropy values leading up to saturation being affected only modestly. Figure 9a shows this behavior for the North Pacific. It is only when we consider a layer extending to 1000m, well below the mixed layer, that we find substantial increases. By contrast, for tropical regions, with their much shallower mixed layers, our choice of the 0-300m layer has had a substantial effect. As depicted in Fig. 9b, in the Tropical Pacific the saturation time is reduced from the T0-300 value of 7 years to a value of 3 years when T0-100 is employed. (Further reduction to a 50m layer has little effect (not shown).) This suggests that our tropical results may overstate the predictability of that portion of the ocean that readily communicates with the surface, and that the contrast between extratropical and tropical predictability mentioned in section 3b is probably even greater than indicated by our T0-300 results.

Another aspect of the state vector that can affect predictability results is its truncation. One might speculate that a more severe truncation than the one we have employed could enhance predictability given that the leading EOFs of geophysical fields often have longer intrinsic time scales than trailing EOFs. To check this possibility we have calculated relative entropy for the A1B experiment truncated to 5 EOFs. The North Atlantic region (Fig. 10a) has behavior typical of most basins. Of course, given the form of (4) it is not surprising that all relative entropy values are smaller than for the case of 15 EOFs, which is also included in the plot. We take this into account in a crude fashion by stretching the scale for the 5 EOF truncation by a factor of three, to make up for the factor of three difference in degrees of freedom that can contribute to information. When this is done it is apparent that information content, saturation time and crossover range are similar for the two truncations. The behavior for the Tropical Pacific (Fig. 10b) is more difficult to decipher quantitatively given the small values of relative entropy resulting from weak projections onto the leading EOFs in this region. The one dramatic effect of the severe truncation is the weakness of the forced response so that forced predictability practically disappears with this truncation. And though the exact saturation range is murky, it is apparent that restricting the state vector to the leading patterns has not enhanced initial value predictability.

A second characteristic of our analysis concerns the confidence we can have in our relative entropy estimates when they are based on a single finite ensemble. The shading in Fig. 11 depicts the range of values of relative entropy that result from 10000 random draws of 10 (top panel) and 20 (bottom) member subensembles from our 40 member A1B experiment for the North Pacific. (Here we have employed a truncation of 5 EOFs rather than our standard 15 so that we can calculate relative entropy for small ensembles.) In the figure the range is given by the 5 and 95 percentile boundaries while the solid line shows the mean at each forecast range. We see that for 10 member ensembles there is a broad range of possible relative entropy values because of sampling fluctuations. For long range forecasts the span of possible bits of information is around 8 while for the first few forecast years when the ensemble is tightly bunched the span is much smaller. This uncertainty produces a range of saturation years that is about 9 years wide. By contrast, for 20 member ensembles' relative entropy is confined to a range of about 4 1/2 bits for extended forecasts and the saturation year is confined to about 5 possible years. Of course for our full 40 member ensemble, uncertainty will be reduced still further, one of the benefits of using a large ensemble. Another worthwhile consequence of using a large ensemble for predictions is that the threshold for significance is reduced as a result of finite sample size error in R being reduced. This will tend to expand the range that forecast distributions can be distinguished from the control distribution. It also means that the emergence of the forced signal from the intrinsic noise can be detected earlier.

The third factor is the sensitivity of initial-value predictability to the structure and amplitude of the ocean initial conditions. As mentioned in Section 2a, here we address this factor in an expedient fashion; we simply repeat our analysis for three additional 40 member ensemble experiments with initial states different from the initial state in the A1B and Commitment experiments and determine how much our plots of relative entropy as a function of forecast range change. These experiments, which include GHG and aerosol concentrations from the SRES A1B scenario, are referred to as experiments 'A1B(II)', 'A1B(III)', and 'A1B(IV)'. A1B(II) and A1B(III) were originally generated for Teng and Branstator's (2010) study of North Pacific modal predictability and were chosen to differ from each other and from our standard experiments in the way they project onto the leading intrinsic mode in that region. Specifically A1B(II) is an ensemble of perturbed integrations starting from the January 1 state in year 2008 of one member of our standard A1B experiment that has a very strong positive projection onto the Pacific Decadal Oscillation. A1B(III) starts from a different member of our standard A1B experiment in January 1, 2008, but here the member has a strong projection onto the second leading pattern of North Pacific variability. Similarly, A1B(IV) starts from a third member of our standard A1B experiment but for January 1, 2010, when this member has a strong projection onto the second EOF of the Atlantic Meridional Overturning Circulation.

When we examine plots of relative entropy for these additional cases (Fig.12), we see that in a few instances there are cases that stand out. The most prominent example is the strong North Pacific predictability in the A1B(II) case. In a more thorough analysis (Teng and Branstator 2010), we have found that this high predictability results from the very large initial projection onto the leading basin EOF of intrinsic T0-300 variability. Indeed the initial projection is in the 99th percentile and produces a signal component in the forecast that lasts for at least a decade. A second example of an outlier is case A1B in the Tropical Atlantic with its unusually late saturation. This is noteworthy because the other three cases are more consistent with the idea referred to in Section 3b that tropical regions have lower initial-value predictability than extratropical regions. With these few exceptions, we find a surprising lack of sensitivity of the saturation range to the initial conditions. This insensitivity contrasts with previous studies that examined the Atlantic overturning circulation (Collins et al. 2006) and found initial conditions with a strong overturning circulation had enhanced predictability. The discrepancy may be caused by model differences, insufficient sample size in the previous work, the different state vectors used to represent the system, or simply the small number of initial states used in both studies.

5. Summary and discussion

Motivated by the current interest in decadal climate predictions and the hypothesized improvement in these predictions that might result from ocean initialization, we have quantified the forecast range that initial states can potentially influence in forecasts made with a coupled climate model similar to models that will be used in decadal prediction studies. We have done this by analyzing the evolution of ensembles of initially similar states.   To summarize our main ideas and results, we plot in Fig. 13 relative entropy values that quantify predictability properties of the two climate change experiments that were the focus of our study. The values in this figure are the sum of relative entropy[3] in eight basins (Fig. 2) that span the world ocean. The relative entropy for raw forecasts (black curves) has the characteristic U shape that we also saw for individual basins (Fig. 5). When we separate this into initial value (blue) and forced (green) influences, we find the decrease in relative entropy at the beginning of the forecasts corresponds to the loss of information from the initial state one expects in a chaotic system while the increase after 6-7 years results from the influence of external forcing. These two processes prompted us to use two time scales to quantify the limit of influence from ocean initial states. The first is the time at which information from the initial value becomes undetectable. We referred to this as the *saturation* range, which for the global ocean is about 12 years in our experiments. The second time scale is the range at which initial condition information becomes smaller than the information that results from external forcing. We called this the *crossover* range; it occurs at year 7. By using relative entropy we can also quantify the information provided by the initial value at any point in the forecast. For example, the initial state provides at least 10 bits of information for almost a decade.

While most previous predictability studies have focused only on the saturation range, because of the design of the experiments we analyzed, we were able to recognize

[3] Strictly speaking, relative entropy is not additive unless one is combining values derived from state vectors whose elements are uncorrelated, but in calculations that are beyond the scope of this paper, we have found that adding together relative entropy values from our separate basins is a good approximation to the true relative entropy found from a combined state vector.

the importance of the crossover range when considering forecasts in a global warming context. Although, like several other studies, we found the ocean initial condition may provide predictability for a decade or so, the crossover range suggests that for some SRES forcing scenarios after 7 years the external forcing can bring more information than the ocean initial state. Consequently, in the model we have analyzed, 10-30 year predictions fall into the category of “boundary condition problems” rather than “initial value problems” if predictive skill comes solely from upper ocean heat content. Therefore, predictions in this range must rely on accurate estimates of future external forcing, rather than on estimates of the present ocean state.

Another significant finding from our study results from considering both mean and spread when we measured predictability. As discussed in Sec.3b, for most basins more information is contained in the mean than in the spread of predicted distributions after the first one or two years. Indeed, based on the results in Fig. 13, for the global ocean between year 2 and the year of saturation, 70% of the information is in the ensemble mean and only 30% in the spread. This result suggests that the many investigations that have focused on spread in their assessment of decadal predictability have neglected a major contributor.

Another potential advance in our study is the variable used to represent the state of the climate system. We found subsurface temperature is more predictable than SST. One way to quantify this difference is to note that if we redraw Fig. 13’s summary of predictability for the global ocean but use SST rather than T0-300, we find saturation occurs about 3 years earlier and crossover happens 1 year earlier. In Section 4 we reported even greater initial-value predictability when we used layers that were deep enough to extend below the mixed layer. In preliminary research we have found evidence that on decadal timescales predictability in the mixed layer is associated with predictability of surface conditions in CCSM3. Whether predictability produced below the mixed layer has this same property is open to question.

Any predictability study that uses ensemble experiments as its basic methodology suffers because it must draw conclusions using just a small number of cases. Our investigation has this shortcoming, but to the extent we have been able to test the robustness of our results, we have found them to be insensitive to four key factors. First, as Fig. 13 shows, we found both limits of initial-value predictability to be insensitive to the two scenarios used to drive the model. This insensitivity was aided by the similarity in the forced response during the first 10 years of prediction (Fig. 8) and by the fact that initial-value predictability did not last much longer than a decade. Second, predictability properties were rather insensitive to the four particular initial ocean states used for our ensemble experiments. The only situation for which we found a very pronounced departure in saturation range was in a case that had an extremely high initial anomaly in the North Pacific. Third, our results suggest predictability is not dramatically higher for the most prominent intrinsic patterns of a basin than for patterns that explain somewhat less variance. This insensitivity is consistent with Teng and Branstator's (2010) finding that the leading intrinsic propagating mode in CCSM3's North Pacific does not have unusually high predictability. But given the high amplitude of the leading modes this result does not rule out their important role in predictions. Fourth, in results not described earlier in the body of our paper, applying temporal filtering to T0-300 (e.g. 5-year or 10-year running means) does not alter our major conclusions. Indeed when running n-year averages are employed saturation ranges tend to be extended by no more than about n/2 years. This increase in the saturation range is simply what is to be expected from predictable years being including in running averages for n/2 years after saturation occurs in the raw fields. Time averaging may produce more dramatic increases in investigations where smaller ensemble averages are used than we have employed because then it serves the purpose of reducing the unpredictable components that one would prefer to remove through ensemble averaging. Similarly, our use of depth-averaged quantities means that time-averaging has less of an effect on our results.

