

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

Gang Wang · Dietmar Dommenges ·
Claudia Frauen

Received: 2 May 2013 / Accepted: 17 April 2014 / Published online: 9 May 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract The natural sea surface temperature (SST) variability in the global oceans is evaluated in simulations of the Climate Model Intercomparison Project Phase 3 (CMIP3) and CMIP5 models. In this evaluation, we examine how well the spatial structure of the SST variability matches between the observations and simulations on the basis of their leading empirical orthogonal functions-modes. Here we focus on the high-pass filter monthly mean time scales and the longer 5 years running mean time scales. 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. Some models show good skill in simulating the observed spatial structure of the SST variability in the tropical domains and less so in the extra-tropical domains. However, most models show substantial deviations from the observations and from each other in most domains and particularly in the North Atlantic and Southern Ocean on the longer (5 years running mean) time scale. In many cases the simple spatial 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 spatial red noise null hypothesis. The CMIP models tend to largely overestimate the effective spatial number degrees of freedom and simulate too strongly localized patterns of SST variability at the wrong locations with structures that are different from the observed. However, the CMIP5 ensemble shows some

improvement over the CMIP3 ensemble, mostly in the tropical domains. Further, the spatial structure of the SST modes of the CMIP3 and CMIP5 super ensemble is more realistic than any single model, if the relative explained variances of these modes are scaled by the observed eigenvalues.

Keywords CMIP · Climate variability · Model evaluation · Eigenvalue projection

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 are different in their structures and physical parameterizations and have shown significant disagreement and uncertainties on their performance (e.g. Gleckler et al. 2008; Jamison and Kravtsov 2010). The aim of the study presented here is to evaluate the CMIP models skill in simulating the natural internal spatial structure of sea surface temperature (SST) variability in all major ocean basins (tropical Indo-Pacific, North Pacific, tropical and North Atlantic and the Southern Ocean, see Fig. 1 for domain boundaries). This should guide the climate community in the understanding of natural modes of SST variability and support the development of seasonal to decadal forecasting systems.

Previous model evaluations 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

Electronic supplementary material The online version of this article (doi:10.1007/s00382-014-2154-0) contains supplementary material, which is available to authorized users.

G. Wang (✉) · D. Dommenges · C. Frauen
School of Mathematical Sciences, Monash University,
Clayton, VIC, Australia
e-mail: gang.wang@monash.edu

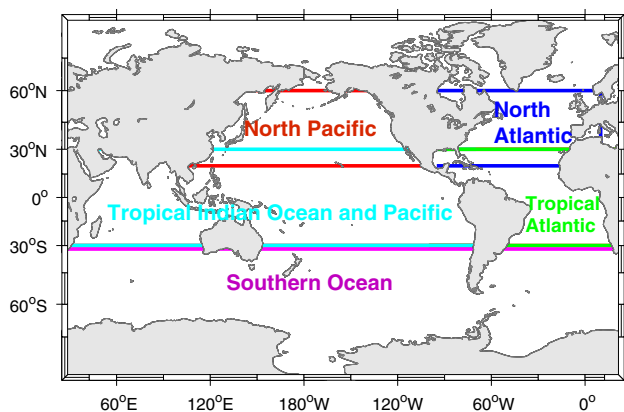


Fig. 1 Illustration of the domain boundaries for the EOF-analyses

2010; Xavier et al. 2010). Missing is a model evaluation of the SST variability on the global scale.

In the study presented here we will base our model evaluation on the comparison of the leading empirical orthogonal function (EOF) modes of SST variability for the different major ocean basins in the model simulations and observations. We will consider the EOF modes of shorter time scales of monthly mean high-pass filtered variability and on longer time scales of 5 years low-pass filtered SST to get some understanding of how the evaluation may change with different time scales. The method of comparing EOF-modes is based on the studies of Dommenget (2007) and Bayr and Dommenget (2014). Similar methods are also discussed in Jolliffe (2002, chapter 13.5) and Krzanowski (1979). This method allows quantifying the agreement in the multi-variate spatial structure of SST variability in a systematic and objective 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 SST variability, as they would result from simplified physical processes such as isotropic diffusion (red noise) or atmospheric forcings only (slab ocean).

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

2 Data and method

2.1 Data

The observed global monthly mean SSTs are taken from the Hadley Centre Sea Ice and SST data set (HadISST,

referred as “observations” below; Rayner et al. 2003) from 1900 to 1999, and the NOAA Extended Reconstructed sea surface temperature data set (ERSST, Smith et al. 2008) was chosen 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. Tables 1 and 2 list all available simulations for this study.

An output of a simple slab ocean coupled experiment is also used to compare versus the CMIP models in this study. The atmospheric component of the model is based on the UK Meteorological Office Unified Model general circulation model with HadGEM2 atmospheric physics (Davies et al. 2005; Martin et al. 2010, 2011). For our study the atmospheric resolution is reduced to N48 ($3.75^\circ \times 2.5^\circ$). For regions with all-year open ocean conditions the model is coupled to a simple slab ocean model (e.g. Washington and Meehl 1984; Dommenget and Latif 2002; Murphy et al. 2004; Dommenget 2010) and otherwise a SST and sea ice climatology based on the HadISST data set is prescribed. A flux correction scheme is used to force the model SST to closely follow the HadISST SST climatology. We take 500 years output from this simulation and divide the data into five 100-year chunks for the analysis. Unless otherwise noted we show the mean result of the five subsamples for the slab ocean data analysis.

All data sets (models and observations) were analysed for the period 1900–1999, interpolated to a common 2.5° latitude \times longitude grid and linearly detrended to remove the global warming signal prior to the analysis. We used a high-pass filter (cut off at 5 years) to obtain the high frequency monthly mean SST anomalies (SSTA) for each model and the observations individually, referred as high pass below. A 5 years-running mean was also used to get the low frequency annual mean SSTA on decadal or longer time scales, referred as low pass below.

In addition to analysing the models individually we also combined all the CMIP3 and CMIP5 ensembles to super model data sets to provide a synthesis. The CMIP3 SSTA (computed for each model individually) were concatenated to generate a CMIP3 super 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 were also constructed. It has to be noted here that combining the anomalies of simulations, which have different modes of variability, will lead to some changes in the characteristics in the EOF-modes. We will point out some of these limitations through out the analysis part.

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

Table 1 List of CMIP3 models

Number	Originating group(s)	Country	Model
1	Bjerknes Centre for Climate Research	Norway	BCCR-BCM2.0
2	Canadian Centre for Climate Modelling & Analysis	Canada	CGCM3.1 (T47)
3	Canadian Centre for Climate Modelling and Analysis	Canada	CGCM3.1 (T63)
4	Météo-France/Centre National de Recherches Météorologiques	France	CNRM-CM3
5	CSIRO Atmospheric Research	Australia	CSIRO-Mk3.0
6	CSIRO Atmospheric Research	Australia	CSIRO-Mk3.5
7	US Dept. of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.0
8	US Department of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.1
9	NASA/Goddard Institute for Space Studies	USA	GISS-AOM
10	NASA/Goddard Institute for Space Studies	USA	GISS-EH
11	NASA/Goddard Institute for Space Studies	USA	GISS-ER
12	LASG/Institute of Atmospheric Physics	China	FGOALS-g1.0
13	Instituto Nazionale di Geofisica e Vulcanologia	Italy	INGV-SXG
14	Institute for Numerical Mathematics	Russia	INM-CM3.0
15	Institut Pierre Simon Laplace	France	IPSL-CM4
16	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC3.2 (hires)
17	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC3.2 (medres)
18	Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-OM
19	Meteorological Research Institute	Japan	MRI-CGCM2.3.2
20	National Center for Atmospheric Research	USA	CCSM3
21	National Center for Atmospheric Research	USA	PCM
22	Hadley Centre for Climate Prediction and Research/Met Office	UK	UKMO-HadCM3
23	Hadley Centre for Climate Prediction and Research/Met Office	UK	UKMO-HadGEM1

EOF-modes, assuming that the leading EOF-modes give a good representation of the large-scale SST variability. This is done by defining the EOF-modes of one data set as the reference modes and projecting 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 projected data set. This concept is based on Dommenget (2007) and Bayr and Dommenget (2014) and briefly summarized here:

An EOF eigenvector (mode) of the reference data set **A**, \vec{E}_i^A and its corresponding eigenvalue (EV) e_i^A are compared with the eigenvector \vec{E}_j^B and eigenvalue e_j^B of another data set **B** by projecting the eigenvectors \vec{E}_i^A onto the \vec{E}_j^B ,

$$c_{ij} = \frac{\vec{E}_i^A \vec{E}_j^B}{|\vec{E}_i^A| |\vec{E}_j^B|} \tag{1}$$

where c_{ij} is the uncentered pattern correlation coefficient describing the spatial similarity between the two EOF-patterns. The projected explained variance (PEV) of mode \vec{E}_i^A in data set **B**, $pe_i^{A \rightarrow B}$, is estimated by the accumulation of all eigenvalues of **B** (Dommenget 2007):

$$pe_i^{A \rightarrow B} = \sum_{j=1}^N c_{ij}^2 e_j^B \tag{2}$$

The value $pe_i^{A \rightarrow B}$ represents the total variance of data set **B** that is explained by the reference mode \vec{E}_i^A , with N the number of EOF-modes considered. In this paper N is set to be 60, which is sufficient to give us stable results. Increasing N does not change the outcomes in any of the domains. The $pe_i^{A \rightarrow B}$ values do not need to be monotonically decrease, as the e_i^A do, since an EOF-mode of **A** may explain more variance in the data set **B** than it does in **A** and vice versa.

