An Evaluation of the CMIP3 and CMIP5 simulations in their skill of simulating the spatial structure of SST variability

By
Gang Wang, Dietmar Dommengen and Claudia Frauen

Draft for Climate Dynamics, February 18, 2013

ABSTRACT

The natural sea surface temperature (SST) variability in the global oceans are evaluated in simulations of the CMIP Phase 3 (CMIP3) and CMIP Phase 5 (CMIP5) models, respectively. In this evaluation, we examine how well the spatial structure of the SST variability match between the observation and simulations on the basis of their Empirical Orthogonal Functions (EOF)-modes. We will compare the models and observations against simple null hypotheses, such as isotropic diffusion (red noise) or a Slab Ocean model, to illustrate the models skill in simulating realistic patterns of variability. Although some models show good skill in simulating the observed spatial structure of the SST variability in some domains, most models show substantial errors. In many cases the simple red noise null hypothesis is closer to the observed structure than most models, despite the fact that the observed SST variability shows significant deviations from this simple red noise null hypothesis. The CMIP models tend to largely overestimate the effective spatial number degrees of freedom and simulate too strong localized patterns of SST variability at the wrong locations with the wrong structure. The most uncertain domains are the North Atlantic and the Southern Ocean on 5yrs running mean time scales. However, the CMIP5 ensemble shows some improvement over the CMIP3 ensemble, mostly in the tropical domains.
1. Introduction

The Coupled Model Intercomparison Project (CMIP) presents a highly valued resource to the climate science research for the understanding of natural variability and future climate change [Meehl et al., 2007; Taylor et al., 2012]. However, the models of CMIP, which are from various institutions and different in their structures and physical parameterizations, have shown significant disagreement and uncertainties on their performance. The aim of the study presented here is to evaluate the CMIP models skill in simulating the natural internal spatial structure of SST variability in all major ocean basins (tropical Indo-Pacific, North Pacific, tropical and North Atlantic and the Southern Ocean). This should guide the climate community for the understanding of natural modes of SST variability and should help in the development of seasonal to decadal forecasting systems.

Previous model evaluations have mostly focused on the mean state climate [Taylor, 2001, Boer and Lambert 2001, Murphy et al. 2004, Gleckler et al., 2008], the general strength of climate variability [Boer and Lambert 2001, Gleckler et al., 2008] or on some regional aspects of climate variability [Guilyardi, 2006, Zhou et al. 2009, Jamison and Kravtsov, 2010, Xavier et al., 2010], but have so far not focused on SST variability on the global scale.

In the study presented here will base our model evaluation on the comparison of the empirical orthogonal function (EOF) modes of SST variability for the different major ocean basins in all the model simulations and observations. The method of comparing EOF-modes is based on the study of Dommenget [2007] and Bayr and Dommenget [2013]. This method allows quantifying the agreement in the multi-variate spatial structure of SST variability in a systematic way. An important aspect in such an evaluation is to put the results of this relatively abstract and complicated analysis into the perspective of some simple null hypotheses, which should help to guide the researchers in evaluating the significance of the results. The simple null hypotheses used in this study describe the spatial structure of the SST as they would result from simplified physical processes such as isotropic diffusion (red noise) or atmospheric forcings only.

The paper is organized as follows: firstly the section 2 presents the data and methodology used. Section 3 introduces the null hypotheses chosen for the evaluation and section 4 shows results of the EOF-mode comparison, which is the main of this study. Finally a summary and discussion are provided in section 5.

2. Data and Method

2.1 Data
The main observed global monthly mean SSTs are from the Hadley Centre Sea Ice and SST data set (HadISST, referred as “observation” below; Rayner et al., 2003), and we also have chosen the NOAA Extended Reconstructed sea surface temperature data set [ERSST, Smith et al., 2008] as an auxiliary. Model simulations are taken from the CMIP3 and CMIP5 databases (Meehl et al., 2007; Taylor et al., 2012). Our analysis focuses on the 20th century SST simulations corresponding to the scenarios of “20c3m” in CMIP3 and “historical run” in CMIP5, respectively. Table 1 and 2 list all available simulations for this study.

An output of a Slab Ocean coupled experiment is also used to compare versus the CMIP models in this study. [Washington and Meehl, 1984; Dommenger and Latif, 2002; Murphy et al., 2004; Dommenger, 2010]. We take 500 years output from this simulation and divide the data to five 100-year chunks for the analysis. The model uses the atmospheric component of the ACCESS model in N48 resolution and 38 vertical levels coupled to a simple slab ocean model for open ocean conditions and sea ice climatology is prescribed.

All data sets (models and observations) were analyzed for the period 1900 to 1999, interpolated to a common 2.5° latitude × longitude grid and linearly detrended to remove the global warming signal prior to the analysis. Monthly mean SST anomalies were defined for each model and the observations individually. All the CMIP3 SST anomalies are concatenated to generate a super CMIP3 model with 2300 years of data. Similarly, a CMIP5 super model with 2300 years of data and a CMIP3+5 with 4600 years of data are also constructed.

2.2 Comparison of EOF modes

We base our comparison of the spatial structure of SST variability in different data sets on the comparison of the EOF-modes, assuming that the leading EOF-modes give a good representation of the large-scale SST variability. This can be done by defining the EOF-modes of one data set as the reference modes and project the EOF-modes of the other data set onto these modes to estimate the amount of variance that the reference EOF-modes explain in this data set. This concept is based on Dommenger [2007] and Bayr and Dommenger [2013].