By contrast, we did discover two factors that do influence estimates of decadal initial-value predictability. First, there are substantial variations in the predictability properties of different basins. The contrast between higher predictability in the extratropics and lower predictability in the tropics was noteworthy in most basins and experiments. This could be associated with large basin-to-basin variations in intrinsic timescales (Fig. 2). We also noticed substantial variations in predictability on sub-basin scales. Second, predictability limits, when they are defined in terms of statistical comparisons of finite ensembles to a background distribution, are affected by ensemble size. This is particularly true for the saturation and emergence times and happens because the smaller the ensemble, the larger will be the relative entropy from randomly chosen states (Fig. 11).

As we emphasized in the introduction and in Section 2a, when considering our results one should remember that predictability is an inherent property of a dynamical system and thus our results are valid only for CCSM3 and not for other models or for nature. But in conjunction with predictability estimates for other models, our estimates help to define the range of possible predictability behavior that nature may have. Also, no matter what nature's predictability is, the inherent limits of any model will affect the limits of its skill. Therefore these limits should be kept in mind when designing and interpreting the results of decadal prediction experiments like the upcoming Coupled Model Intercomparison Project Phase 5 (Taylor et al. 2009). In fact, the prudent approach would appear to be for groups engaged in decadal prediction to carry out similar determinations of their model's predictability before undertaking extensive prediction experiments. Also, when applying our results it is best to remember that we have used an unusually large ensemble, we have concentrated on depth-averaged, subsurface conditions, and we have employed a perfect model assumption. All of these factors mean that the two limits of basin scale CCSM3 predictability that we estimated are likely to be longer than its range of skillful prediction. On the other hand, our plots of sub-basin scale features (Fig.7) suggest that on these scales there may be isolated features whose initial-value predictability is longer than the basin and global limits we have emphasized. Moreover, CCSM3's North Atlantic oceanic variability in our long control integration may be influenced more by atmospheric noise than in nature (Danabasoglu 2008) and it peaks at higher frequencies in the decadal range than does the corresponding variability in the short record from nature. Both factors mean the predictability we have found for CCSM3 may be less than nature's predictability in this region, predictability that future models may achieve as they become more realistic.

Acknowledgment

The authors acknowledge helpful comments from three anonymous reviewers, useful discussions with numerous colleagues, and support from the DOE under Cooperative Agreement No. DE-FC02-97ER62402. NCAR is sponsored by the National Science Foundation.

Appendix A The Fitting Interval for the Mean Forced Response

In our method for separating initial-value predictability and forced predictability we must choose a starting point for the fitting interval. Ideally this would be set to the range at which the influence of the initial state has become small enough to not affect the fit. On the other hand the result of the fitting procedure affects which portion of the forecast appears to be influenced by the initial condition. Given this circularity, we have reasoned that the best strategy is to make sure that our results are not sensitive to the starting point of the fitting interval.

One extreme is to use the interval from 2000 to 2061 for the fit. In this case one is underestimating the influence of the initial value because some of the variability it produces may be incorporated into the forced component. The other extreme we have considered is to use the interval 2020-2061. We have seen no evidence that the initial conditions influence this interval but with this choice there is a risk of overestimating the initial value influence because of the large departures from the forced component that may result from the 21 year extrapolation that is involved when estimating $\overline{T}_{1999}$. For Northern Hemisphere extratropical basins we find little difference for this range of starting points. As an example we show in the top panel of Fig. A1 the relative entropy for the initial value component of the A1B ensemble for the North Atlantic when three fitting intervals are used. Clearly the rate of information loss and the saturation range show only small variations for different fitting intervals. For some other basins the choice of fitting intervals makes a discernible difference. For example when we use the interval that starts in 2020 for the Tropical Pacific (bottom panel of Fig. A1), there is a noticeable increase in relative entropy compared to using the other two intervals. Considering that the interval beginning in 2000 is a very conservative choice that certainly underestimates predictability, and given that using the 2010 starting point only increases relative entropy by 1 to 2 bits, we conclude our choice of using the 2010 to 2061 interval for all basins is prudent.

References

Alexander, M., and coauthors, 2006: Extratropical atmosphere-ocean variability in CCSM3. *J. Climate*, **19**, 2496-2525.

Alexander, M., L. Matrosova, C. Penland, J.D. Scott, and P. Chang, 2008: Forecasting Pacific SSTs: Linear inverse model predictions of the PDO. *J. Climate*, **21**, 385-402.

Boer, G.J. 2000: A study of atmosphere-ocean predictability on long time scales. *Climate Dyn*., **16**, 469-477.

Boer, G. J. 2009: Changes in interannual variability and decadal potential predictability under global warming. *J. Climate*, **22**, 3098-3109.

Boer, G. J. 2010: Decadal potential predictability of 21st c. *Climate Dyn*., submitted.

Brohan, P., J. J. Kennedy, I. Harris, S.F.B. Tett, and P.D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.*, **111**, D12106, doi:10.1029/2005JD006548.

Bryan, F.O., and coauthors, 2006: Response of the North Atlantic thermohaline circulation and ventilation to increased carbon dioxide in CCSM3. *J. Climate*, **19**, 2382-2397.

Collins, M. 2002: Climate predictability on interannual to decadal time scales: The initial value problem. *Climate Dyn*., **19**, 671-692.

Collins, M., and M. R. Allen, 2002: Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate*, **15**, 3104-3109.

Collins, M., and B. Sinha B, 2003: Predictability of decadal variations in the thermohaline circulation and climate. *Geophs. Res. Lett*. **30**. doi:10.1029/2002GL016504

Collins M., and coauthors, 2006: Interannual to decadal climate predictability in the North Atlantic: A multimodel-ensemble study. *J. Climate*, **19**, 1195-1203.

Collins W.D., and coauthors, 2006, The Community Climate System Model version 3 (CCSM3). *J. Climate*, **19**, 2122-2143.

Danabasoglu, G., 2008: On multi-decadal variability of the Atlantic overturning circulation in the Community Climate System Model version 3. *J. Climate*, **21**, 5524-5544.

Griffies, S.M., and K. Bryan, 1997a: A predictability study of simulated North Atlantic multidecadal variability. *Climate Dyn.*, **13**, 459-488.

Griffies, S.M., and K. Bryan, 1997b: Predictability of North Atlantic multidecadal climate variability. *Science*, **275**, 181-184.

Gu, D. F., and S.G.H. Philander, 1995: Secular changes of annual and interannual variability in the tropics during the past century. *J. Climate.*, **8**, 864-876.

Hawkins, E, and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull Am Meteorol Soc*, doi:10.1175/2009BAMS2607.1

Hurrell J.W., and coauthors, 2009: Decadal climate prediction: opportunities and challenges. Community White Paper, OceanObs '09 (http://www.oceanobs09.net/globe/?=97).

Karoly, D. J. and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Climate*, **18**, 4337-4343.

Keenlyside N., M. Latif, J. Junclaus, L. Kornblueh, and E. Roeckner 2008: Advancing decadal climate scale prediction in the North Atlantic. *Nature*, **453**, 84-88.

Kleeman, R. 2002, Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci*. **59**, 2057-2072.

Knutson, T. R., T. L. Delworth, K. W. Dixon, and R. J. Stouffer, 1999: Model assessment of regional surface temperature trends (1949-1997). *J. Geophys. Res.*, **104**, 30981-30996.

Kwon, Y.O., and C. Deser, 2007: North Pacific decadal variability in the Community Climate System Model version 2. *J. Climate*, **20**, 2416-2433.

Latif, M., and T. P. Barnett, 1994: Causes of decadal climate variability over the North Pacific and North America. *Science*, **266**, 634-637.

Latif, M., 2006: On North Pacific multidecadal climate variability. *J. Climate*, **19**, 2906-2915.

Latif, M., M. Collins, H. Pohlmann, and N. Keenlyside, 2006: A review of predictability studies of Atlantic sector climate on decadal scales. *J. Climate*, **19**, 5971-5987.

Leith, C. E., 1975, Climate response and fluctuation dissipation. *J. Atmos. Sci.*, **32**, 2022-2026.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos Sci*, **20**, 130-141.

Majda, A,, R. Abramov, and M. Grote, 2005: Information Theory and Stochastics for Multiscale Nonlinear Systems. American Mathematical Society, Providence.

Mantua, N.J., S.R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific decadal climate oscillation with impacts on salmon. *Bull. Amer. Meteorol. Soc*., **78**, 1069-1079.

Meehl, G. A., and coauthors, 2005: How much more global warming and sea level rise? *Science,* **307**, 1769 – 1772.

Meehl, G. A., and coauthors, 2006: Climate change projections for the twenty-century and climate change commitment in the CCSM3. *J. Climate*, **19**, 2597-2626.

Meehl, G.A., H. Teng and G. Branstator, 2006: Future change of El Nino in two global coupled climate models. *Climate Dyn*., doi: 10.1007/s00382-005-0098-0.

Meehl G. A., and coauthors, 2009: Decadal prediction: Can it be skillful? *Bull Am Meterol Soc*, doi: 10.1175/2009BAMS778.1.

Newman, M. 2007: Interannual to decadal predictability of tropical and North Pacific sea surface temperatures. *J Climate*, **20**, 2333-2356.

Pohlmann, H, and coauthors, 2004: Estimating the decadal predictability of a coupled AOGCM. *J. Climate*, **17**, 4463-4472.

Pohlmann, H., J. H. Jungclaus, A. Kohl, D. Stammer, and J. Marotzke, 2009: Initialized decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926-3938.

Saravanan, R. and J. C. McWilliams, 1998: Advective ocean-atmosphere interaction: An analytical stochastic model with implications for decadal variability. *J. Climate*, **11**, 165-188.

Schneider, N., A. J. Miller, and D. W. Pierce, 2002: Anatomy of North Pacific decadal variability. *J. Climate*, **15**, 586-605.

Smith, D., and coauthors, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**,796-799.

Solomon, A., and coauthors, 2009: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: Implication for prediction. *Bull Am Meterol Soc*, submitted.

Sugiura, N., and coauthors, 2009: Potential for decadal predictability in the North Pacific region. *Geophs. Res. Lett*., doi:10.1029/2009GL039787.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the CMIP5 experimental design. http://www-pcmdi.llnl.gov/.