We illustrate this method in a simple example of comparing two data sets: In Fig. 2a–f we show the leading EOF modes of the North Pacific for the observations and the GFDL-cm2.1 model simulation. We note here that the two data sets have slightly different leading modes of variability and that the explained variances of each of these modes are also slightly different. To compare the overall spatial structure of variability in the two data sets we choose the observed EOF-modes as the reference modes (\vec{E}_i^A) and project the EOF-modes of the GFDL-cm2.1

Table 2 List of CMIP5 models

Number	Originating group(s)	Country	Model
1	CSIRO and BOM	Australia	ACCESS1.0
2	Beijing Climate Center, China Meteorological Administration	China	BCC-CSM1.1
3	National Center for Atmospheric Research	USA	CCSM4
4	Météo-France/Centre National de Recherches Météorologiques	France	CNRM-CM5
5	Canadian Centre for Climate Modelling and Analysis	Canada	CanESM2
6	Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM3
7	Geophysical Fluid Dynamics Laboratory	USA	GFDL-ESM2G
8	Geophysical Fluid Dynamics Laboratory	USA	GFDL-ESM2M
9	NASA/Goddard Institute for Space Studies	USA	GISS-E2-H
10	NASA/Goddard Institute for Space Studies	USA	GISS-E2-R
11	Hadley Centre for Climate Prediction and Research/Met Office	UK	HadCM3
12	Hadley Centre for Climate Prediction and Research/Met Office	UK	HadGEM2-CC
13	Hadley Centre for Climate Prediction and Research/Met Office	UK	HadGEM2-ES
14	Institute for Numerical Mathematics	Russia	INM-CM4
15	Institut Pierre Simon Laplace	France	IPSL-CM5A-LR
16	Institut Pierre Simon Laplace	France	IPSL-CM5A-MR
17	Institut Pierre Simon Laplace	France	IPSL-CM5B-LR
18	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	Japan	MIROC5
19	Max Planck Institute for Meteorology	Germany	MPI-ESM-LR
20	Max Planck Institute for Meteorology	Germany	MPI-ESM-P
21	Meteorological Research Institute	Japan	MRI-CGCM3
22	Norwegian Climate Centre	Norway	NorESM1-M
23	Norwegian Climate Centre	Norway	NorESM1-ME

model simulation (\vec{E}_i^B) onto these modes. Figure 3a shows the eigenvalues of the observed EOF-modes (e_i^A) against the PEV from the GFDL-cm2.1 model simulation ($pe_i^{A \rightarrow B}$). The leading observed mode (EOF-1 in Fig. 2a), for instance, explains only half as much variance in the GFDL-cm2.1 model simulation (red line in Fig. 3a) than it does in the observations (black line in Fig. 3a). In turn, the observed EOF-10, for instance, mode has more relative explained variance in the GFDL-cm2.1 simulation. Therefore, comparing the overall spatial structure of variability essentially means to estimate how much variance each of the reference modes explains in both data sets, and essentially quantify the discrepancy based on the mismatch of the variances explained.

We 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 (grey band in Fig. 3a) in the explained variances is quantified by a normalized root-mean-square error (RMSE_{EOF}) between the e_i^A and $pe_i^{A \rightarrow B}$ values:

$$RMSE_{EOF}(A, B) = \sqrt{\frac{\sum_{i=1}^{N_A} (pe_i^{A \rightarrow B} - e_i^A)^2}{\sum_{i=1}^{N_A} (e_i^A)^2}} \quad (3)$$

The normalization allows a better comparison of the RMSE_{EOF} values among different domains with different sampling uncertainties. Here N_A corresponds to the number of EOF modes considered. The sum RMSE_{EOF} (A, B) is dominated by the mismatches in the leading modes between e^A and $pe^{A \rightarrow B}$, as they have larger uncertainties. Subsequently the uncertainties in EOF 1–3 are what dominates RMSE_{EOF} (A, B). Most of the results presented in this study are indeed very much the same if only the first two leading EOFs are considered. However, in domains with larger effective spatial number degrees of freedom, $N_{spatial}$ (Bretherton et al. 1999) the higher order modes will also contribute to RMSE_{EOF} (A, B) significantly.

In our analysis we choose N_A equal to $N_{spatial}$ to provide a consistent estimate over the whole multi-variate variability in each domain:

$$N_{spatial} = \frac{1}{\sum (e_i)^2} \quad (4)$$

$N_{spatial}$ varies considerably from domain to domain and for the different time-scales considered (see Fig. 4). For the domains and time-scales we considered in this analysis

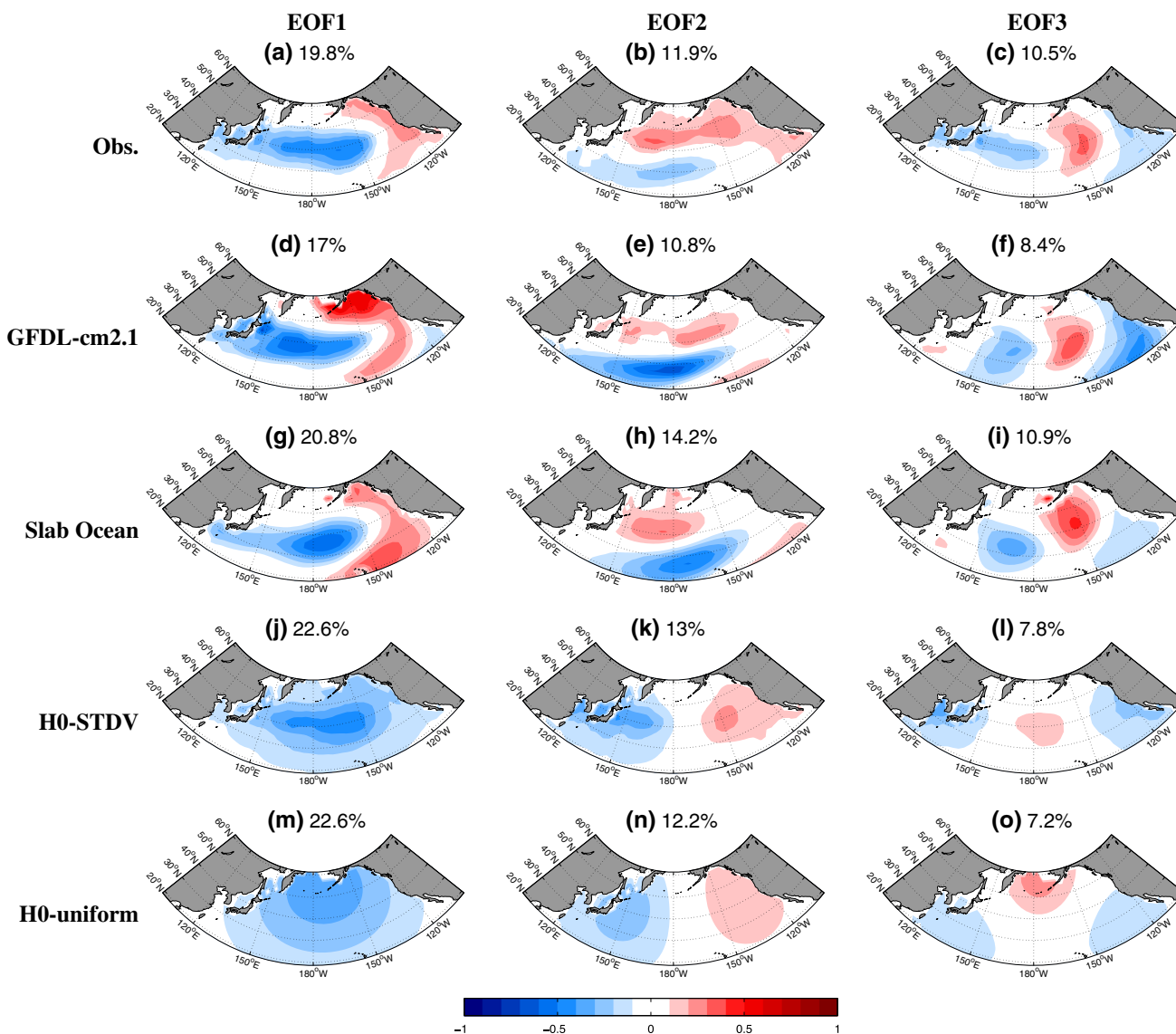


Fig. 2 First three leading EOF patterns of detrended monthly SSTa in the North Pacific after 5-year high-pass filter for **a–c** observations (HadISST); **d–f** GFDL-cm2.1; **g–i**: slab ocean experiment; **j–l** fitted isotropic diffusion process with inhomogeneous standard deviation

forcing; **m–o** as in **j–l** but with homogeneous forcing. The values in the headings of each panel are the explained variances of each EOF-mode

$N_{spatial}$ is between 3 (e.g. North Atlantic longer time scale) and 50 (e.g. South Ocean shorter time scale).

Figure 3 shows the $RMSE_{EOF}$ values for two examples. A $RMSE_{EOF}$ value of 100 % corresponds to errors that are as big as the eigenvalues. Thus, the $RMSE_{EOF}$ value (32 %, see Fig. 3a) of the GFDL-cm2.1 model relative to the observations reflects an uncertainty of the leading EOF-modes of about 32 % of the eigenvalues, which is a substantial uncertainty. Similarly, the relative smaller $RMSE_{EOF}$ value (17 %) in Fig. 3b, which represents an uncertainty of about 17 % within the leading modes, suggests less fluctuation and better matches between the PEV of GFDL-cm2.1 and the eigenvalues of the CMIP3 + 5 super model.