An EOF eigenvector (mode) of the reference data set \( A \), \( \vec{E}_i^A \) and its corresponding eigenvalue \( \vec{\sigma}_i^A \) are compared with the eigenvector \( \vec{E}_j^B \) and eigenvalue \( \vec{\sigma}_j^B \) of another data set \( B \) by projecting the eigenvectors \( \vec{E}_i^A \) onto the \( \vec{E}_j^B \).
\[ c_{ij} = \frac{e_i^A e_j^B}{\| e_i^A \| \| e_j^B \|} \]  

(1)

where \( c_{ij} \) is the uncentered pattern correlation coefficient between the two EOF-patterns. The projected explained variance of mode \( E_i^A \) in data set \( B \), \( pe_i^{AB} \), is estimated by the accumulation of all eigenvalues of \( B \) [Dommenget, 2007]:

\[ pe_i^{AB} = \sum_{j=1}^{N} c_{ij}^2 e_j^B \]  

(2)

The value \( pe_i^{AB} \) represents the total variance of data set \( B \) that can be explained by the reference mode \( E_i^A \), with \( N \) the number of EOF-modes considered.

We can illustrate this method of comparing two data sets with an example: In Figure 1 we show the leading EOF modes of the North Pacific for the observations and the GFDL-cm2.1 model simulation. We can note here that the two data sets have slightly different leading modes of variability and that the explained variance of each of these modes is also slightly different. To compare the overall spatial structure of variability in the two data sets we chose the observed EOF-modes as the reference modes \( \bar{E}_i^A \) and project the EOF-modes of the GFDL-cm2.1 model simulation \( \bar{E}_i^B \) onto these modes. Fig. 2a shows the eigenvalues of the observed EOF-modes \( \bar{e}_i^A \) against the project explained variance from the GFDL-cm2.1 model simulation \( pe_i^{AB} \). The leading observed mode (Fig. 1a), for instance, explains only half as much variance in the GFDL-cm2.1 model simulation (red line in Fig. 2a) than it does in the observations (black line in Fig. 2a). So comparing the overall spatial structure of variability essentially means to estimate how much variance each of the reference modes explain in both data sets.

We can further see in this comparison that the explained variances of the observed leading modes are significantly less in the GFDL-cm2.1 model simulation. This overall mismatch in the explained variances can be quantified by a normalized root-mean square error (RMSE$_{EOF}$) between the \( \bar{e}_i^A \) and \( pe_i^{AB} \) values:

\[ RMSE_{EOF}(A,B) = \sqrt{\frac{\sum_{i=1}^{N_A} (e_i^{AB} - e_i^A)^2}{N_A}} \frac{\sqrt{\sum_{i=1}^{N_A} (e_i^A)^2}}{N_A} \]  

(3)

The normalization allows a better comparison of the RMSE$_{EOF}$ values between different domains with different sampling uncertainties. Here \( N_A \) corresponds to the number of EOF modes considered. In all following analysis we chose \( N_A \) equal to the effective number degrees of freedom, \( N_{spatial} \), also known as the number of independent modes after Bretherton et al. [1999]:
The sum $RMSE_{EOF}(A, B)$ is dominated by the mismatch in the leading modes, as they have larger uncertainties. Uncertainties in the higher order modes are less important in this estimate. Fig. 2 shows the $RMSE_{EOF}$ values for two examples. A $RMSE_{EOF}$ value of 1 corresponds to errors that are as big as the eigenvalues. Thus the $RMSE_{EOF}$ value (0.41, see Fig. 2a) of the GFDL-cm2.1 model relative to the observations reflects an uncertainty of the leading EOF-modes of about 40% of the eigenvalues, which is a substantial uncertainty.

3 Formulation of null hypotheses

The comparison of the spatial patterns of SST variability in different data sets in this study is based on projecting EOF-modes and estimating the $RMSE_{EOF}$ values. These $RMSE_{EOF}$ values are quite abstract values and it is important to put these values into perspective with some simple null hypotheses to understand the significance of these values. We therefore formulate a number of theoretical reference null hypotheses: first we estimate the $RMSE_{EOF}$ for sampling uncertainties after North et al. [1982]. We then formulate three simple physical models for the spatial pattern of SST variability: the first is the slab ocean modes of variability. The second and third physical models are based on the assumption that the spatial patterns of SST variability are just a reflection of isotropic diffusion.

3.1 Uncertainty of EV

North et al. [1982] gives the statistical uncertainties of the eigenvalues $e_i$ due to sampling errors:

$$
\delta e_i = e_i \left(2/N_{sample}\right)^{1/2} \tag{5}
$$

where $N_{sample}$ is the number of independent samples. In this study $N_{sample}$ is estimated as $N/t_d$, while $N$ is the length of the time series and $t_d$ is the average e-folding decorrelation time based on first five leading Principal Components (PCs). To maintain consistency, we estimate $N$ from the shorter time series of the CMIP model rather than the much long references. For the $RMSE_{EOF}$ for sampling uncertainties is then

$$
RMSE_{EOF}(\delta e_i) = \sqrt{\frac{\sum_{i=1}^{N_A}(\delta e_i)^2}{N_A}} / \sqrt{\frac{\sum_{i=1}^{N_A}(e_i^2)}{N_A}} \tag{6}
$$
Thus, if two data sets are just different stochastic realizations of the same process (have the same spatial patterns of SST variability), then we expect the RMSE$_{EOF}$ to be in average $RMSE_{EOF}(\delta e_i)$.  

**Fig. 3** illustrates an example based on subsampling the CMIP5 super model. Here we computed the EOF-modes for the North Pacific for monthly mean SST (Fig. 3a) and for 5yrs running mean SST (Fig. 3b). For subsampling the CMIP5 super ensemble we split the data from each model into average chunks with 5yrs/60mon data and concatenated the chunks into the CMIP5 super subsamples. Here the total numbers of samples $N$ is set to be 1380 (60*23) for monthly mean and 115 (5*23) for the 5yrs running mean. For monthly mean analysis, $t_d = 8.8$, which represents the average decorrelation time of leading five PCs, then we get $N_{sample} = 1380/8.8 = 156.8$. Similarly $N_{sample} = 115/5.8 = 19.8$ for the 5yrs running mean.  