Teng, H., and G. Branstator, 2010: Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Climate Dyn*., in press, doi: 10.1007/s00382-010-0749-7.

Troccoli, A., and T. N. Palmer, 2007: Ensemble decadal predictions from analysed initial conditions. *Phil. T R Soc A*, **365**, 2179-2191.

Figure captions

Fig.1. a) Schematic diagram of time-evolving distributions under changing external forcing. The yellow/red shading represents the climatological distribution ($P_c(t)$). It is independent of any particular initial state. $P_e(t)$ is an ensemble of predicted states evolving from a specific tight cluster of initial conditions. $P_e(t)$ eventually converges to $P_c(t)$ as the influence of the particular initial conditions is lost. The red dashed and black solid lines represent the time-evolving means of the two distributions. b) Scatter plot of annual mean T0-300 at 40^o-45^oW, 40^o-45^oN in the A1B (red) and Commitment (blue) ensembles. The red dashed and blue solid lines denote the two ensembles' time-evolving climatological mean estimated via eqn (1) and (2), respectively. The black open circle at the origin denotes the year 1999 climatological mean estimated by extrapolating the A1B linear trend (eqn(1)). c) Same as b) but for globally averaged T0-300.

Fig.2. The lower panels are variance spectra of T0-300 at four different 5^ox5^o boxes from the control run. The two green dashed lines are first-order autoregressive model (AR1) spectra and 90% significance level based on the 1-yr lag autocorrelations of the time series. The two black vertical reference lines denote frequencies corresponding to 10- and 30-year periods. The spectra are based on 9-point modified Daniell smoothing. Spectra at each gridpoint are used to calculate a variance-weighted mean frequency, which is then converted to a period and plotted in the upper panel. The location of the four selected boxes in the lower panel is marked with a cross, and the ocean basins outlined with the dashed lines are the ocean basins used throughout this study.

Fig.3. Root mean square difference (RMSD) in SST (red) and T0-300 (blue) between all pairs of members from the 40-member A1B (solid lines) and Commitment (dashed lines) ensembles in the North Pacific (120E-110W, 20N-65N) , the North Atlantic (80W-0, 20N-75N), the Tropical Pacific (120E-80W, 20S-20N) and the Tropical Atlantic (80W-0, 20S-20N). Grey dashed lines correspond to 5th percentile of the averaged RMSD between pairs of states from randomly drawn 40 member ensembles from the CCSM3 control run, in SST (thick) and T0-300 (thin), which represent the corresponding 95% significance level.

Fig.4. Root mean square (RMS) amplitude of the initial-value component in the A1B (solid) and Commitment (dashed) ensembles in four ocean basins. Grey dashed lines correspond to the 95th percentile of averaged RMS in SST (thick grey line) and T0-300 (thin grey line), from ensembles of 40 random states in the control run.

Fig.5. Relative entropy of T0-300 anomalies relative to the 1999 climatology (black), relative entropy of the initial-value component (blue), and of the forced component (green). Solid and dashed lines denote the A1B and Commitment ensembles, respectively. Red dashed line indicates the 95th percentile of relative entropy from ensembles of 40 random states in the control run, and red dotted line indicates the relative entropy value with 3 bits more information than random states at the 95% significance level. Each ocean basin is represented by its leading 15 EOFs.

Fig.6. Signal (blue) and dispersion (black) components of relative entropy of the T0-300 initial-value component. The short and long dashed lines are the 95% significance level for the signal and dispersion components, respectively, derived from the control run.

Fig.7. 40-member ensemble mean of T0-300 initial-value component from the Commitment (left) and A1B (right) ensembles in 2004 (year5), 2009 (year10) and 2014 (year15). Stippling indicates the 95th percentile of averages from 40 random states in the control run.

Fig.8. Forced component from the Commitment (left) and A1B (right) ensembles in 2004 (year5) and 2009 (year10). Stippling is as in Fig. 7.

Fig.9. Relative entropy of the T0-300 initial-value component in the A1B ensemble. Orange, green, blue and black lines correspond to upper-100m, 300m, 500m, and 1000m ocean temperature, respectively. Red dashed line is the 95th percentile of relative entropy from 40 random states (which is the virtually the same for all variables because relative

entropy is a standardized quantity). The upper and lower panels are for the North Pacific, and the Tropical Pacific, respectively.

Fig.10. Relative entropy of the T0-300 initial-value component (blue) and forced component (green) in the North Pacific (upper panel) and the Tropical Pacific (lower panel). Solid and dashed lines correspond to relative entropy of the leading 5 (solid line, left y-axis) and 15 (dashed, right y-axis) EOFs, respectively.

Fig.11. Relative entropy of the T0-300 initial-value component in the leading 5 EOFs in the North Pacific. The upper and lower panels are relative entropy values resulting from 10000 random draws of 10- (upper) and 20- (lower) member subensembles from the 40-member A1B ensemble. The black line and shading indicate the average, and the 5th –to-95th percentile range, respectively. Red dashed line is the 95% significance level derived from the control run.

Fig.12. Relative entropy of the T0-300 initial-value component in four ocean basins in four 40-member ensembles that feature different ocean/land/sea ice initial conditions: the A1B ensemble (orange), A1B (II) (black), A1B(III) (green), and A1B(IV) (blue). The red dashed line indicates the 95% significance level derived from the control run. All relative entropy values are calculated from the leading 15 EOFs. X-axis denotes year since branching.

Fig.13. Relative entropy of T0-300 anomalies relative to climatology of year 1999 (black), relative entropy of the initial-value component (blue), and of the forced component (green) in the A1B (solid) and Commitment (dashed) ensembles. The red dashed line indicates the 95% significance level derived from the control run, and the red dotted line indicates the relative entropy value with 10 bits more information than the 95% significance value. All relative entropy values are sums from the 8 ocean basins as outlined in Fig.2, and each basin is represented by its leading 15 EOFs.

A1. Relative entropy of the T0-300 initial-value component in the A1B ensemble in the North Atlantic (upper) and Tropical Pacific (lower). Variations in the three solid lines result from using different time-evolving climatological means as the fitting period of both eqn (1) and (2) is set to 2000-2061 (black), 2010-2061 (blue) and 2020-2061 (green). The relative entropy values are derived from the leading 15 EOFs and the red dashed lines denotes the 95% significance level.

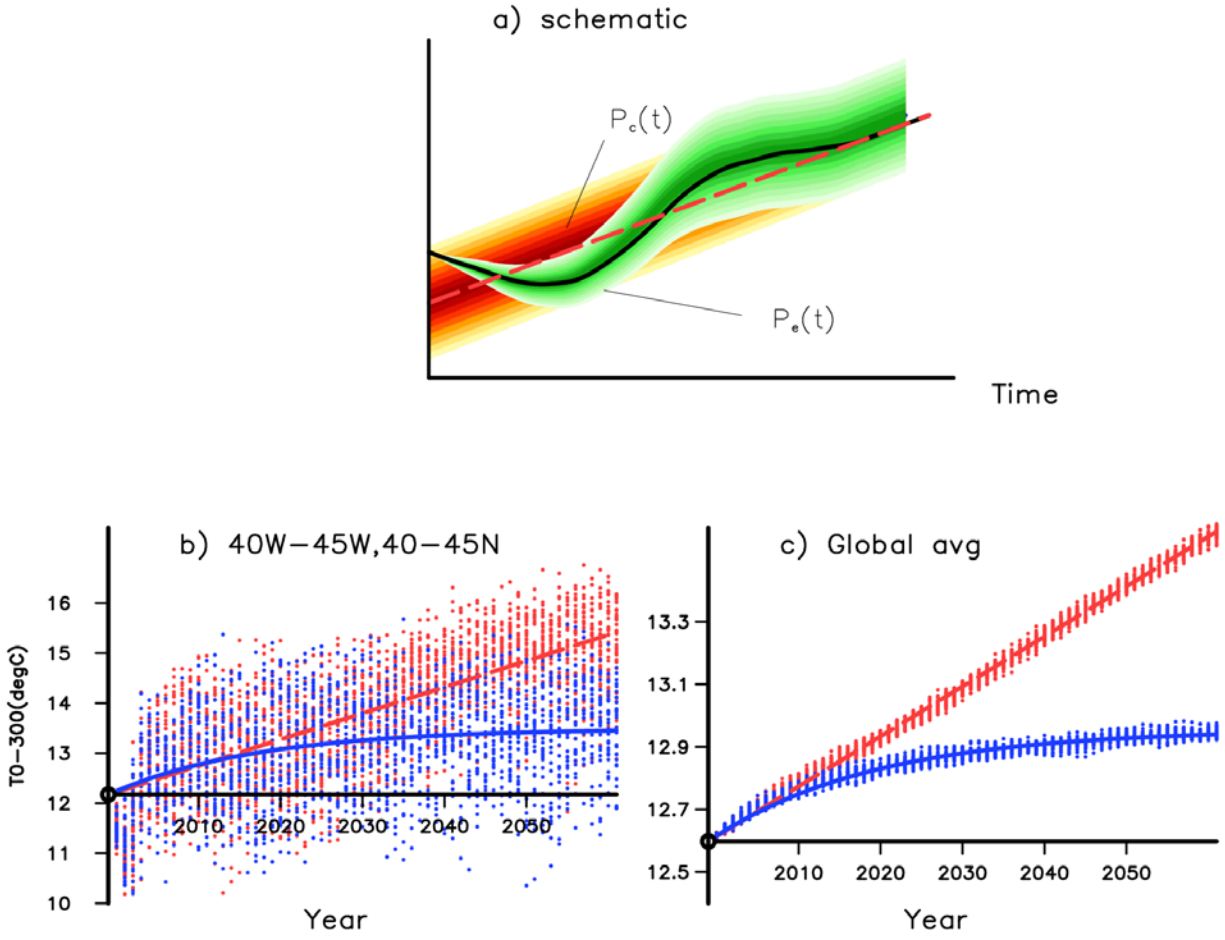


Fig.1. a) Schematic diagram of time-evolving distributions under changing external forcing. The yellow/red shading represents the climatological distribution ($P_c(t)$). It is independent of any particular initial state. $P_e(t)$ is an ensemble of predicted states evolving from a specific tight cluster of initial conditions. $P_e(t)$ eventually converges to $P_c(t)$ as the influence of the particular initial conditions is lost. The red dashed and black solid lines represent the time-evolving means of the two distributions. b) Scatter plot of annual mean T0-300 at 40°-45°W, 40°-45°N in the A1B (red) and Commitment (blue) ensembles. The red dashed and blue solid lines denote the two ensembles' time-evolving climatological mean estimated via eqn (1) and (2), respectively. The black open circle at the origin denotes the year 1999 climatological mean estimated by extrapolating the A1B linear trend (eqn(1)). c) Same as b) but for globally averaged T0-300.