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 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 patterns of SST variability: the first is the slab ocean models modes of

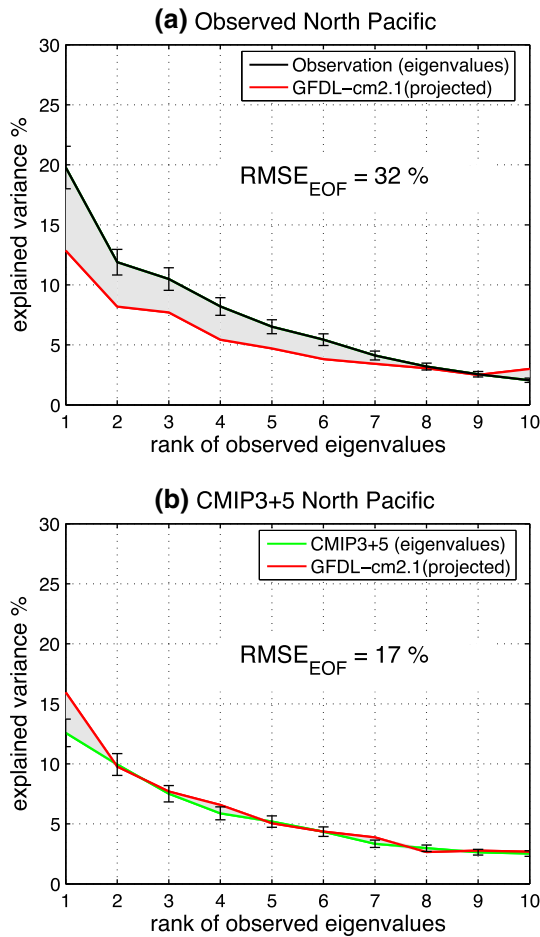


Fig. 3 **a** Eigenvalue spectrum of the observed leading EOF-modes in the North Pacific and the PEV values of the GFDL-CM2.1. The bars mark the sampling uncertainty interval of the eigenvalues after North et al. (1982) **b** as **a** but for the leading EOF-modes of the CMIP3 + 5 super model in the North Pacific

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 with two different assumptions for the spatial distribution of variances.

3.1 Sampling uncertainties of eigenvalues

North et al. (1982) give the statistical uncertainties of the eigenvalues e_i due to sampling errors:

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

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

$$RMSE_{EOF}(\delta e_i^A) = \sqrt{\frac{\sum_{i=1}^{N_A} (\delta e_i^A)^2}{N_A}} / \sqrt{\frac{\sum_{i=1}^{N_A} (e_i^A)^2}{N_A}} = \sqrt{\frac{2}{N_{sample}^A}} \tag{6}$$

Thus, the $RMSE_{EOF}(\delta e_i)$ could be deemed as the confidence level of a data set. If two data sets are just different stochastic realizations of the same process (have the same spatial patterns of SST variability), we expect the $RMSE_{EOF}$ to be in average $RMSE_{EOF}(\delta e_i)$.

Figure 5 illustrates an example based on subsampling the CMIP5 super model. Here we computed the EOF-modes for the North Pacific for high-pass (5 years) and low-pass (5 years) SST. For subsampling the CMIP5 super ensemble we split the data from each model into average chunks with 5 years/60 months data and concatenated the chunks into the CMIP5 super subsamples. Here, the total number of samples N_{len} is set to be 1,380 (60 * 23) for the high pass and 115 (5 * 23) for the low pass. For high-pass data analysis the average decorrelation time of the leading five principal components $t_d = 8.8$. Therefore, we get $N_{sample} = 1,380/8.8 = 156.8$. Similarly $N_{sample} = 115/5.8 = 19.8$ for the low-pass analysis.

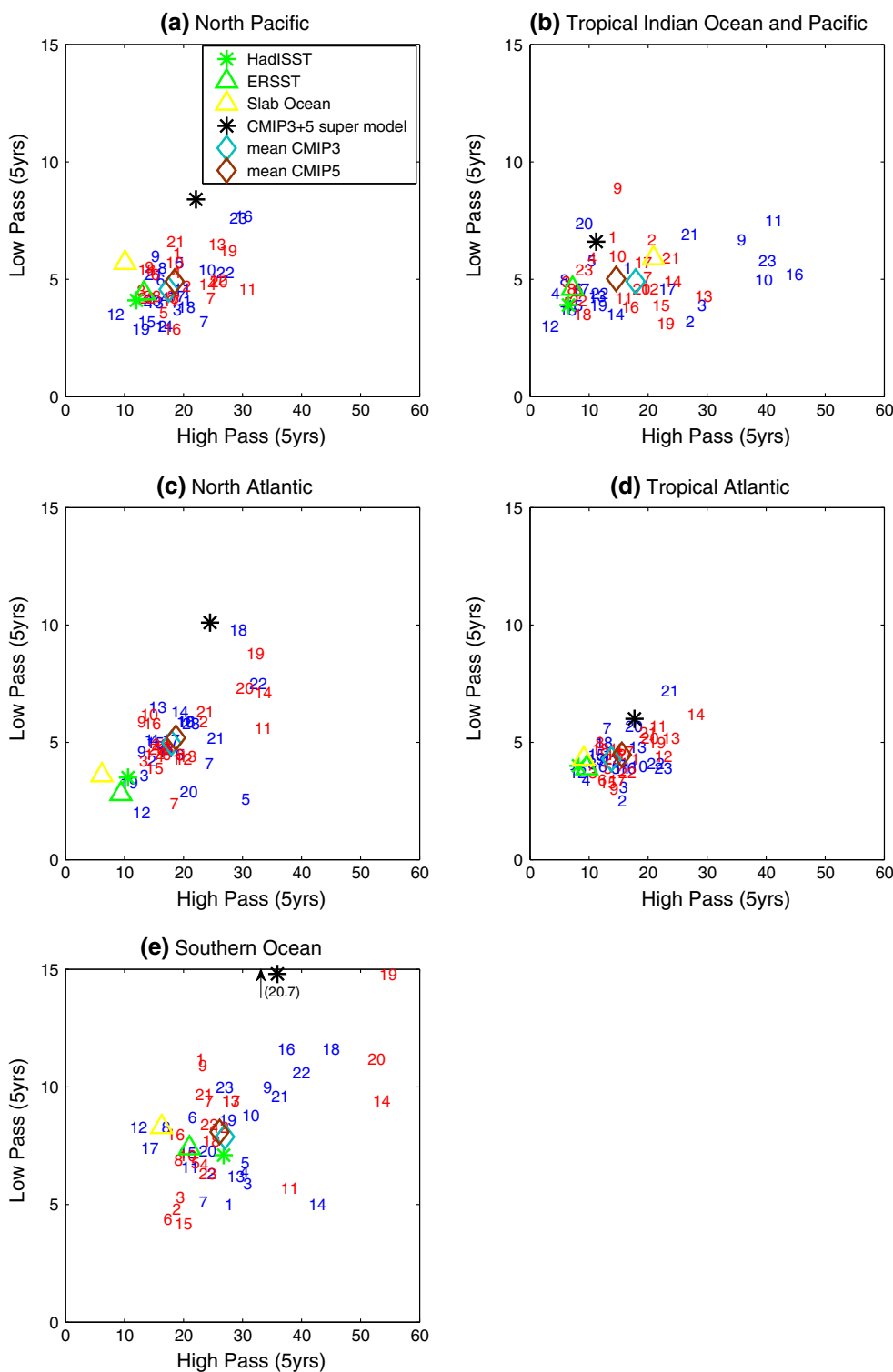
We note from Fig. 5a, b that the subsamples $pe_i^{A \rightarrow B}$ values fluctuate around the e_i^A values of the CMIP5 super model, caused here by sampling uncertainties and not by differences in the physical system. Subsequently, the $RMSE_{EOF}$ values of the subsamples fluctuate around the $RMSE_{EOF}(\delta e_i)$ (see Fig. 5c). For the 5 years-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 as they are from the same data sets.

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) estimates the spatial structure of the SST variability that is caused by the atmosphere only. Compared to a fully CGCM, the slab ocean model only introduces the error of atmospheric simulation without significant error addition or amplification by the coupling process or ocean dynamics. It is therefore a good null hypothesis to evaluate the models: if the spatial structure of the SST variability agrees better with the slab ocean model (smaller $RMSE_{EOF}$) than with a CGCM, this is indicating that the coupling procedures and ocean dynamics are causing unrealistic SST patterns.