We can note that the subsamples $p e_i^{AB}$ values fluctuate around the $\delta e_i^A$ values of the CMIP5 super model, as it is expected from the $\delta e_i$ values. Subsequently, the RMSE$_{EOF}$ values of the subsamples fluctuate around the $RMSE_{EOF}(\delta e_i)$. For the 5yrs running mean SST it seems that the subsamples fluctuate less than expected by the $RMSE_{EOF}(\delta e_i)$, which could indicate that our subsampling of the models is not quite representative of the sampling uncertainties in the CMIP5 super model.  

**3.2 Slab Ocean Model**  
The spatial structure of the SST variability is a result of the coupled dynamics between atmosphere and oceans. A slab ocean model coupled to an AGCM (see data section for model details) can estimate the spatial structure of the SST variability that is caused by the atmosphere only. It is therefore a good null hypothesis to evaluate the role of ocean dynamics: if the spatial structure of the SST variability agrees better with the slab ocean model than with a CGCM, than it is indicating that the ocean dynamics are causing unrealistic SST patterns.  

**Fig. 4** shows the observed EOF-modes of the monthly mean SST variability in the North Pacific in comparison with the EOF-modes of the ACCESS-slab simulation, for an example. We can note here that the EOF-modes of the slab simulations are already quite realistic, suggesting that much of the large-scale structure of the SST variability is to first order simulated in the slab simulation, which is consistent with what has been found in other studies as well [e.g. Pierce et. al, 2001].  

**3.3 Isotropic diffusion**  
Cahalan et al., [1996] and Dommenget [2007] used the null hypothesis of isotropic
diffusion to explain the leading EOF-modes of climate variability. The isotropic diffusion process leads to EOF-modes that are a hierarchy of multi-poles, starting with a monopole (largest scale) and followed by a dipole, followed by multi-poles with increasing complexity (smaller scales). It essentially represents a spatial red noise process [Dommenget 2007]. Fig 4 (last row) illustrates the EOF-mode of isotropic diffusion for the North Pacific domain. The EOF-modes are pure geometric deconstructions of the domain not considering any structure in the SST standard deviation (STDV, Fig. 5), but assuming the same effective number degrees of freedom [after Bretherton et al. 1999] as observed. Thus the spectrum of the explained variance of the eigenvalues has a structure similar to what is observed. We refer to this null hypothesis as $H_0^{\text{uniform}}$.

We can slightly alter the isotropic diffusion process by also assuming the observed inhomogeneous SST STDV field and therefore focusing the leading EOF-modes onto regions where the observed SST STDV is large [see Dommenget 2007 for details]. This concept has also been applied to study the structure of the Indian Ocean SST variability, for instance, by Dommenget [2011]. Fig 4 (third row) illustrates the EOF-mode of isotropic diffusion with observed SST STDV for the North Pacific domain. The EOF-modes are still a hierarchy of multi-poles, but now the nodes are centered on regions of large SST STDV. These EOF-modes now have some more realistic features. We refer to this null hypothesis as $H_0^{\text{STDV}}$.

4. Comparison of the EV-spectrum

In this section we present the main results of this analysis, which is based on the comparison of the projected $pe_i^{AB}$ values (referred as EV-spectrum below) of the SST variability in different ocean basins. We start this analysis with a more comprehensive discussion of the North Pacific to illustrate the method. We then compare all model simulations with the observed EOF-modes for all ocean basins. The analysis is then repeated by pairwise comparisons of the CMIP model simulations to evaluate the uncertainties within the model ensemble members.

4.1 North Pacific

Fig. 6 shows the EV-spectrum of the observed monthly mean SSTA in the North Pacific region together with the projected $pe_i^{AB}$ values for all CMIP5 models and the four different null hypotheses references. A few points can be noted here:

- ERSST is close to the observations (HADISST) as they are basically the same observed data. The differences are mostly within the sampling uncertainties.
However, there is some indication that the two data sets are not totally in agreement.

- All the models and null hypotheses underestimate the first PC of observation, namely the PDO pattern (see Fig. 1a) and most of the other leading modes.
- The deviations of the individual models from the observed EV-spectrum are much larger than expected by the sampling uncertainties $\delta e_i$. The slab ocean simulation is more similar to the observation than most models.
- The deviations of $H_{0_{STDV}}$ are about as strong as for most of the CMIP5 model simulations. However, the deviations of $H_{0_{uniform}}$ appear to be larger than those of most CMIP5 models.

The results of the EV-spectrum are quantified by the RMSE$_{EOF}$ values in Fig. 7 for the monthly mean SST as shown in Fig. 7 on the x-axis and for the 5yrs running mean SST on the y-axis. The results of the CMIP3 models are shown as well. In addition to what we have already concluded above for the monthly mean SST modes we can note the following points:

- The ERSST is close to expected sampling uncertainties for both time scales. Although, this indicates that the two observational data sets have good agreements in this domain, it has to be noted here that the two data sets containing the same samples (same observations). Thus an even better agreement should have been possible.
- The model errors relative to the observations are in the range of 0.3 to 0.8. Thus they in average have errors in the leading modes of variability in the order of 30-80% of the eigenvalues. These are substantial errors.
- The RMSE$_{EOF}$ values in Fig. 7 are almost linearly distributed, indicating that in this region the models show similar performance for monthly mean and 5yrs running mean variability. Most models seem better than the $H_{0_{uniform}}$; however, half of the models are not as good as the $H_{0_{STDV}}$ hypothesis.
- The slab ocean simulation is closer to the observation than most models in the monthly mean modes, but is about average for the 5-years running mean modes.
- The mean position of the CMIP3 models is close to that of the CMIP5 models, implying similar skill in this region, but most of the outliers with very large errors are in the CMIP3 ensemble.