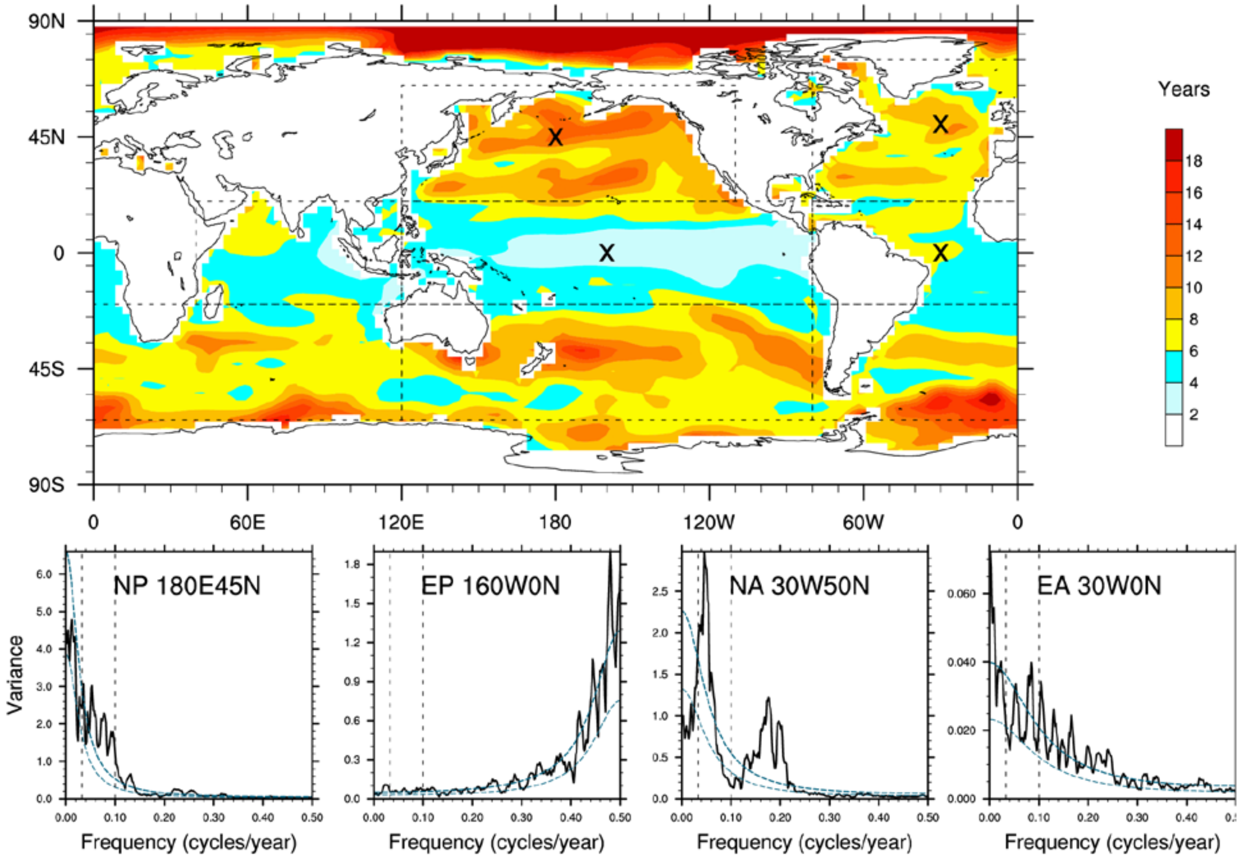


Fig.2. The lower panels are variance spectra of T0-300 at four different 5^{o}x5^o boxes from the control run. The two green dashed lines are first-order autoregressive model (AR1) spectra and 90% significance level based on the 1-yr lag autocorrelations of the time series. The two black vertical reference lines denote frequencies corresponding to 10- and 30-year periods. The spectra are based on 9-point modified Daniell smoothing. Spectra at each gridpoint are used to calculate a variance-weighted mean frequency, which is then converted to a period and plotted in the upper panel. The location of the four selected boxes in the lower panel is marked with a cross, and the ocean basins outlined with the dashed lines are the ocean basins used throughout this study.

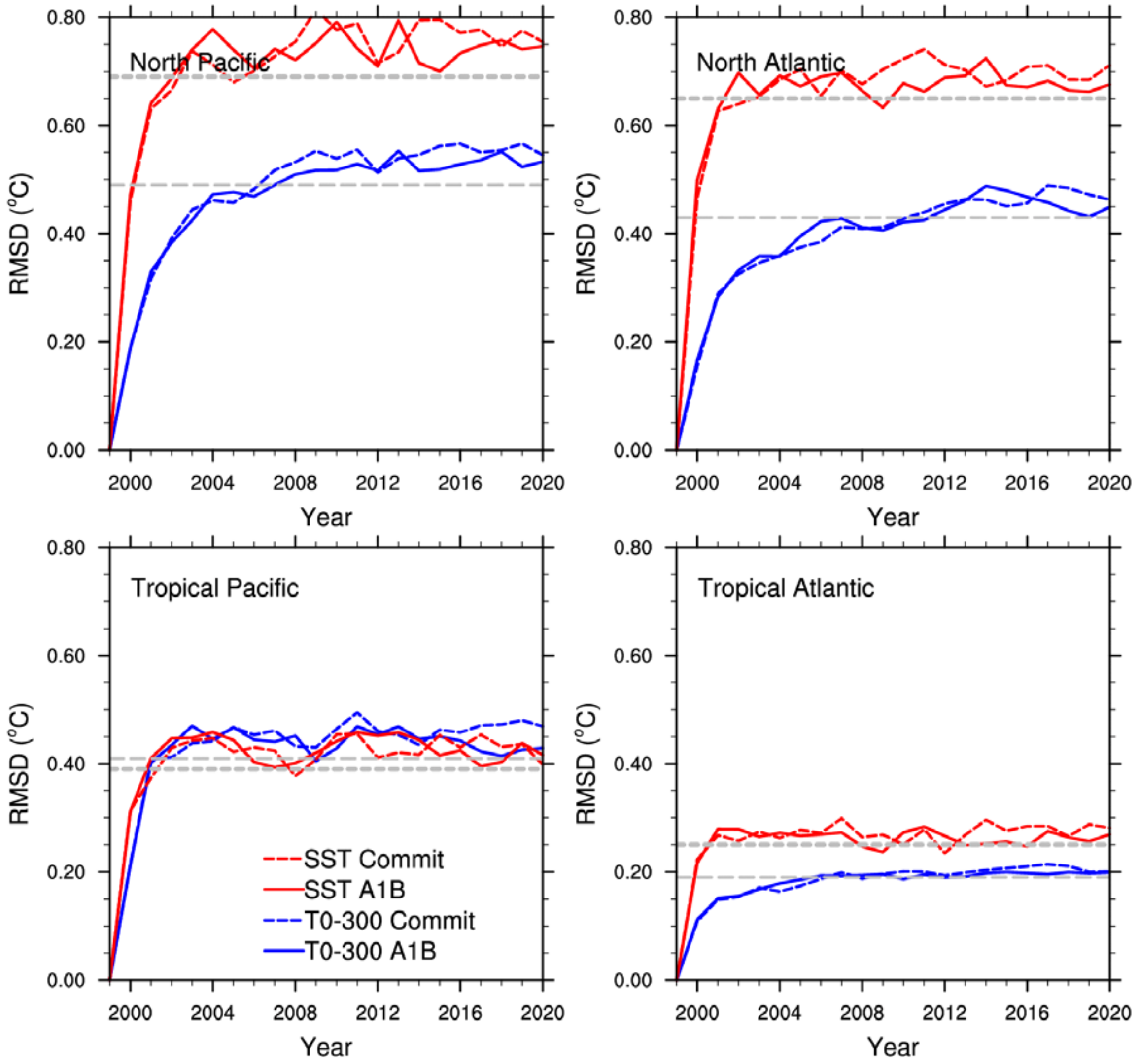


Fig.3. Root mean square difference (RMSD) in SST (red) and T0-300 (blue) between all pairs of members from the 40-member A1B (solid lines) and Commitment (dashed lines) ensembles in the North Pacific (120E-110W, 20N-65N) , the North Atlantic (80W-0, 20N-75N), the Tropical Pacific (120E-80W, 20S-20N) and the Tropical Atlantic (80W-0, 20S-20N). Grey dashed lines correspond to 5th percentile of the averaged RMSD between pairs of states from randomly drawn 40 member ensembles from the CCSM3 control run, in SST (thick) and T0-300 (thin), which represent the corresponding 95% significance level.