Figure 2 shows the observed EOF-modes of the high-pass monthly mean SST variability in the North Pacific in

Fig. 4 N_{spatial} values for high-pass and low-pass EOF-analysis of SSTA in **a** North Pacific; **b** tropical Indian Ocean and Pacific; **c** North Atlantic; **d** tropical Atlantic; **e** Southern Ocean. The *black asterisk* in **e** is out of range with a real ordinate value of 20.7. The N_{spatial} values of all CMIP3 and CMIP5 model are listed in Supplemental Table 1

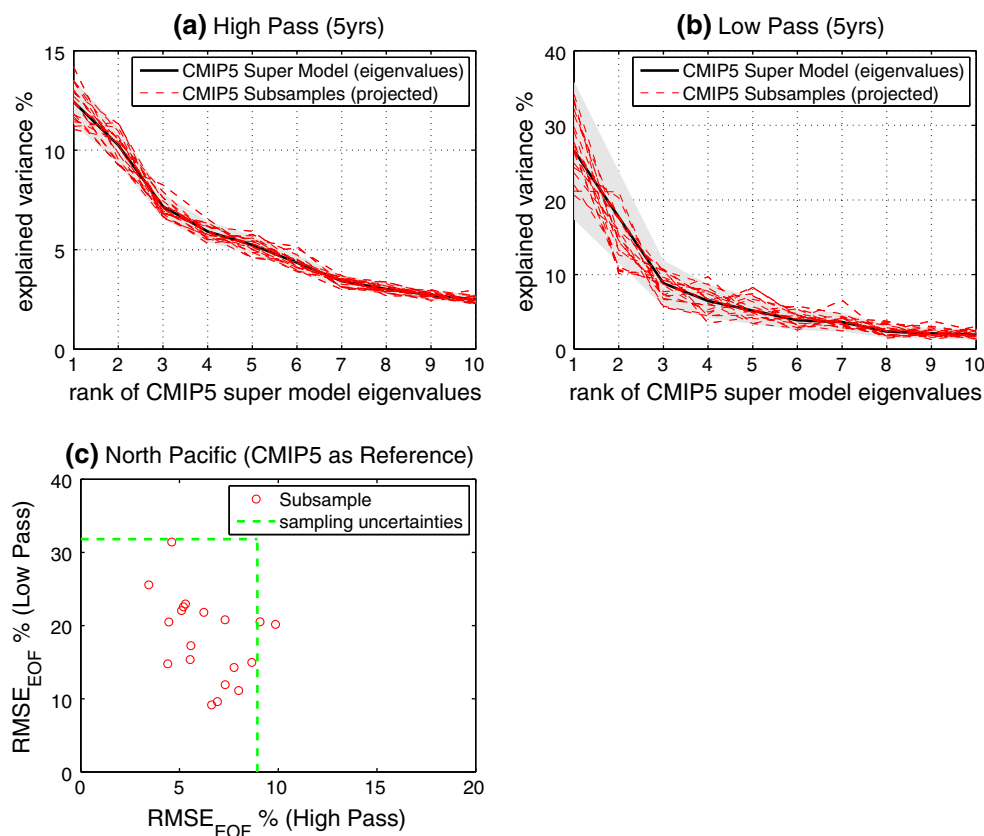


comparison with the EOF-modes of the slab ocean simulation, for an example. The EOF-modes of the slab simulation are already quite realistic, suggesting that much of the large-scale structure of the SST variability is to first order simulated in the slab simulation, which is consistent with what has been found in other studies as well (e.g. Pierce et al. 2001).

3.3 Isotropic diffusion

Cahalan et al. (1996) and Dommenges (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,

Fig. 5 Eigenvalue spectrum of CMIP5 super model and the PEV of its subsamples in the North Pacific. The *shaded area* marks the uncertainty interval of the eigenvalues after North et al. (1982) **a** high-pass result; **b** low-pass result; **c** RMSE_{EOF} values of the subsamples in **a**, **b**



starting with a monopole (largest scale), followed by a dipole, and followed by multi-poles with increasing complexity (smaller scales). It essentially represents a spatial red noise process (Dommenget 2007). Figure 2m–o illustrates the EOF-modes 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. 6), but assuming the same $N_{spatial}$ 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 $H0_{uniform}$.

This null hypothesis is further extended by also using the observed inhomogeneous SST STDV field (see Fig. 6a, b) 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).

Much of the spatial structure in SST variability is already highlighted by the spatial structure in the SST STDV field (see Fig. 6a, b). The regions with large SST STDV will be the regions where most of the leading EOF-modes have large variance. Here we can already see that the SST STDV deviation is different on different time scales, which will be reflected in slightly or significantly different SST modes. The CMIP3 and CMIP5

ensembles capture most of the main observed structures and even the slab ocean simulation captures some of it (see Fig. 6c–h).

Figure 2j–l illustrates the EOF-modes 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 modes are centred on regions of large SST STDV (for comparison with STDV of observations see Fig. 6a). These EOF-modes now have some more realistic features. We refer to this null hypothesis as $H0_{STDV}$. Thus, we list the comparison objects in Fig. 2 essentially from most realistic or complex (observed) to least or simplest ($H0_{uniform}$), as the modes are becoming more unrealistic from top to bottom.

4 Comparison of the eigenvalue (EV)-spectrum

In this section we present the main results of this analysis, which is based on the comparison of the $pe_i^{A \rightarrow B}$ values (referred as EV-spectrum below) of the SST variability in different ocean basins. We will define SST variability on two different time scales to highlight potential differences in the variability modes.

First we like to illustrate that the EOF-modes in the high-pass and the low-pass SST variability are indeed

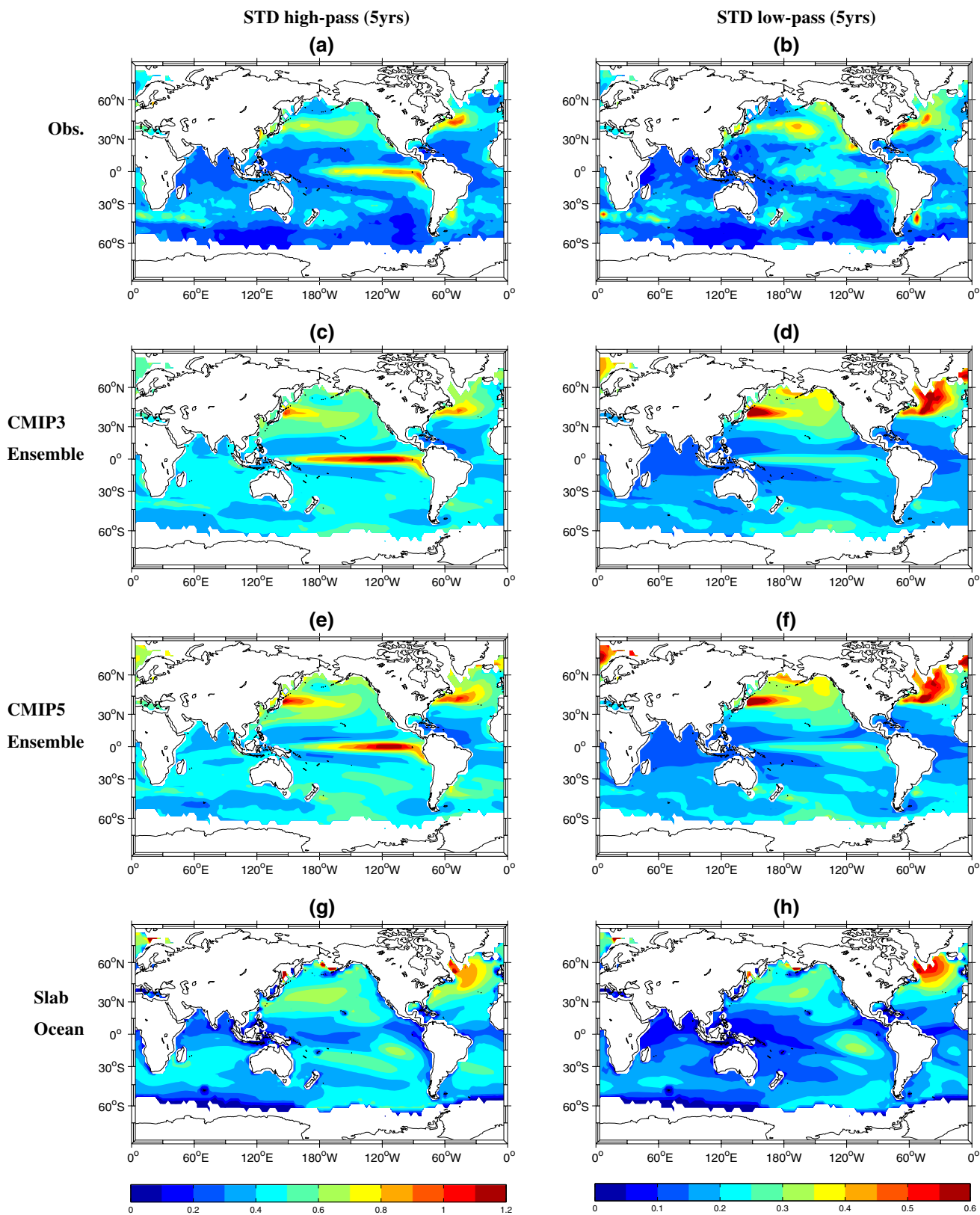


Fig. 6 Standard deviation fields of SSTA for **a, b** HadISST; **c, d** CMIP3 ensemble mean; **e, f** CMIP5 ensemble mean; **g, h** slab ocean experiment result. **a, c, e, g** after high-pass filter; **b, d, f, h** after low-pass filter