We can pick out a few models to illustrate how the modes of variability in some models deviate from those observed. In Fig. 8 we show the leading modes of the two worst (GISS-EH and BCM2.0), the two best models (MRI-CGCM2.3.2 and CCSM4) and the
CMIP3 and CMIP5 super models in the North Pacific modes comparison. The following can be noted here:

- The two worst models both show leading EOF-modes that are quite different in structure from the observed. In particular, these modes show quite localized peaks that are unrelated to any observed structures.
- The two best models show leading EOF-modes that are similar in structure to the observed and that have similar amount of explained variance.
- The modes of the super models are very similar to the observed and have very smooth structures with no strong localized features. They also tend to explain less variance than observed. The reduced variance of the leading modes relative to the observed, and to what individual models show, reflects the fact that the super models are based on ensembles of individual models that have different localized structures (modes), which leads to less explained variance of the leading modes.

4.2 Uncertainties in the SST modes in the global oceans

The analyses are now extended to all other ocean domains (displayed in Fig. 9) in Fig. 10. To sum up, we also average the results, and get the global summary of the EV-RMSE diagram in Fig. 10a. The results show a number of interesting aspects. We start the discussion with a focus on the individual domains, starting with the Indo-Pacific domain:

- On the monthly mean time scales the spread in the quality of the models is very large. A few models are close to the observed modes and most models are quite different from the observed modes. On the longer 5yrs time scales the models spread is smaller and many models seem to be closer to the observed modes.
- The slab model is quite different from the observed modes on the monthly mean time scales and the longer 5yrs time scales. The El Nino dynamics dominate the modes of the Indo-Pacific domain and these dynamics are not simulated in the slab model, which may explain why the slab model is quite different from the observed modes.
- The simple Null hypothesis $H_0^{STDV}$ performs better than most models, but the $H_0^{uniform}$ hypothesis is clearly worth than most models and very different from the observed structure.
- The CMIP3 ensemble has much more outliers, in particular on the shorter time scales, than the CMIP5 ensemble.

In the North Atlantic the picture is quite different:
• The most remarkable feature is that both simple null hypotheses are closer to the observer modes than most of the CGCM simulations. First of all this is due to the fact that the observed modes in the North Atlantic are indeed more similar to the simple null hypotheses than they are in the Indo-Pacific domain. But still it indicates that the CGCM simulations have substantial problems in simulating these simple modes of variability.

• The agreement with the observed modes is better on the shorter time scales than on the longer 5yrs time scales. This is the opposite of what we see in the Indo-Pacific domain.

• The slab model performs better than any of the CGCM simulation on the monthly mean time scales and still better than many models on the longer time scales.

• There is no substantial different in the performance of the CMIP3 and CMIP5 ensembles.

The tropical Atlantic has again some interesting features:

• Notable is that the CGCM simulations are in average closer to the observed modes on the longer and also on the shorter time scales than in any other domain. On the longer time scales this domain actually seem to be the only domain where the CGCM simulations are mostly in agreement with the observed modes.

• The CMIP5 simulations show a clear improvement over the CMIP3 simulations for the longer 5yrs time scales. More than in any other domain. This is even more remarkable considering the already good fit to the observations in the CMIP3 simulations and also considering the much better performance than in any of the other domains.

• The slab model is better than most CGCM simulations on the monthly mean time scales, but worth than all models on the longer time scales.

• Both the H0STDV and the H0uniform hypotheses are very close to the observations on both time scales and closer than most of the CGCM simulations. This indicates that knowing the SST STDV field and assuming a multi-pole deconstruction, as it follows from isotropic diffusion process, would already explain most of the SST variability in this domain.

Finally, the Southern Ocean:

• This domain shows the overall largest deviations from the observed modes, with all models disagreeing with the observations substantially.
• There appears to be no substantial difference between the CMIP3, CMIP5 and the slab simulations.
• Similar to the North Pacific both null hypotheses are substantially different from the observed modes on both time scales. However, the $H_{0_{STDV}}$ hypothesis is closer to the observed modes than any CGCM simulation on the longer time scales and close than most on the shorter time scales.

The summary of all individual domains (Fig.10a) shows the average skill of the models:

• First of all we can note that the models skill on the short and longer time scales are roughly linearly related. Model that are close to the observed modes on the shorter time scales tend to be close to the observed modes on the longer time scales as well.
• Basically all model show significant deviations in the spatial structure from the observed modes. These deviations are in the order of 50% (0.5) of the eigenvalues on the leading modes. This means they in average under/overestimate some of the leading modes by a factor of $1.5/0.5$, which is a substantial error.
• In the global average some models are clearly much closer to the observed modes (e.g. the CCCSM4 model is closest) and some models substantially deviate from the observed modes (e.g. all the CMIP3 GISS models). However, the spread in the global average is not as big as in the individual domains, indicating that model that have big RMSE$_{EOF}$ in some domains have often smaller RMSE$_{EOF}$ in other domains.
• The CMIP5 ensemble appears to be slightly closer to the observed modes than the CMIP3 ensemble on both time scales. However, the supermodel modes are very similar in their structure and skill relative to the observed modes.
• The slab simulation is of similar skill on the shorter time scales, but has less skill than most models on the longer time scales. However, on both time scales the slab simulation is not consistent (RMSE$_{EOF}$ is larger than expected by sampling uncertainties) with observations.
• The simple $H_{0_{STDV}}$ hypothesis is in average closer to the observed modes than any CGCM simulation on the longer time scales and close than most on the shorter time scales. Even the $H_{0_{uniform}}$ hypothesis is better than many models. This suggest that knowing the effective number of spatial degrees of freedom, the domain geometry and, most importantly, knowing the SST STDV field, and assuming a modal structure resulting from isotropic diffusion already describes the observed spatial structure of SST variability better than most of the CGCM simulations.
The above analysis has shown that the CMIP model simulations have substantial errors in simulating the observed spatial structure of SST variability. We can have a closer look at the leading EOF-modes of the model simulations to understand what it is that the models do differently if compared to the observations:

- First we can note that $N_{\text{spatial}}$ is larger than observed in most CMIP models and for all domains and on both time scales, (see Fig. 11). It is also larger than in the slab simulation. This suggests that the simulated leading modes of variability explain in average less variance than observed and are on smaller spatial scales (the patterns are more localized) than observed. This can also be seen by visual inspection of the leading modes of all the model simulations (see supplemental Figs. 1-10).
- Further we can notice that the patterns of the leading modes in the model simulations are different from those observed (see supplemental Figs. 1-10). They are often quite localized patterns of scales much smaller than the domain size. The observations also do have such localized modes, but these are often at different locations than in the models and are of different structure and amplitude. Thus the models produce a double error: They simulate significant localized structures at the wrong locations and with the wrong structure and subsequently miss the observed localized structures at the right locations and with the right structure. The simple null hypotheses and the slab simulation do not have these localized structures and therefore do not have these double errors.

In the analysis it is also noticeable that the super model ensembles do not perform significantly better than most of the models. This is quite different from many other inter model comparisons (e.g. seasonal forecasting skills or mean state errors), where the ensemble mean outperforms the individual models. The modes of variability or the spatial structure of internal SST variability does not average out to be more realistic in an ensemble super model. If models have different modes of variability, than the super model will have all of these modes, but each with a smaller eigenvalue, which increases $N_{\text{spatial}}$ of the ensemble super model, (see Fig. 11). However, we can illustrate that the ensemble super model is indeed improving to most individual models if we replace the eigenvalues in eq. [2] against observed eigenvalues. These scaled values  are shown in Figs. 7 and 10. These scaled values are much closer to the observed spatial structure than most models. It illustrates that the leading patterns in the ensemble super model are indeed realistic, but that just the relative explained variance of each mode are
underestimated by the ensemble super model due to the artificial diversity in the individual models.

Finally, we can also discuss the similarity in the two different observations:

- For most domains there is a relative good agreement between the ERSST and the HadISST datasets on both time scales. This indicates that we have some relative good confidence in the spatial structure of SST variability in these domains.
- The best agreement is in the North Atlantic, which seems to be consistent with the larger database existing in this relatively well-observed domain.
- Strong disagreement exists in the Southern Oceans. Here the spatial structure of the observed SST variability is very uncertain. On the longer time scales the uncertainties in the leading modes are in the order of 40-60% of the eigenvalues, which is a substantial uncertainty. Again, this seems to be consistent with the lag of sufficient observations in this domain.

### 4.3 Comparison between models

In the above section we evaluated the models against the observation, which illustrated some substantial differences of the model’s spatial structures in SST variability relative to the observed. We also noticed that the leading modes of the CMIP3+5 super model models have much smaller explained variance as the observed modes, illustrating a larger diversity of modes in the model ensembles relative to the observed. This indicates that the model have strong differences in the spatial structures in SST variability between each other. We can quantify these model-to-model differences by repeating the above analysis by pairwise comparison of the EOF-modes in the CMIP3 and 5 ensembles.

Fig. 11 shows the RMSE_{EOF} values as in Fig. 10, but for the average of all pairwise comparisons between all CMIP3 and 5 models. Thus the reference modes in these comparisons are the EOF-modes from each of the CMIP3 and 5 models. Here small RMSE_{EOF} values suggest small differences in the spatial structures in SST variability of the model relative the spatial structures in SST variability of all the other models and vise versa for large RMSE_{EOF} values. The following features can be noted in this comparison:

- All models show RMSE_{EOF} values larger than expected from sampling uncertainties (North-RMSE). Thus the models substantially disagree on each other in terms of the spatial structures in SST variability. The errors are in the order of 40-60% of the eigenvalues.
- The largest model internal spread is in the North Atlantic and the Southern Ocean. This is similar to what we found in the comparison with the observations.
• In global average the model-to-model spread is similar in all models, indicating that there is no model that is closest to all the other models.
• The CMIP5 models tend to be slightly closer to all the other models than the CMIP3 models on both time scales and for all domains.
• In the tropical Indo-Pacific domain several CMIP3 models have substantially larger RMSE_{EOF} values than most of the other models. This suggest that these models are indeed quite different from the over all model ensemble. All these models

5. Summary and Discussion
In the study presented here we evaluated the skill of the CMIP3 and CMIP5 model simulations to simulate the observed spatial structure of SST variability on interannual and decadal time scales. This comparison was based on a quantitative comparison of the leading EOF-modes in the five major ocean basins (tropical Indo-Pacific, North Pacific, tropical and North Atlantic and the Southern Ocean) with the observed EOF-modes and those of simplified null hypotheses. The study illustrated a number of interesting aspects in the skill of the model simulations, but also about the observed spatial structure of SST variability. For the observed spatial structure of SST variability we can list the following main findings:

• By comparing the observed modes with those of the simple isotropic diffusion null hypothesis we can note that for most domains and time scales the observed spatial structure of SST variability is significantly different from isotropic diffusion. Thus the observed modes of variability have non-trivial structure. In particular on the monthly mean time scales in the tropical Indo-Pacific and Atlantic and on both time scales in the North Pacific and Southern ocean. The longer time scales of the tropical and North Atlantic are, however, remarkably similar to the simple large-scale multi-pole modes of the isotropic diffusion process.
• The effective number of spatial degrees of freedom (N_{spatial}) are between 5-10 for most domains on the monthly meantime scales and smaller (<5) on the longer 5yrs time scales. The Indo-Pacific, which is the largest domain, has the smallest N_{spatial}, whereas the Southern ocean, which is similar in size to the tropical Indo-Pacific domain, has the largest N_{spatial} on both time scales, marking the most complex spatial structure in SST variability.
• The comparison of the two observational datasets suggests that the modes of SST variability are relatively well known for most domains, but not for the Southern ocean. Here the uncertainty in the SST modes is quite substantial, even in two
datasets that contain the same observations.

We can start the summary and discussion of the model results with some positive findings:

- Some models have a quite realistic spatial structure of the SST variability in some domains at some time scales. In particular in the monthly mean time scales in the tropical Pacific and also (for some models) in North Pacific. On the longer 5yrs time scales most models simulate the tropical Atlantic and Indo-Pacific SST variability with quite realistic spatial structure. The good performance of these models in these domains is in particular notable, as these models also out perform the simple null hypotheses, indicating they can indeed simulate non-trivial spatial structure of SST variability.

- The CMIP5 ensemble does show some improvement over the CMIP3 ensemble. The most significant improvements are seen in the two tropical domains. In the tropical Atlantic the CMIP5 ensemble as a whole is shifted towards more realistic variability on the longer time scales and in the tropical Indo-Pacific the CMIP5 ensemble has improved on both time scales, but mostly by a lag of very bad models and not by an improvement of the best models.

The most important findings of this study are, however, the substantial limitations that the CMIP3 and CMIP5 model ensembles have in simulating the spatial structure of SST variability:

- Most CMIP models in most domains on both the monthly mean and the 5yr running mean time scales have less skill in simulating the spatial structure of SST variability than the simple isotropic diffusion (red noise) null hypothesis $H_{0_{\text{STDV}}}$. And in many cases they have less skill then the Slab Ocean simulation. On the shorter time scales the models are only better in the North Pacific and the tropical Indo-Pacific. The tropical Atlantic region is the only region in which the CMIP ensembles perform equally good as the $H_{0_{\text{STDV}}}$ null hypothesis on the longer time scales.

- The models largely over estimate the effective number of spatial degrees of freedom ($N_{\text{spatial}}$) in all domains and particular on the shorter time scales. Thus the models produce more complex spatial structure in the SST variability, with more localized smaller scale patterns.

- The models do not only disagree largely with the observations, but also with each other. The mismatch between the models is as big as the mismatch with the
observations. The largest uncertainties are in the North Atlantic and in the Southern Ocean on the longer time scales. Here the mismatch between models is larger than relative to the simple H0STDV null hypothesis.

Several aspects of this analysis indicate that the models limited skill is caused by ocean dynamics: The first piece of evidence comes from the relative good performance of the Slab Ocean simulation, which does not simulate any ocean dynamics, but performs more realistic in the simulation of the shorter time scales in the extra tropical domains than most CGCM simulations. The most remarkable difference to the CGCM simulations here is the much more realistic low $N_{\text{spatial}}$ values in all extra-tropical domains. This may indicate that the ocean GCM simulations of extra-tropical dynamics causes significant problems. In particular it seem to generate much more complex small-scale SST variability that is inconsistent with observations. This is also related to the second piece of evidence pointing towards problems in the ocean GCM simulations: The models produce to many small, localized modes of variability that are at the wrong positions with the wrong structures. Such modes do not exist in the Slab Ocean simulation nor in the simple null hypotheses. This result seems to be consistent with what we know from the dynamics of the atmosphere and oceans: Atmospheric meso-scale internal variability is on a much larger scale, than that of the oceans. Indeed current state of the art CGCMs do not resolve oceanic meso-scale dynamics. The coarse resolution of the ocean models (not resolving the meso-scale dynamics) may potentially be one of the main problems in the CGCMs.

However, it also need to be noted that the CGCMs need to simulate a correct mean SST climatology in order to simulate the correct spatial structure of the SST variability. In particular in the extra-tropical domains, SST variability is often a reflection of variability relative to fronts in either the ocean (e.g. between different gyres) or the atmosphere (e.g. jet stream). The variability in the position or the strength of the fronts is a significant part of the SST variability. CGCM simulations that simulate the positions of these fronts incorrectly will not be able to simulate the spatial structure of SST variability correctly. Here the Slab Ocean simulation has a significant advantage, as is has the right mean SST climatology by construction, due to the use of flux correction terms.

Acknowledgements
References


Table Captions

Table 1 List of models from the CMIP3 analyzed in this study

Table 2 List of models from the CMIP5 analyzed in this study
Figure Captions

Fig.1. First three EOF patterns of detrended monthly SSTA under Observation and GFDL-CM2.1.

Fig.2. (a) EV-spectrum projected to the observation. The bars mark the uncertainty interval of the eigenvalues after North et al. (1982) (b) Similar as (a) but projected to CMIP3+5 super model.

Fig. 3 (a) and (b). EV-spectrum of CMIP5 super model and the projected ev-spectrum of its subsamples. The shaded area marks the uncertainty interval of the eigenvalues after North et al. (1982). (c) RMSE_EOF distrubition.

Fig.4. Leading PC structures of monthly SSTA of (a). Observation (HadISST); (b). slab ocean experiment; (c). fitted isotropic diffusion process with inhomogeneous standard deviation forcing (d); as in (c) but with homogeneous forcing.