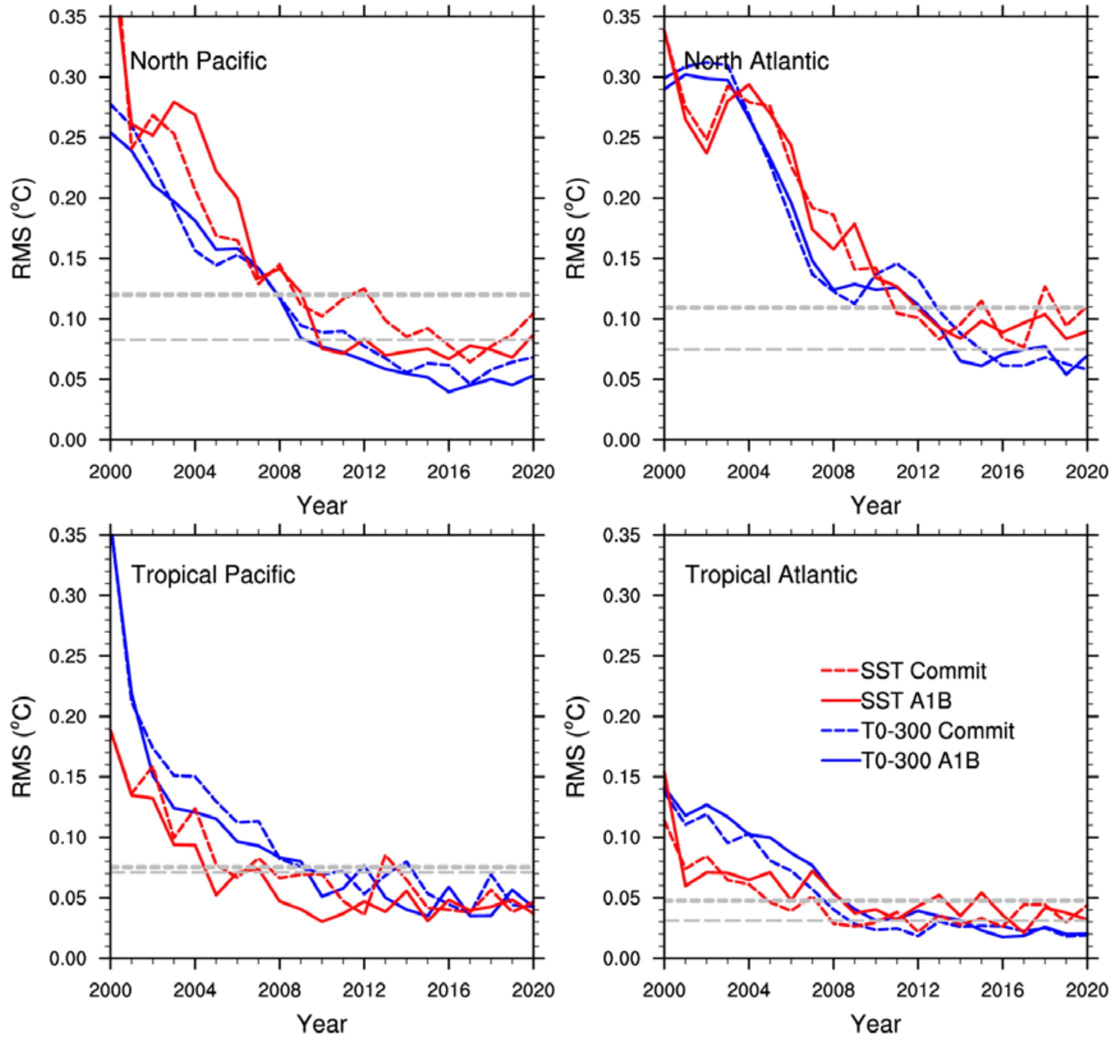


Fig.4. Root mean square (RMS) amplitude of the initial-value component in the A1B (solid) and Commitment (dashed) ensembles in four ocean basins. Grey dashed lines correspond to the 95th percentile of averaged RMS in SST (thick grey line) and T0-300 (thin grey line), from ensembles of 40 random states in the control run.

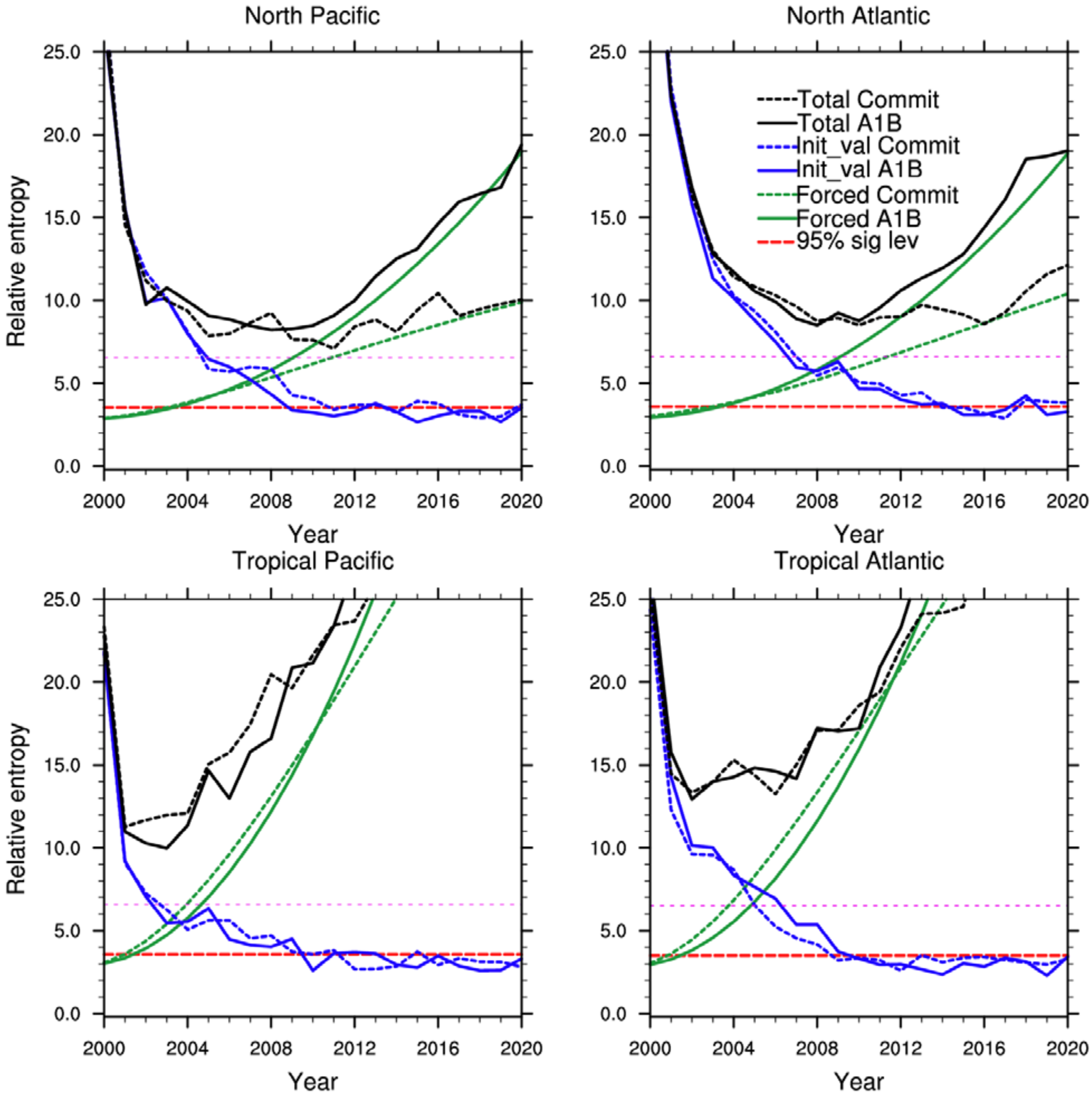


Fig.5. Relative entropy of T0-300 anomalies relative to the 1999 climatology (black), relative entropy of the initial-value component (blue), and of the forced component (green). Solid and dashed lines denote the A1B and Commitment ensembles, respectively. Red dashed line indicates the 95th percentile of relative entropy from ensembles of 40 random states in the control run, and red dotted line indicates the relative entropy value with 3 bits more information than random states at the 95% significance level. Each ocean basin is represented by its leading 15 EOFs.

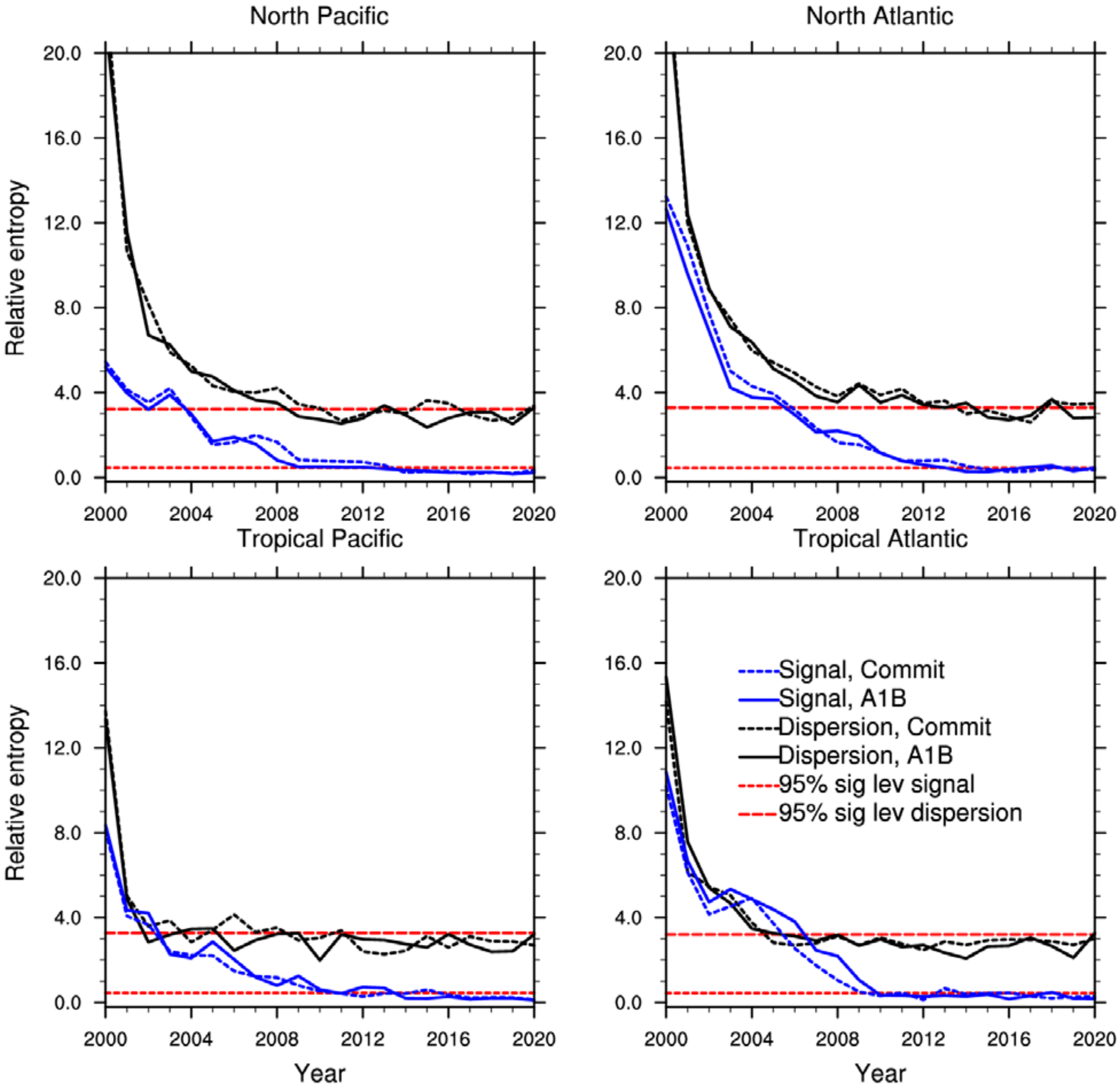


Fig.6. Signal (blue) and dispersion (black) components of relative entropy of the T0-300 initial-value component. The short and long dashed lines are the 95% significance level for the signal and dispersion components, respectively, derived from the control run.