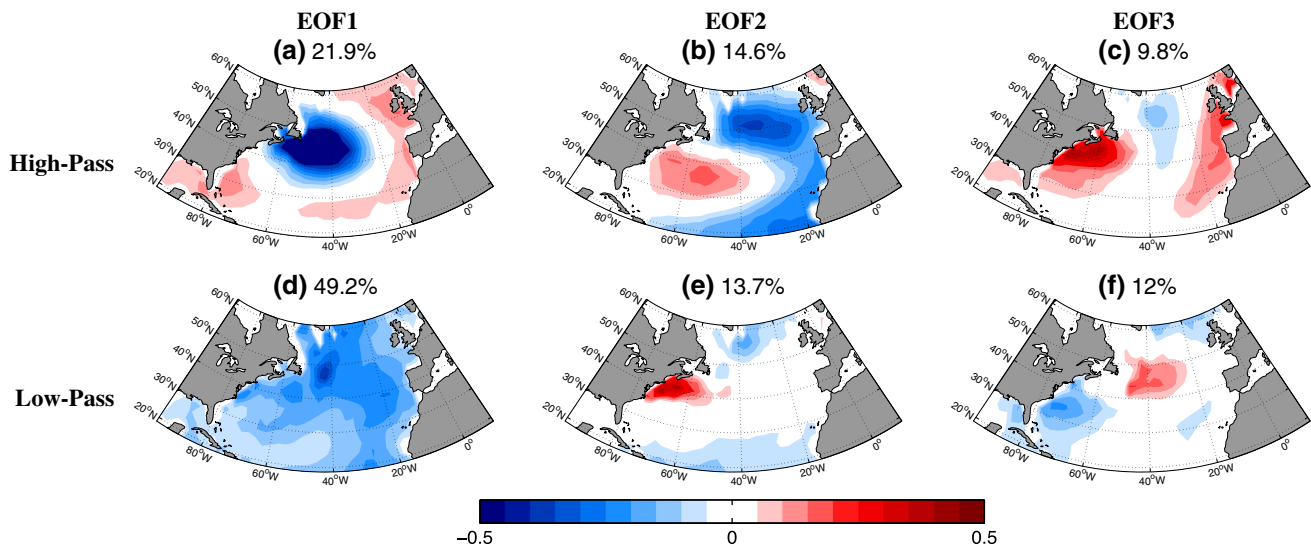


Fig. 7 First three leading EOF patterns of observed SSTA in the North Atlantic for **a–c** after high-pass filter; **d–f** after low-pass filter

different. Figure 7 shows the different EOF-modes structures on high-pass and low-pass scale in the North Atlantic. The following is noted here:

- The spatial patterns are quite different between the two time scales. The high-pass, for instance, reveals three tri-pole modes from EOF-1 to EOF-3. However, none of them is strongly related to the leading EOF-modes of the low-pass modes.
- The eigenvalues of the low-pass modes are much larger than their counterpart of high-pass ones. Subsequently $N_{spatial}$ is much larger for the high-pass SST ($N_{spatial} = 10$) than for the low-pass SST ($N_{spatial} = 3$). Thus the high-pass SST has more complex variability modes than the low-pass SST.

Similar findings can be made for all domains, but the EOF-modes of the different time scales may be more similar in the other domains than they are in the North Atlantic.

We start the main 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

Figure 8 shows the EV-spectrum of the observed high-pass SSTA in the North Pacific region together with the projected $pe_i^{A \rightarrow B}$ values for all CMIP models and the four different null hypotheses references. A few points should be noted here:

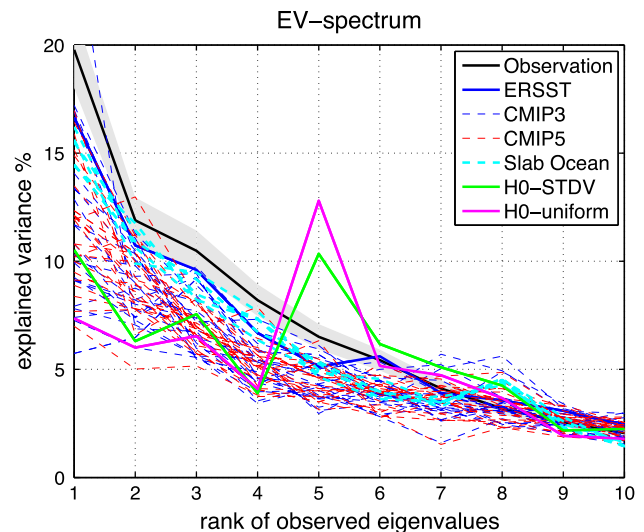
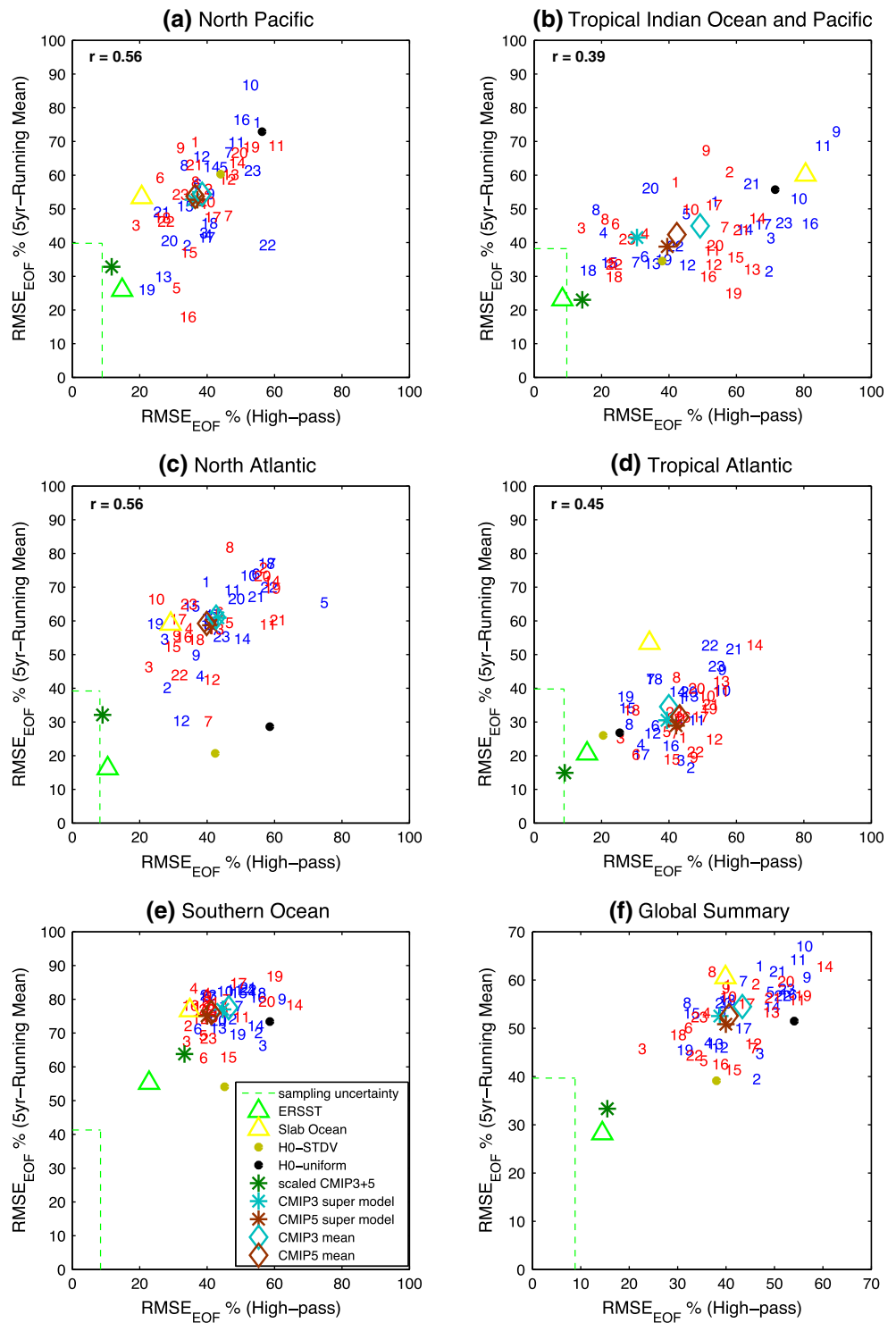


Fig. 8 Eigenvalue spectrum of the observed leading EOF-modes in the North Pacific as in Fig. 3a, but compared with the PEV values for all CMIP models, the slab ocean simulations and the isotropic diffusion null hypotheses. The shaded area marks the uncertainty interval of the eigenvalues after North et al. (1982)

- 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 observations, namely the Pacific decadal oscillation (PDO) pattern (e.g. Mantua et al. 1997) and most of the other leading modes.
- The deviations of the individual models from the observed EV-spectrum are much larger than expected

Fig. 9 $RMSE_{EOF}$ values relative to the observations for high-pass and low-pass EOF-analysis of **a** North Pacific; **b** tropical Indian Ocean and Pacific; **c** North Atlantic; **d** tropical Atlantic; **e** Southern Ocean; **f** global summary. Noting that the axis range in **(f)** is different from that in **a–e**. *Blue numbers* are CMIP3 models and *red numbers* are CMIP5. *Diamonds* represents the average position of the CMIP models. *Blue and red stars* are the results of the CMIP3 and CMIP5 super models, respectively. The *green star* is the result of the CMIP3 + 5 super model scaled with observed eigenvalues. The letter “*r*” shows the correlation coefficient of the $RMSE_{EOF}$ values between two time scales based on models only. The $RMSE_{EOF}$ values of all CMIP3 and CMIP5 model are listed in Supplemental Table 2



by the sampling uncertainties δe_i . The slab ocean simulation is closer to the observations than most models.