Fig.5. Standard deviation fields of monthly SSTA of (a). HadISST; (b). CMIP3 ensemble mean; (c). CMIP5 ensemble mean; (d). slab ocean experiment result.

Fig. 6. EV-spectrum of the observed EOF-modes and all pev values of all the models and null hypotheses relative to the observed EOF-modes. The shaded area marks the uncertainty interval of the observed eigenvalues after North et al. (1982).

Fig. 7 EV-RMSE errors spreading monthly mean vs. 5yrs. Blue numbers are cmip3 models listed in table.1 and red numbers are cmip5 models in table.2. “◇” represents average position of the CMIP models. “*” is the result of the super model.

Fig. 8 Leading EOF-modes of monthly SSTA of GISS-EH, BCM2.0, MRI-CGCM2.3.2, CCSM4, CMIP3 super model and CMIP5 super model.

Fig. 9 Distribution of the domains.

Fig. 10. As in Fig. 6 but in (a). Global summary; (b). Tropical Indian Ocean and Pacific; (c). North Atlantic; (d). Tropical Atlantic; (e). Southern Ocean.

Fig. 11 As in Fig. 6 but projected to the CMIP3+5 super model.

Fig. 12 As in Fig. 6 but projected to the (a) CMIP3 super model; (b) CMIP5 super model.
Table. 1 List of models from the CMIP3 analyzed in this study

<table>
<thead>
<tr>
<th>Number</th>
<th>Originating Group(s)</th>
<th>Country</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bjerknes Centre for Climate Research</td>
<td>Norway</td>
<td>BCCR-BCM2.0</td>
</tr>
<tr>
<td>2</td>
<td>Canadian Centre for Climate Modelling &amp; Analysis</td>
<td>Canada</td>
<td>CGCM3.1 (T47)</td>
</tr>
<tr>
<td>3</td>
<td>Canadian Centre for Climate Modelling &amp; Analysis</td>
<td>Canada</td>
<td>CGCM3.1 (T63)</td>
</tr>
<tr>
<td>4</td>
<td>Météo-France / Centre National de Recherches Météorologiques</td>
<td>France</td>
<td>CNRM-CM3</td>
</tr>
<tr>
<td>5</td>
<td>CSIRO Atmospheric Research</td>
<td>Australia</td>
<td>CSIRO-Mk3.0</td>
</tr>
<tr>
<td>6</td>
<td>CSIRO Atmospheric Research</td>
<td>Australia</td>
<td>CSIRO-Mk3.5</td>
</tr>
<tr>
<td>7</td>
<td>US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory</td>
<td>USA</td>
<td>GFDL-CM2.0</td>
</tr>
<tr>
<td>8</td>
<td>US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory</td>
<td>USA</td>
<td>GFDL-CM2.1</td>
</tr>
<tr>
<td>9</td>
<td>NASA / Goddard Institute for Space Studies</td>
<td>USA</td>
<td>GISS-AOM</td>
</tr>
<tr>
<td>10</td>
<td>NASA / Goddard Institute for Space Studies</td>
<td>USA</td>
<td>GISS-EH</td>
</tr>
<tr>
<td>11</td>
<td>NASA / Goddard Institute for Space Studies</td>
<td>USA</td>
<td>GISS-ER</td>
</tr>
<tr>
<td>12</td>
<td>LASG / Institute of Atmospheric Physics</td>
<td>China</td>
<td>FGOALS-g1.0</td>
</tr>
<tr>
<td>13</td>
<td>Instituto Nazionale di Geofisica e Vulcanologia</td>
<td>Italy</td>
<td>INGV-SXG</td>
</tr>
<tr>
<td>14</td>
<td>Institute for Numerical Mathematics</td>
<td>Russia</td>
<td>INM-CM3.0</td>
</tr>
<tr>
<td>15</td>
<td>Institut Pierre Simon Laplace</td>
<td>France</td>
<td>IPSL-CM4</td>
</tr>
<tr>
<td>16</td>
<td>Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)</td>
<td>Japan</td>
<td>MIROC3.2 (hires)</td>
</tr>
<tr>
<td>17</td>
<td>Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)</td>
<td>Japan</td>
<td>MIROC3.2 (medres)</td>
</tr>
<tr>
<td>18</td>
<td>Max Planck Institute for Meteorology</td>
<td>Germany</td>
<td>ECHAM5/MPI-OM</td>
</tr>
<tr>
<td>19</td>
<td>Meteorological Research Institute</td>
<td>Japan</td>
<td>MRI-CGCM2.3.2</td>
</tr>
<tr>
<td>20</td>
<td>National Center for Atmospheric Research</td>
<td>USA</td>
<td>CCSM3</td>
</tr>
<tr>
<td>21</td>
<td>National Center for Atmospheric Research</td>
<td>USA</td>
<td>PCM</td>
</tr>
<tr>
<td>22</td>
<td>Hadley Centre for Climate Prediction and Research / Met Office</td>
<td>UK</td>
<td>UKMO-HadCM3</td>
</tr>
<tr>
<td>23</td>
<td>Hadley Centre for Climate Prediction and Research / Met Office</td>
<td>UK</td>
<td>UKMO-HadGEM1</td>
</tr>
<tr>
<td>Number</td>
<td>Originating Group(s)</td>
<td>Country</td>
<td>Model</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------</td>
<td>---------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>1</td>
<td>CSIRO and BOM</td>
<td>Australia</td>
<td>ACCESS1.0</td>
</tr>
<tr>
<td>2</td>
<td>Beijing Climate Center, China Meteorological Administration</td>
<td>China</td>
<td>BCC-CSM1.1</td>
</tr>
<tr>
<td>3</td>
<td>National Center for Atmospheric Research</td>
<td>USA</td>
<td>CCSM4</td>
</tr>
<tr>
<td>4</td>
<td>Météo-France / Centre National de Recherches Météorologiques</td>
<td>France</td>
<td>CNRM-CM5</td>
</tr>
<tr>
<td>5</td>
<td>Canadian Centre for Climate Modelling and Analysis</td>
<td>Canada</td>
<td>CanESM2</td>
</tr>
<tr>
<td>6</td>
<td>Geophysical Fluid Dynamics Laboratory</td>
<td>USA</td>
<td>GFDL-CM3</td>
</tr>
<tr>
<td>7</td>
<td>Geophysical Fluid Dynamics Laboratory</td>
<td>USA</td>
<td>GFDL-ESM2G</td>
</tr>
<tr>
<td>8</td>
<td>Geophysical Fluid Dynamics Laboratory</td>
<td>USA</td>
<td>GFDL-ESM2M</td>
</tr>
<tr>
<td>9</td>
<td>NASA / Goddard Institute for Space Studies</td>
<td>USA</td>
<td>GISS-E2-H</td>
</tr>
<tr>
<td>10</td>
<td>NASA / Goddard Institute for Space Studies</td>
<td>USA</td>
<td>GISS-E2-R</td>
</tr>
<tr>
<td>11</td>
<td>Hadley Centre for Climate Prediction and Research / Met Office</td>
<td>UK</td>
<td>HadCM3</td>
</tr>
<tr>
<td>12</td>
<td>Hadley Centre for Climate Prediction and Research / Met Office</td>
<td>UK</td>
<td>HadGEM2-CC</td>
</tr>
<tr>
<td>13</td>
<td>Hadley Centre for Climate Prediction and Research / Met Office</td>
<td>UK</td>
<td>HadGEM2-ES</td>
</tr>
<tr>
<td>14</td>
<td>Institute for Numerical Mathematics</td>
<td>Russia</td>
<td>INM-CM4</td>
</tr>
<tr>
<td>15</td>
<td>Institut Pierre Simon Laplace</td>
<td>France</td>
<td>IPSL-CM5A-LR</td>
</tr>
<tr>
<td>16</td>
<td>Institut Pierre Simon Laplace</td>
<td>France</td>
<td>IPSL-CM5A-MR</td>
</tr>
<tr>
<td>17</td>
<td>Institut Pierre Simon Laplace</td>
<td>France</td>
<td>IPSL-CM5B-LR</td>
</tr>
<tr>
<td>18</td>
<td>Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology</td>
<td>Japan</td>
<td>MIROC5</td>
</tr>
<tr>
<td>19</td>
<td>Max Planck Institute for Meteorology</td>
<td>Germany</td>
<td>MPI-ESM-LR</td>
</tr>
<tr>
<td>20</td>
<td>Max Planck Institute for Meteorology</td>
<td>Germany</td>
<td>MPI-ESM-P</td>
</tr>
<tr>
<td>21</td>
<td>Meteorological Research Institute</td>
<td>Japan</td>
<td>MRI-CGCM3</td>
</tr>
<tr>
<td>22</td>
<td>Norwegian Climate Centre</td>
<td>Norway</td>
<td>NorESM1-M</td>
</tr>
<tr>
<td>23</td>
<td>Norwegian Climate Centre</td>
<td>Norway</td>
<td>NorESM1-ME</td>
</tr>
</tbody>
</table>
Fig. 1. First three EOF patterns of detrended monthly SSTA under Observation and GFDL-CM2.1.
Fig. 2. (a) EV-spectrum projected to the observation. The bars mark the uncertainty interval of the eigenvalues after North et al. (1982) (b) Similar as (a) but projected to CMIP3+5 super model.