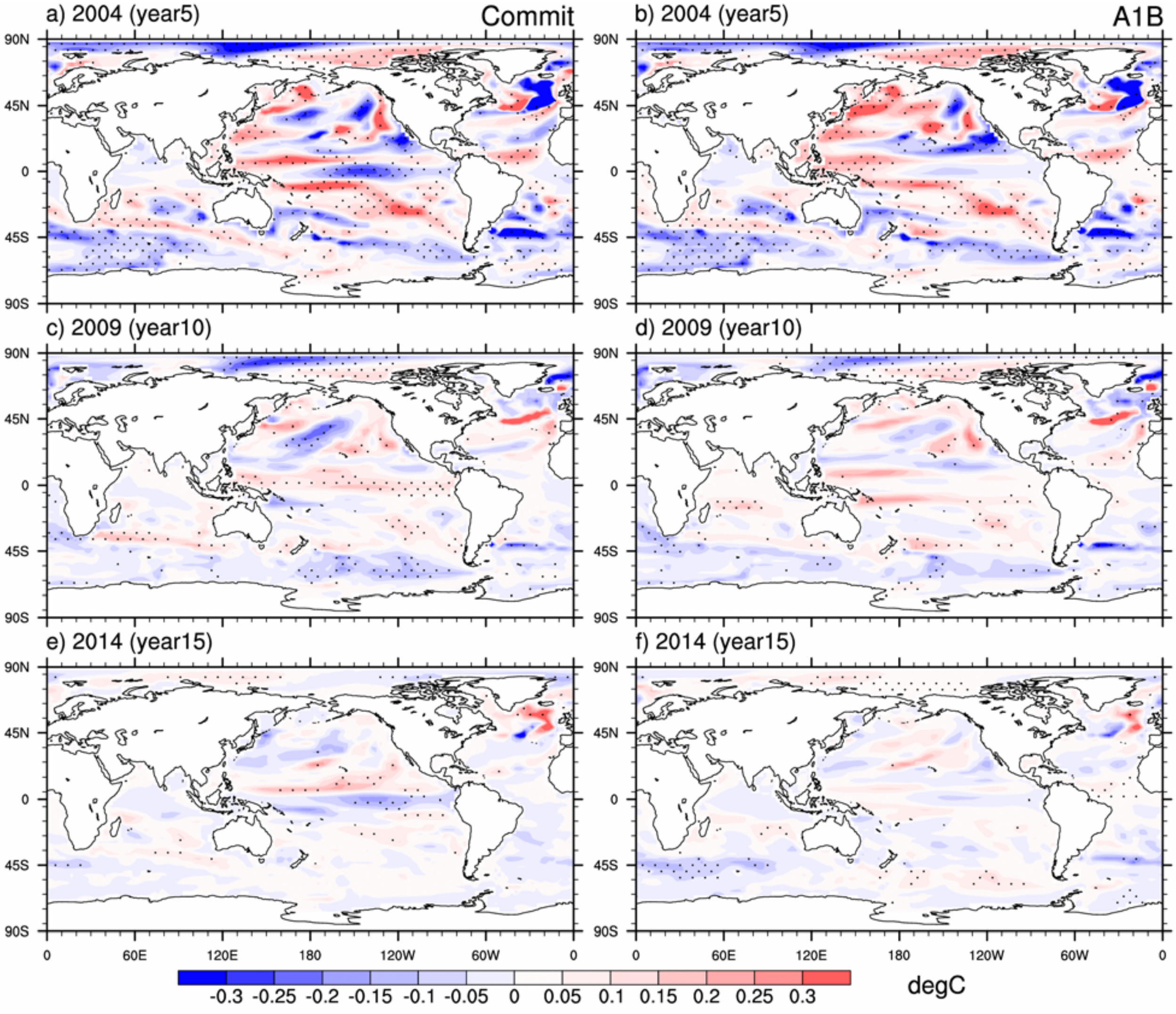


Fig.7. 40-member ensemble mean of T0-300 initial-value component from the Commitment (left) and A1B (right) ensembles in 2004 (year5), 2009 (year10) and 2014 (year15). Stippling indicates the 95th percentile of averages from 40 random states in the control run.

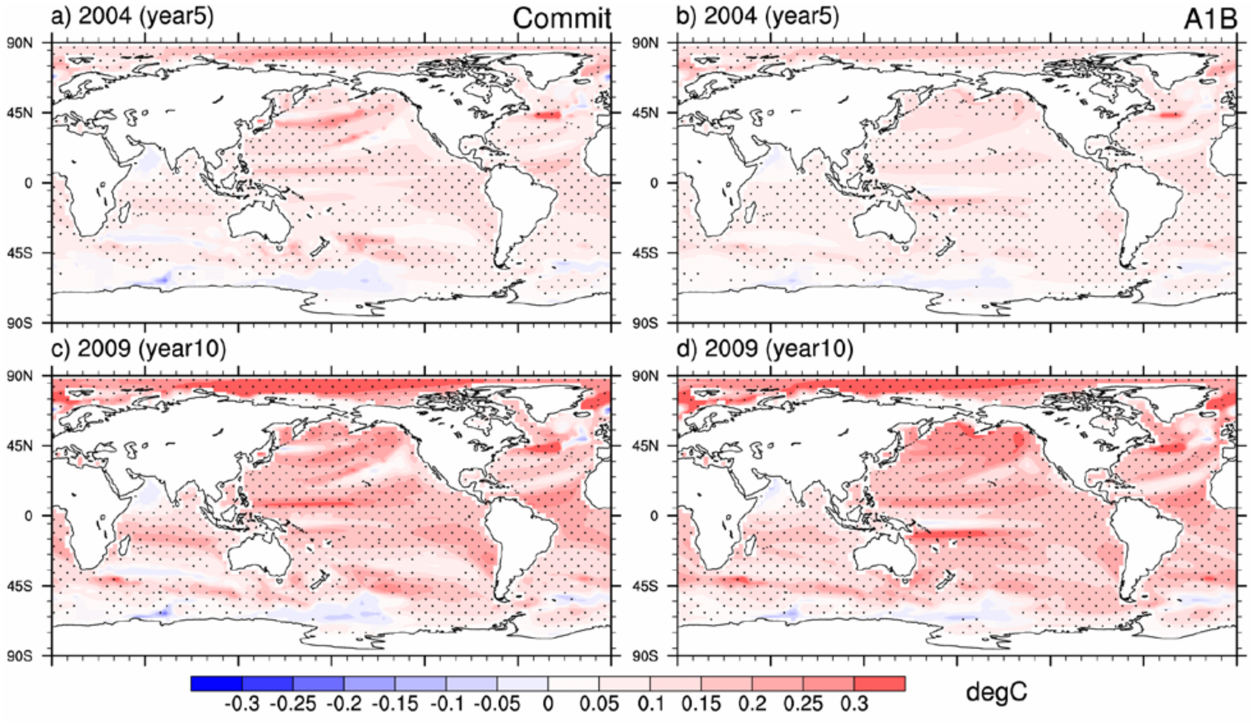


Fig.8. Forced component from the Commitment (left) and A1B (right) ensembles in 2004 (year5) and 2009 (year10). Stippling is as in Fig. 7.

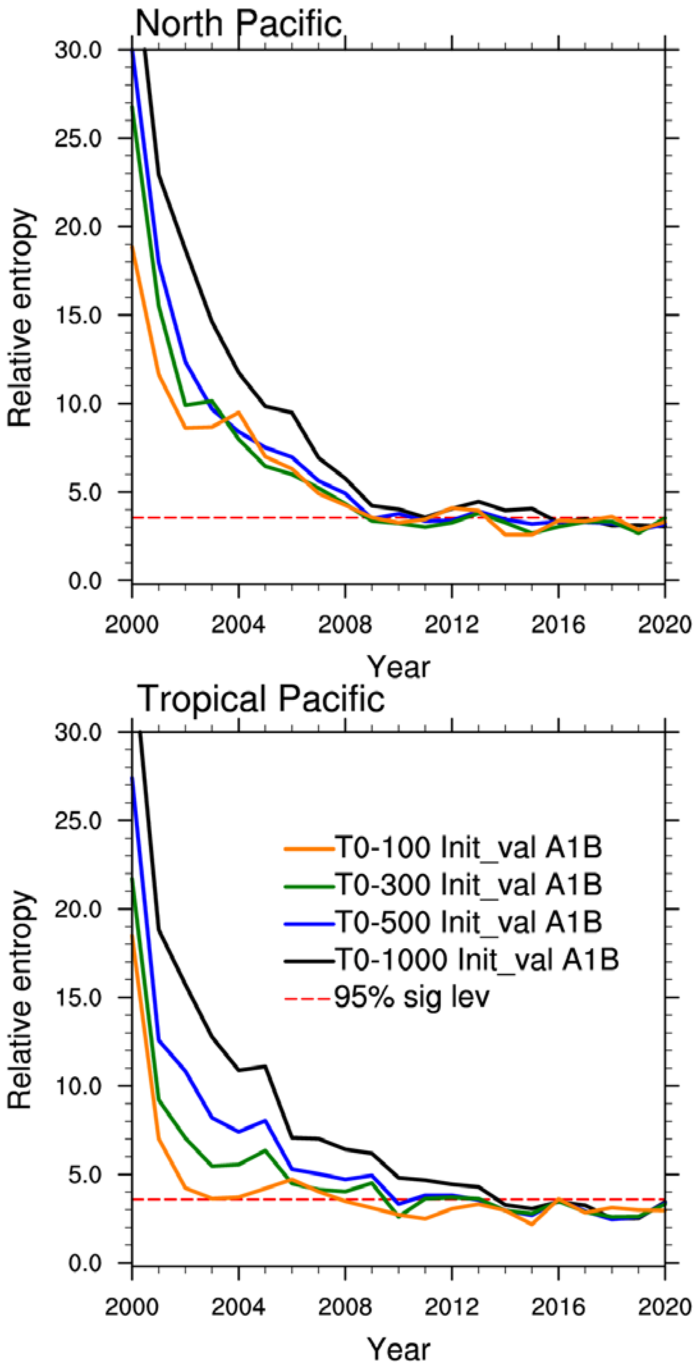


Fig.9. Relative entropy of the T0-300 initial-value component in the A1B ensemble. Orange, green, blue and black lines correspond to upper-100m, 300m, 500m, and 1000m ocean temperature, respectively. Red dashed line is the 95th percentile of relative entropy from 40 random states (which is the same for all four variables because relative entropy is a standardized value). The upper and lower panels are for the North Pacific, and the Tropical Pacific, respectively.