- The deviations of $H0_{STDV}$ are about as strong as for most of the CMIP model simulations. However, the deviations of $H0_{uniform}$ appear to be larger than those of most CMIP models. The H0 curves both have a peak at

the EOF-5. This is due to the similarity of the observed EOF-5 with a basin wide monopole (not shown), which is the leading mode in isotropic diffusion (both $H0_{STDV}$ and $H0_{uniform}$).

The results of the EV-spectrum are quantified by the $RMSE_{EOF}$ values for the high-pass monthly mean SST as

shown in Fig. 9a on the x-axis and for the low-pass SST on the y-axis. In addition to what we have already concluded above for the high-pass SST modes we should 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 contain 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 30–80 % of the eigenvalues. These errors are substantial.
- The RMSE_{EOF} values for the different time scales in Fig. 9a have a mostly linear relationship with a correlation coefficient 0.6 indicating that in this region the models show similar performance for high-frequency and low-frequency variability. However, there are also significant deviations from the linear relationship, which indicates that some models are performing good on one time scale but not as good on the other time scale.
- Most models seem better than the $\text{H0}_{\text{uniform}}$; however, half of the models are not as good as the H0_{STDV} hypothesis.
- The slab ocean simulation is closer to the observations than most models in the high-pass variability, but is about average for the low-pass variability.
- 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 deviations are in the CMIP3 ensemble.

We pick out a few models to illustrate how the modes of variability in some models deviate from those observed. In Fig. 10 we show the leading modes of the two models that deviate the most (HadCM3 and BCM2.0), the two models closest to the observations (MRI-CGCM2.3.2 and CCSM4) and the CMIP3 and CMIP5 super models in the North Pacific modes comparison. The following is noted here:

- The two models that deviate the most both show leading EOF-modes that are somewhat different in structure from the observed. For instance, they tend to have the negative anomaly centres for the PDO-like mode (EOF-1) more to the west and more focused on a small region than in the observed PDO (EOF-1). Further, the eigenvalues of EOF-1 are much smaller than observed.
- The two models closest to the observations 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 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. To summarize, we also average the results and get the global summary of the RMSE_{EOF} diagram in Fig. 9f. The results show a number of interesting aspects. We start the discussion with a focus on the individual domains, starting with the tropical Indo-Pacific domain (Fig. 9b):

- On the high-pass time scales the spread in the quality of the models is very large. Some models are close to the observed modes, but most models are quite different from the observed modes. On the low-pass time scale the skills of models are more similar and many models are as close to the observed modes as expected from sampling uncertainties.
- The slab model is quite different from the observed modes on the high-pass time scale and the longer time scale. The El Niño 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 H0_{STDV} performs better than most models, but the $\text{H0}_{\text{uniform}}$ hypothesis clearly deviates more than most models and is very different from the observed structure.
- The CMIP3 ensemble has much more outliers, in particular on the shorter time scale, than the CMIP5 ensemble.

Similar to the North Pacific domain, we picked out a few models to demonstrate the differences of the modes (Supplemental Fig. 1). The model GISS-AOM that deviates the most shows no El Niño pattern. The model ECHAM5-MPI-OM, which is closest to the observed, as well as CMIP super models all display ENSO pattern and Central-Pacific El Niño pattern (Kao and Yu 2009) on the leading two modes, close to the observation.

In the North Atlantic the picture is quite different (Fig. 9c):

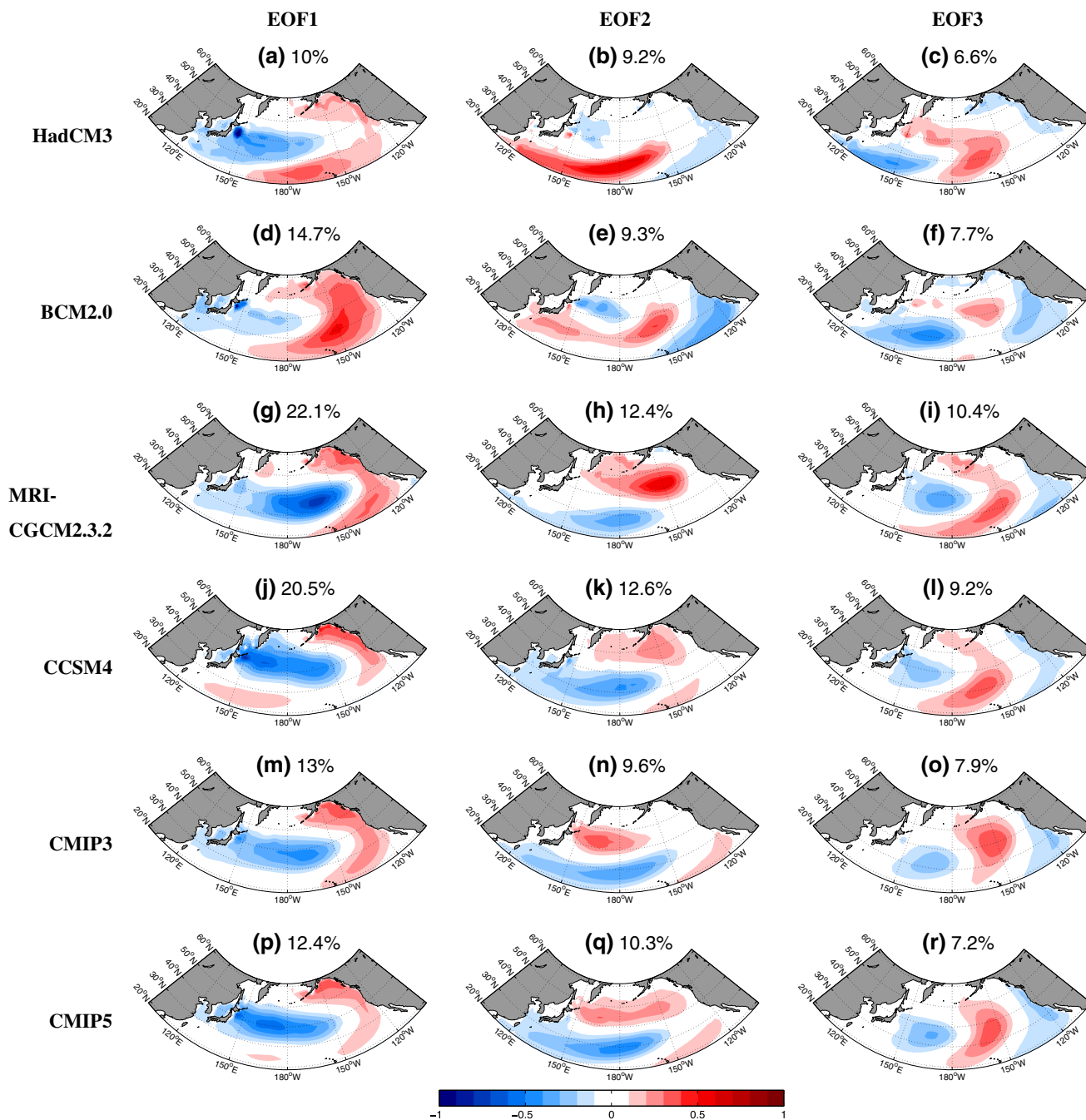


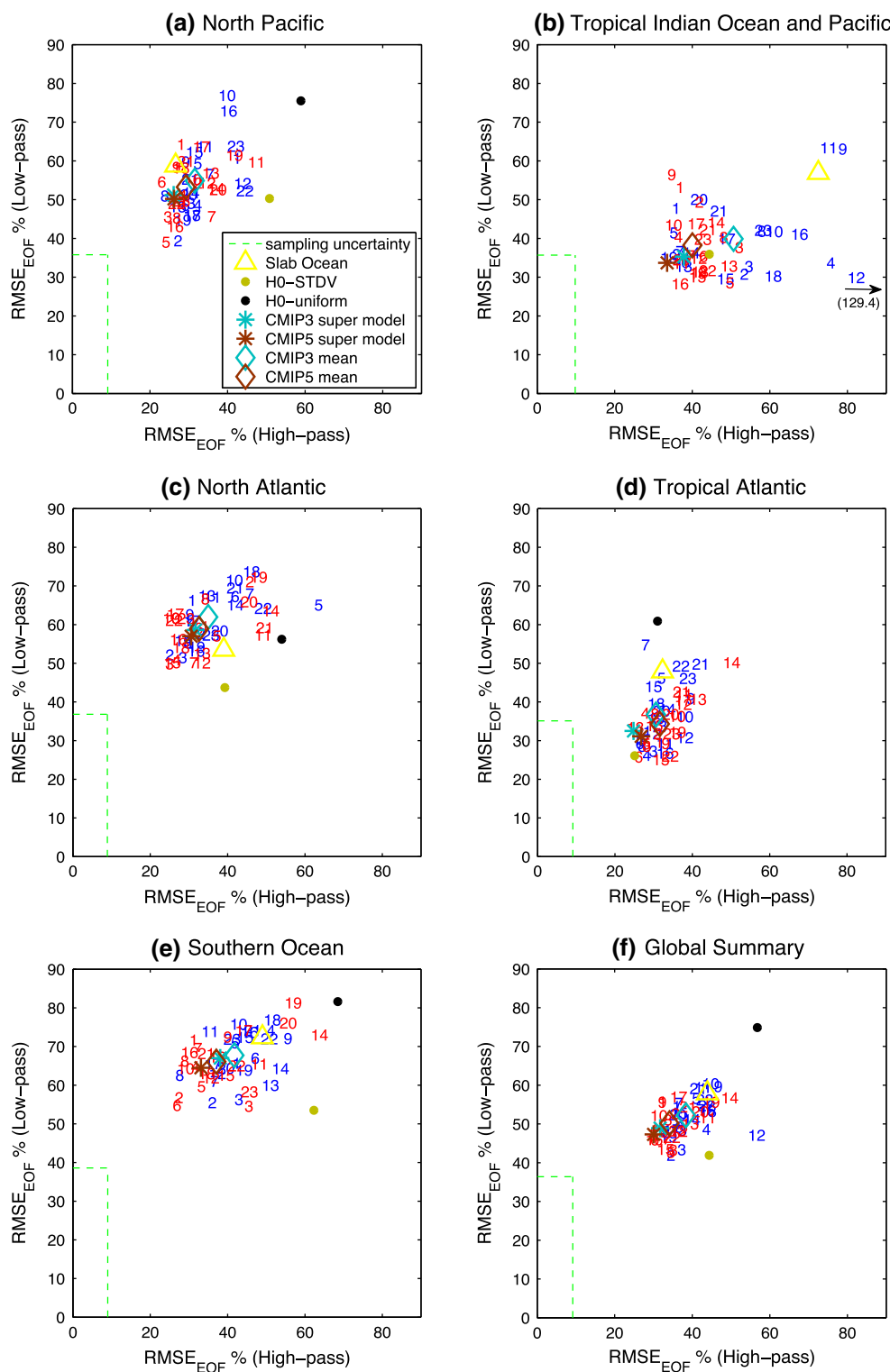
Fig. 10 Leading EOF-modes as in Fig. 2, but for a selection of model simulations