(a) PEV of EOF based on Obs.

(b) PEV of EOF based on CMIP3+5
Fig. 3  EV-spectrum of CMIP5 super model and the projected ev-spectrum of its subsamples. The shaded area marks the uncertainty interval of the eigenvalues after North et al. (1982)
Fig. 4. Leading PC structures of monthly SSTA of (a) Observation (HadISST); (b) slab ocean experiment; (c) fitted isotropic diffusion process with inhomogeneous standard deviation forcing (d); as in (c) but with homogeneous forcing.
Fig. 5. Standard deviation fields of monthly SSTA of (a) HadISST; (b) CMIP3 ensemble mean; (c) CMIP5 ensemble mean; (d) slab ocean experiment result.
Fig. 6. EV-spectrum with all the models and null hypotheses relative to the observed. The shaded area marks the uncertainty interval of the eigenvalues after North et al. (1982).
Fig. 7 EV-RMSE errors spreading monthly mean vs. 5yrs. Blue numbers are cmip3 models listed in table 1 and red numbers are cmip5 models in table 2. "●" represents average position of the CMIP models. Blue and red "*" are the results of the CMIP3 and CMIP5 super models, respectively. Green star is the result of the CMIP3+5 super model with observed eigenvalues.
Fig. 8 Leading PC structures of monthly SSTA of GISS-EH, BCM2.0, MRI-CGCM2.3.2, CCSM4, CMIP3 super model and CMIP5 super model.
Fig. 9 Distribution of the domains.
Fig. 10. As in Fig. 6 but in (a). Global summary; (b). Tropical Indian Ocean and Pacific; (c). North Atlantic; (d). Tropical Atlantic; (e). Southern Ocean.
Fig. 11. N_spatial value monthly mean vs. 5yrs. in (a). Global summary; (b). North Pacific; (c). Tropical Indian Ocean and Pacific; (d). North Atlantic; (e). Tropical Atlantic; (f). Southern Ocean.
Fig. 12. Similar as Fig.10 but projected to models.