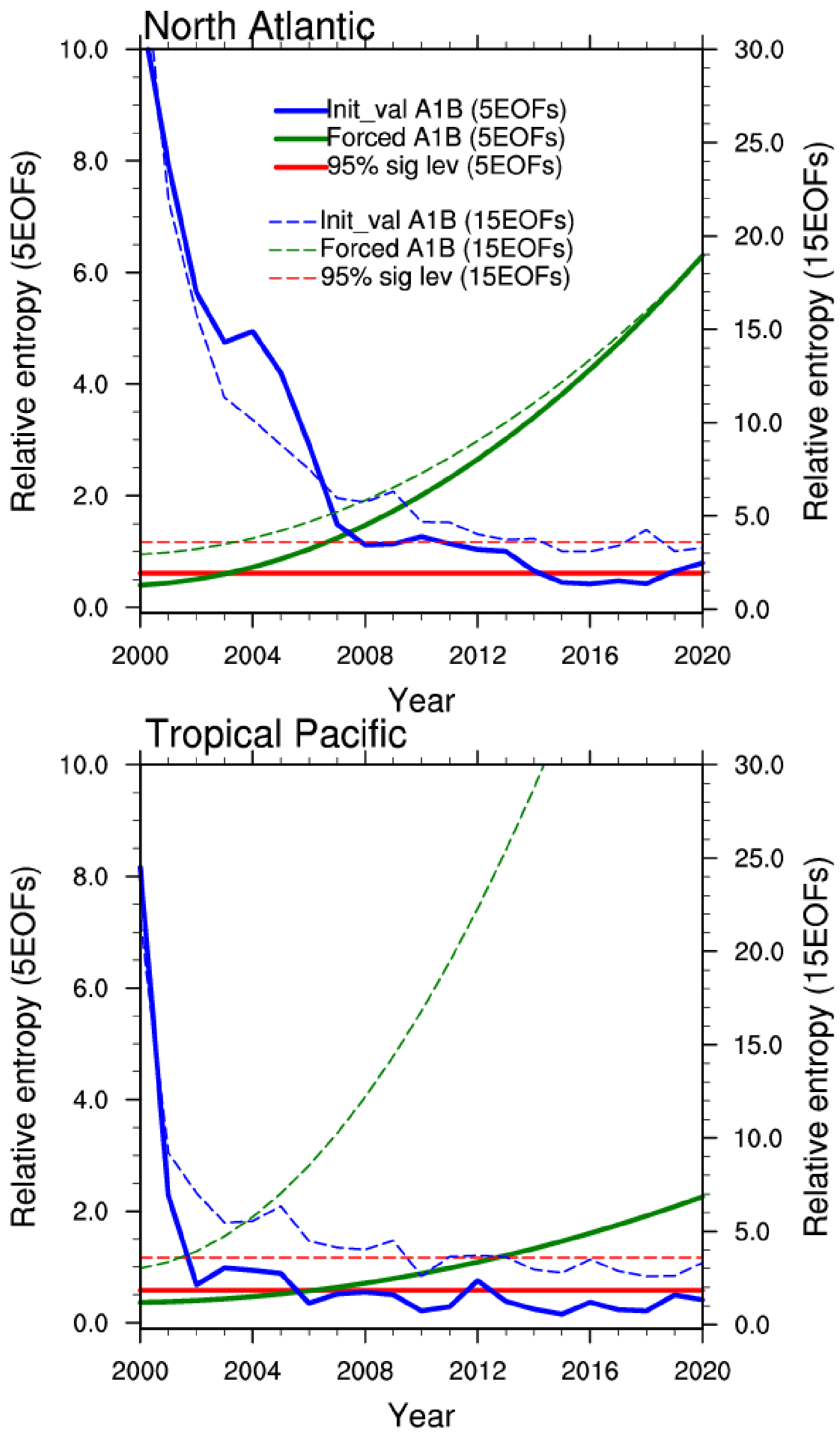


Fig.10. Relative entropy of the T0-300 initial-value component (blue) and forced component (green) in the North Pacific (upper panel) and the Tropical Pacific (lower panel). Solid and dashed lines correspond to relative entropy of the leading 5 (solid line, left y-axis) and 15 (dashed, right y-axis) EOFs, respectively.

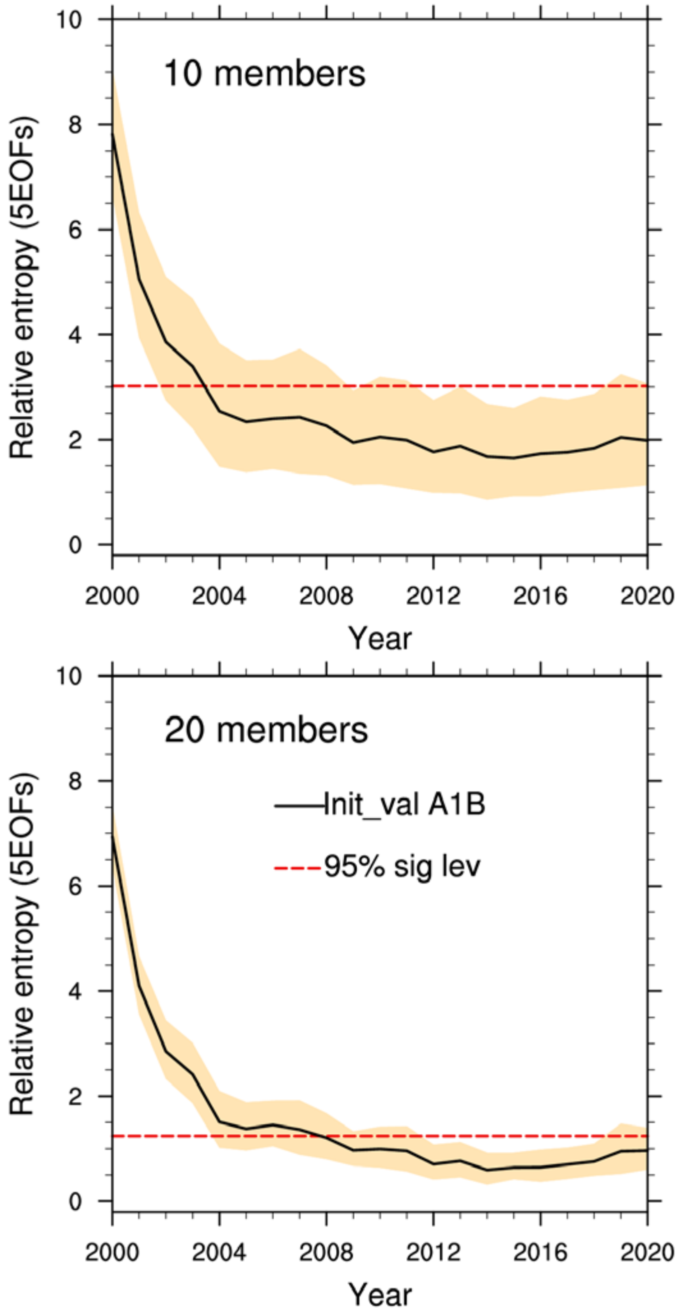


Fig.11. Relative entropy of the T0-300 initial-value component in the leading 5 EOFs in the North Pacific. The upper and lower panels are relative entropy values resulting from 10000 random draws of 10- (upper) and 20- (lower) member subensembles from the 40-member A1B ensemble. The black line and shading indicate the average, and the 5th -to-95th percentile range, respectively. Red dashed line is the 95% significance level derived from the control run.