- The most remarkable feature is that both simple isotropic diffusion null hypotheses are closer to the observed modes than most of the CGCM simulations especially on long-term scale. First of all this is due to the fact that the observed modes in the North Atlantic are indeed more similar to the isotropic diffusion 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. This conclusion appears to be quite different to what Jamison and Kravtsov (2010) conclude from their analysis of the leading modes in the North Atlantic. However, when we evaluate their Figs. 10–12 we would assume that the quantitative and objective error values $RMSE_{EOF}$ based on their analysis results should be similar to ours.

- The agreement with the observed modes is better on the shorter time scale than on the longer low-pass time

Fig. 11 Mean $RMSE_{EOF}$ values of pairwise model comparisons for high-pass and low-pass EOF-analysis in **a** North Pacific; **b** tropical Indian Ocean and Pacific; **c** North Atlantic; **d** tropical Atlantic; **e** Southern Ocean; **f**: global summary. Noting that the blue “12” in **(b)** is out of range with a real abscissa value of 129.4. The $RMSE_{EOF}$ values of all CMIP3 and CMIP5 model are listed in Supplemental Table 3



- The slab model performs better than any of the CGCM simulations on the high-pass time scale and still better than many models on the longer time scale.

- There is no substantial difference in the performance of the CMIP3 and CMIP5 ensembles.

Supplemental Fig. 2 again shows several models for comparison in the North Atlantic. The tropical Atlantic has again some interesting features (Fig. 9d):

- 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 scale this domain actually seems 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 time scale, which is 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 closer to the observations than most CGCM simulations on the high-pass (5 years) time scale, but deviates more than all models on the longer time scale.
- Both the $H0_{STDV}$ and the $H0_{uniform}$ 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 multipole deconstruction, as it follows from an isotropic diffusion process, would already explain most of the SST variability in this domain.

Supplemental Fig. 3 shows the modes of some models in this domain. Finally, the Southern Ocean (Fig. 9e):

- 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 isotropic diffusion null hypotheses are substantially different from the observed modes on both time scales. However, the $H0_{STDV}$ hypothesis is closer to the observed modes than any CGCM simulation on the longer time scale and closer than most on the shorter time scale.

The Southern Ocean is a special domain for its sparse in situ observation, which introduces non-negligible uncertainties of observed reference. The modes comparison (see Supplemental Fig. 4) is essentially different to other domains, as we can't find too much similarity here between the leading modes of observed and the CMIP super models especially on EOF-1. The Southern Ocean is also one of the largest domains; it involves the more complex extra-tropical dynamics (larger $N_{spatial}$, see Fig. 4) and interactions with sea ice, which may explain the large disagreement to some part.

The summary of all individual domains (Fig. 9f) shows the average skill of the models:

- First of all we note that the models skills on the short and longer time scales are roughly linearly related with

a correlation of 0.6. Models that are close to the observed modes on the shorter time scale tend to be close to the observed modes on the longer time scale as well.

- Basically all models show significant deviations in the spatial structure from the observed modes. These deviations are in the order of 50 % of the eigenvalues on the leading modes. This means they in average under-/over-estimate 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 CCSM4 model is closest) than others 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 models that have big $RMSE_{EOF}$ in some domains often have 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 super model 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 scale, but has less skill than most models on the longer time scale. Nevertheless, on both time scales the slab simulation is not consistent with observations ($RMSE_{EOF}$ is larger than expected by sampling uncertainties).
- The simple $H0_{STDV}$ hypothesis is in average closer to the observed modes than any CGCM simulation on the longer time scale and closer than most simulations on the shorter time scale. Even the $H0_{uniform}$ hypothesis is better than many models. This suggests that knowing the $N_{spatial}$, the domain geometry, the SST STDV field (most importantly) and assuming a modal structure resulting from isotropic diffusion could already describe 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. A closer look at the leading EOF-modes of the model simulations (not all shown, but some are shown in Fig. 10 and in the Supplemental Figs. 1–4) reveals why the models differ from the observations:

- First, we see that $N_{spatial}$ is larger than observed in most CMIP models and for all domains and on both time scales (see Fig. 4). 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 is also seen by visual inspection of the leading modes of all the model simulations (not shown).

- Further, we note that the patterns of the leading modes in the model simulations are often different from those observed. 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 smaller 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 isotropic diffusion null hypotheses and the slab simulation do not have these localized structures and therefore do not have these double errors.

Additionally, it is noticeable that the super model ensembles seem not to 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 (e.g. Tebaldi and Knutti 2007; Reifen and Toumi 2009; Santer et al. 2009; Knutti et al. 2010). 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, then the super model will have all of these modes, but each with a smaller eigenvalue, which increases $N_{spatial}$ of the ensemble super model (see Fig. 4).

However, we can illustrate that the ensemble super model is indeed containing some useful information that improves the presentation of the spatial structures of variability compared to most individual models. If we replace the eigenvalues of the ensemble super model in Eq. (2) against observed eigenvalues, they are much closer to the observed spatial structure than most models. These scaled values are shown in Fig. 9. Even if we replace the eigenvalues of all models against observed eigenvalues, which clearly decreases the $RMSE_{EOF}$ values of most models, the ensemble super models still demonstrate smaller $RMSE_{EOF}$ values than the majority of the individual models, while the CMIP5 super model is a bit closer to the observations than its counterpart of CMIP3 (not shown). It illustrates that the spatial structure of the leading modes of variability in the ensemble super model are indeed realistic, but the relative explained variance of each mode is underestimated by the ensemble super model due to the artificial diversity in the individual models.

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

- For most domains there is a relative good agreement between the ERSST and the HadISST data sets 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 Ocean. Here, the spatial structure of the observed SST variability is very uncertain. On the longer time scale 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 observations, which revealed 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 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 models have strong differences in the spatial structures in SST variability between each other. These model-to-model differences are quantified by repeating the above analysis by pairwise comparison of the EOF-modes in the CMIP3 and 5 ensembles.

Figure 11 shows the $RMSE_{EOF}$ values as in Fig. 9, 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 to the spatial structures in SST variability of all the other models and vice versa for large $RMSE_{EOF}$ values. The following features are noted in this comparison:

- All models show $RMSE_{EOF}$ values larger than expected from sampling uncertainties [$RMSE_{EOF}(\delta e_i)$]. Thus, the models substantially disagree with 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 suggests that these models poorly represent the ENSO pattern and are indeed quite different from the overall model ensemble on the monthly mean scale.

5 Summary and discussion

In the study presented here we evaluated the skill of the CMIP3 and CMIP5 models in simulating the observed spatial structure of SST variability on interannual and decadal time scales. This comparison was based on a quantitative and objective 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. Our main quantitative measures were the $RMSE_{EOF}$ values, which estimates the disagreement in the leading EOF-modes. Although this considers all leading EOF-modes up to the $N_{spatial}$ value of each domain, it essentially focussed on the two to three leading EOF-modes. The higher order EOF-modes are considered as well, they, however, only matter in domains with larger $N_{spatial}$ values. 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 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 both 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 scale in the tropical Indo-Pacific and Atlantic and on both time scales in the North Pacific and Southern Ocean. The longer time scale 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 numbers of spatial degrees of freedom ($N_{spatial}$) are between 5 and 10 for most domains on the monthly mean time scale and smaller (<5) on the longer 5 years time scale. 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 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 this is the case on the monthly mean time scale in the tropical Pacific and also (for some models) in the North Pacific. On the longer 5 years time scale 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 outperform the simple null hypotheses, suggesting they 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 scale and in the tropical Indo-Pacific the CMIP5 ensemble has improved on both time scales, but mostly by a lag of very ‘bad’ models from CMIP3 and not by an improvement of the ‘best’ models from CMIP5.
- The modes of the CMIP3 + 5 super ensemble have very realistic spatial structures, but underestimate the relative importance of each mode, due to the artificial spread between the individual models contributing to it. However, if the eigenvalues are scaled by the observed eigenvalues they are quite close to the observed modes and overall are closer to the observed spatial structure than any individual model. Thus the super ensemble of all the models gives the representation that is closest to the spatial structure of the observed SST variability, but only if the eigenvalues are scaled towards the observed.
- The global summary of the $RMSE_{EOF}$ for all domains in Fig. 9f can be considered the synthesis of the models skills in simulating natural SST variability. In this synthesis the CCSM4 turns out to be the best performing model. The CCSM4 model is performing relatively well in each individual domain.