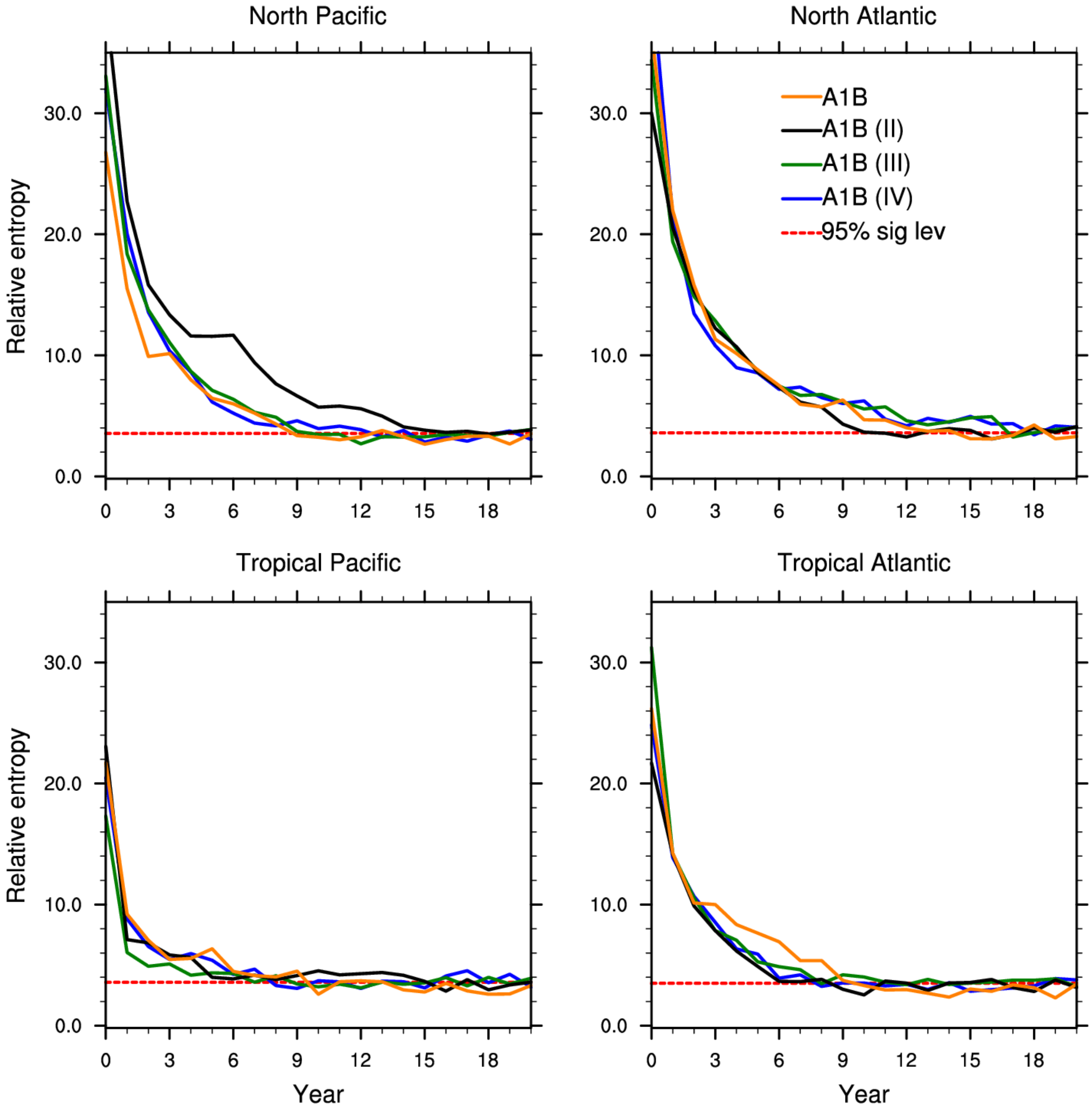


Fig.12. Relative entropy of the T0-300 initial-value component in four ocean basins in four 40-member ensembles that feature different ocean/land/sea ice initial conditions: the A1B ensemble (orange), A1B (II) (black), A1B(III) (green), and A1B(IV) (blue). The red dashed line indicates the 95% significance level derived from the control run. All relative entropy values are calculated from the leading 15 EOFs. X-axis denotes year since branching.

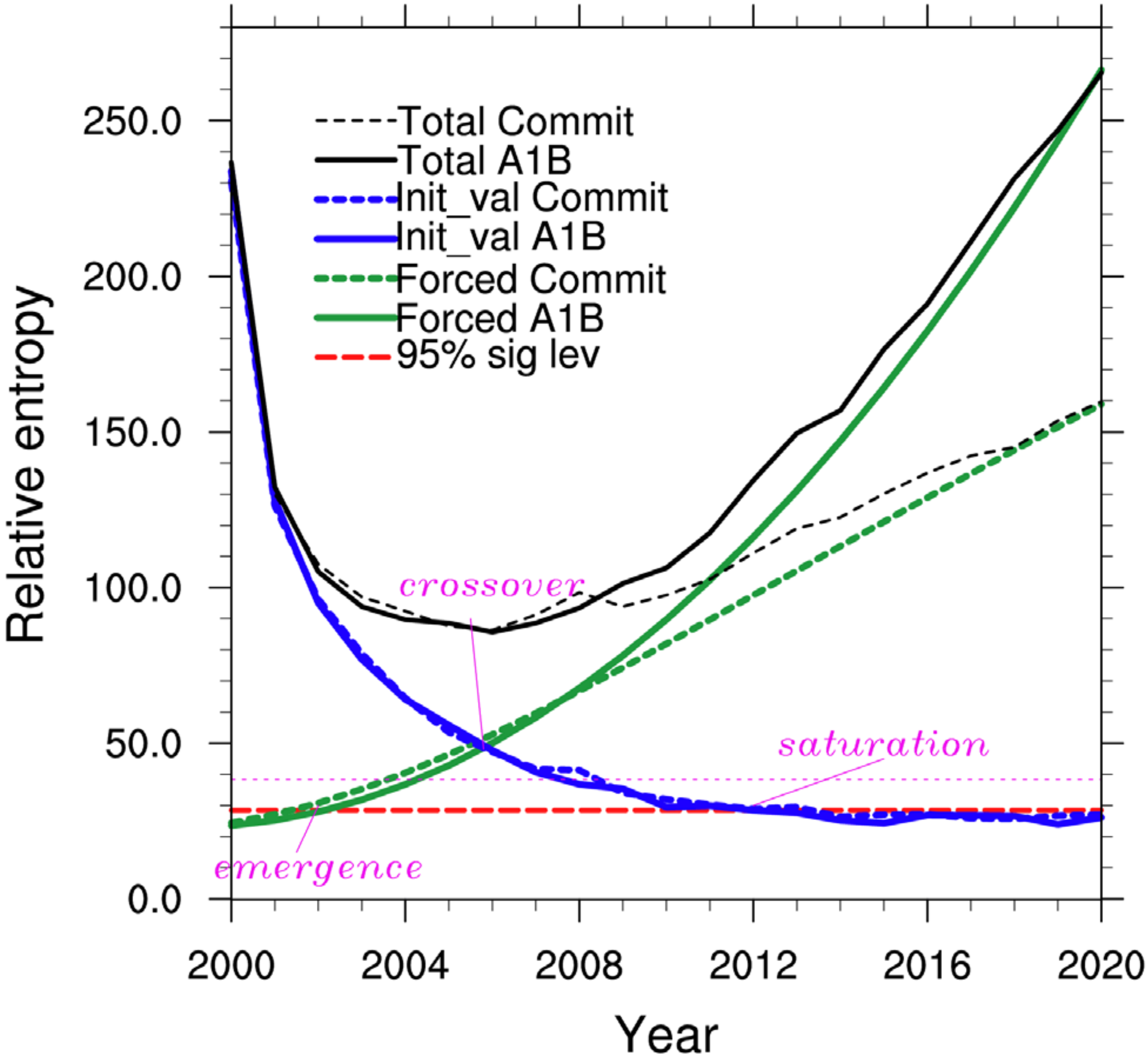
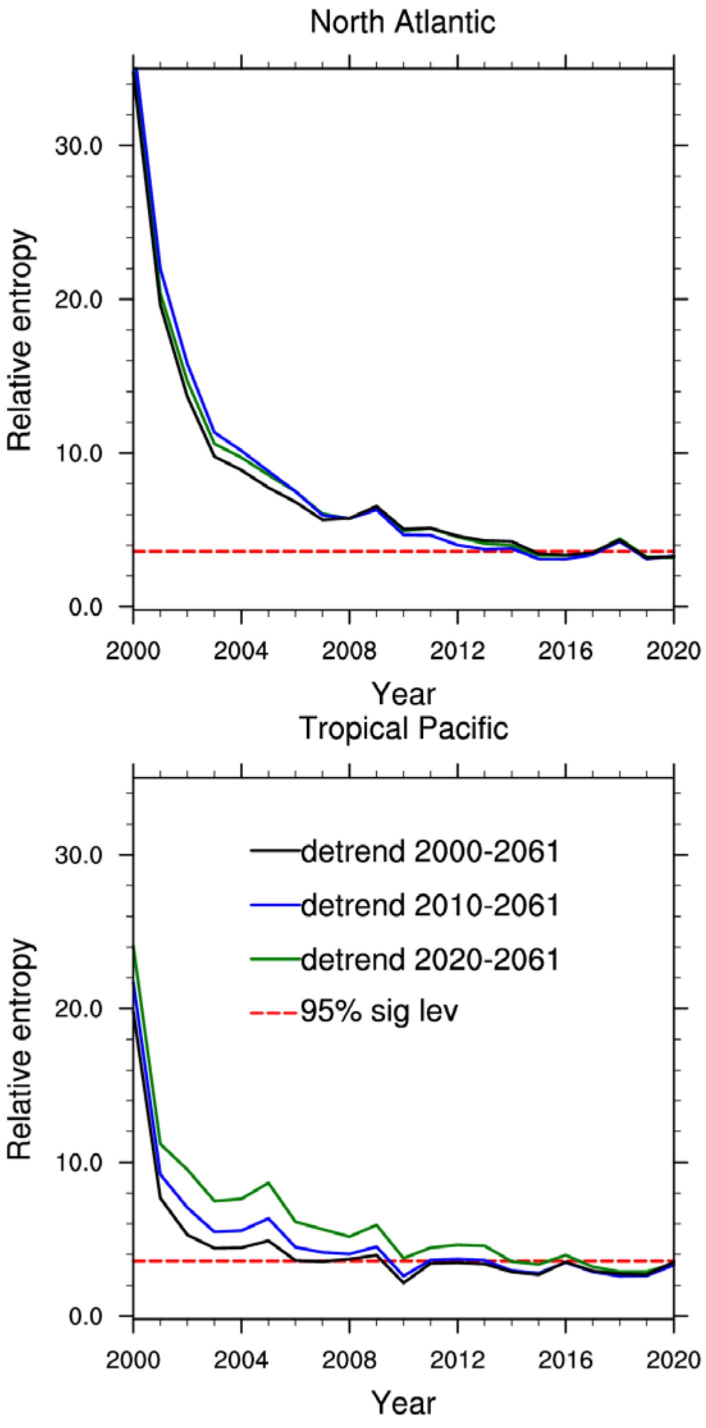


Fig.13. Relative entropy of T0-300 anomalies relative to climatology of year 1999 (black), relative entropy of the initial-value component (blue), and of the forced component (green) in the A1B (solid) and Commitment (dashed) ensembles. The red dashed line indicates the 95% significance level derived from the control run, and the red dotted line indicates the relative entropy value with 10 bits more information than the 95% significance value. All relative entropy values are sums from the 8 ocean basins as outlined in Fig.2, and each basin is represented by its leading 15 EOFs.



A1. Relative entropy of the T0-300 initial-value component in the A1B ensemble in the North Atlantic (upper) and Tropical Pacific (lower). Variations in the three solid lines result from using different time-evolving climatological means as the fitting period of both eqn (1) and (2) is set to 2000-2061 (black), 2010-2061 (blue) and 2020-2061 (green). The relative entropy values are derived from the leading 15 EOFs and the red dashed lines denotes the 95% significance level.