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 5 years 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 than the slab ocean simulation. 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 scale.
- The models largely overestimate the effective number of spatial degrees of freedom (N_{spatial}) in all domains and in particular on the shorter time scale. Thus, the models produce more complex spatial structures 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 scale. Here, the mismatch between models is larger than relative to the simple $H_{0\text{STDV}}$ null hypothesis.
- Much of the disagreement between the models and the observations comes from smaller scale patterns. Often these have different locations, structures and amplitudes in the models relative to the observations or other models.

The somewhat limited skills in the models, particular in the extra-tropics, may require some discussion and some indications are given of what may be some of the problems. However, it needs to be noted here that this study cannot give the answer to these problems. Several aspects of this analysis indicate that the models limited skill is caused by ocean dynamics, coupling processes and possible error amplification: The first piece of evidence comes from the relative good performance of the slab ocean simulation, which does not simulate any variability caused by ocean dynamics, but performs more realistic in the simulation of the shorter time scale 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_{spatial} values in all extra-tropical domains. This may indicate that the ocean and air-sea interaction simulations of extra-tropical dynamics cause significant problems (e.g. Li et al. 2013; Kirtman et al. 2012). In particular, they 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 simulations: The models produce too many

small, localized modes of variability that are at the wrong positions with the wrong structures. Such modes do neither exist in the slab ocean simulation nor in the isotropic diffusion null hypotheses.

This kind of climate drift can be easily found in fully coupled GCMs as the atmospheric, oceanic and coupling processes could all introduce errors and even amplify the errors from each other (e.g. Delecluse et al. 1998; Grenier et al. 2000; Cai et al. 2011; Gupta et al. 2013). Further, 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 may potentially be one of the main problems in the CMIP CGCMs (e.g. Downes and Hogg 2013; Hirota and Takayabu 2013). However, it also needs 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 (e.g. Stockdale 1997; Dommenges 2012). 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 extra-tropical SST variability. CGCMs that simulate the positions of these fronts incorrectly will not be able to simulate the spatial structure of SST variability correctly (e.g. Huang et al. 2007; Brayshaw et al. 2008; Deremble et al. 2012). Here, the slab ocean simulation has a significant advantage, as it has the right mean SST climatology by construction due to the use of flux correction terms. Thus it seems reasonable to assume that the CGCM simulations will improve in the simulation of the spatial structure of SST variability if they would operate at a more realistic mean ocean state. This would in particular benefit seasonal to decadal prediction schemes in which the assimilation of observed ocean states is an important aspect of the overall skill of the predictions.

Acknowledgments We like to thank Tobias Bayr, Johanna Baehr, Katja Lorbacher and Timofej Woyzichowzki for fruitful discussions and comments. The comments of two anonymous referees have helped to improve the presentation of this study substantially. The ARC Centre of Excellence in Climate System Science (CE110001028) and the Deutsche Forschung Gemeinschaft (DO1038/5-1) supported this study. The slab ocean model simulations were computed on the National Computational Infrastructure in Canberra.

References

- Bayr T, Dommenges D (2014) Comparing the spatial structure of variability in two datasets against each other on the basis of EOF-modes. *Clim Dyn* 42(5–6):1631–1648

- Boer GJ, Lambert SJ (2001) Second order space–time climate difference statistics. *Clim Dyn* 17:213–218
- Brayshaw DJ, Hoskins B, Blackburn M (2008) The storm-track response to idealized SST perturbations in an aquaplanet GCM. *J Atmos Sci* 65:2842–2860
- Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Bladé I (1999) The effective number of spatial degrees of freedom of a time-varying field. *J Clim* 12:1990–2009
- Cahalan RF, Wharton LE, Wu W-L (1996) Empirical orthogonal functions of monthly precipitation and temperature over the United States and homogeneous stochastic models. *J Geophys Res* 101:26309–26318
- Cai W, Sullivan A, Cowan T, Ribbe J, Shi G (2011) Simulation of the Indian Ocean Dipole: a relevant criterion for selecting models for climate projections. *Geophys Res Lett* 38:L03704. doi:10.1029/2010GL046242
- Davies T, Cullen MJP, Malcolm AJ, Mawson MH, Staniforth A, White AA, Wood N (2005) A new dynamical core for the Met Office’s global and regional modelling of the atmosphere. *Q J R Meteorol Soc* 131:1759–1782
- Delecluse P, Davey MK, Kitamura Y, Philander SGH, Suarez M, Bengtsson L (1998) Coupled general circulation modeling of the tropical Pacific. *J Geophys Res* 103(C7):14357–14373
- Deremble B, Lapeyre G, Ghil M (2012) Atmospheric dynamics triggered by an oceanic SST front in a moist quasigeostrophic model. *J Atmos Sci* 69:1617–1632
- Dommenget D (2007) Evaluating EOF modes against a stochastic null hypothesis. *Clim Dyn* 28(5):517–531
- Dommenget D (2010) The slab ocean El Niño. *Geophys Res Lett* 37:L20701. doi:10.1029/2010GL044888
- Dommenget D (2011) An objective analysis of the observed spatial structure of the tropical Indian Ocean SST variability. *Clim Dyn* 36:2129–2145
- Dommenget D (2012) Analysis of the model climate sensitivity spread forced by mean sea surface temperature biases. *J Clim* 25:7147–7162
- Dommenget D, Latif M (2002) Analysis of observed and simulated SST spectra in the midlatitudes. *Clim Dyn* 19:277–288
- Downes SM, Hogg AM (2013) Southern Ocean circulation and eddy compensation in CMIP5 models. *J Clim* 26:7198–7220
- Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. *J Geophys Res* 113:D06104. doi:10.1029/2007JD008972
- Grenier H, Le Treut H, Fichefet T (2000) Ocean-atmosphere interactions and climate drift in a coupled general circulation model. *Clim Dyn* 16:701–717
- Guilyardi E (2006) El Niño–mean state–seasonal cycle interactions in a multi-model ensemble. *Clim Dyn* 26:329–348
- Gupta AS, Jourdain NC, Brown JN, Monselesan D (2013) Climate drift in the CMIP5 models. *J Clim* 26:8597–8615
- Hirota N, Takayabu YN (2013) Reproducibility of precipitation distribution over the tropical oceans in CMIP5 multi-climate models compared to CMIP3. *Clim Dyn* 41(11–12):2909–2920
- Huang B, Hu Z-Z, Jha B (2007) Evolution of model systematic errors in the tropical Atlantic basin from coupled climate hindcasts. *Clim Dyn* 28(7–8):661–682
- Jamson N, Kravtsov S (2010) Decadal variations of north Atlantic sea surface temperature in observations and CMIP3 simulations. *J Clim* 23:4619–4636
- Jolliffe I (2002) *Principal component analysis*, 2nd edn. Springer, New York
- Kao H, Yu J (2009) Contrasting Eastern-Pacific and Central-Pacific types of ENSO. *J Clim* 22(3):615–632
- Kirtman BP, Bitz C, Bryan F, Collins W, Dennis J, Hearn N, Kinter JL III, Loft R, Rousset C, Siqueira L, Stan C, Tomas R, Vertenstein M (2012) Impact of ocean model resolution on CCSM climate simulations. *Clim Dyn* 39(6):1303–1328
- Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple models. *J Clim* 23:2739–2758
- Krzanowski WJ (1979) Between-groups comparison of principal components. *J Am Stat Assoc* 74:703–707
- Li JLF, Waliser DE, Stephens G, Lee S, Ecuyer TL, Kato S, Loeb N, Ma Y (2013) Characterizing and understanding radiation budget biases in CMIP3/CMIP5 GCMs, contemporary GCM, and reanalysis. *J Geophys Res Atmos* 118:8166–8184
- Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (1997) A Pacific decadal climate oscillation with impacts on salmon. *Bull Am Meteorol Soc* 78:1069–1079
- Martin GM, Milton SF, Senior CA, Brooks ME, Ineson S, Reichler T, Kim J (2010) Analysis and reduction of systematic errors through a seamless approach to modelling weather and climate. *J Clim* 23:5933–5957
- Martin GM, Bellouin N, Collins WJ, Culverwell ID, Halloran P, Hardiman S, Hinton TJ, Jones CD, McLaren A, O’Connor F, Rodriguez J, Woodward S et al (2011) The HadGEM2 family of met office unified model climate configurations. *Geosci Model Dev Discuss* 4:723–757
- Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multi-model dataset: a new era in climate change research. *Bull Am Meteorol Soc* 88:1383–1394
- Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modeling uncertainties in a large ensemble of climate change simulations. *Nature* 430:768–772
- North GR, Bell TL, Cahalan RF, Moeng FJ (1982) Sampling errors in the estimation of empirical orthogonal functions. *Mon Weather Rev* 110:699–706
- Pierce DW, Barnett TP, Schneider N, Saravanan R, Dommenget D, Mojib L (2001) The role of ocean dynamics in producing decadal climate variability in the North Pacific. *Clim Dyn* 18(1–2):51–70
- Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J Geophys Res* 108(D14):4407. doi:10.1029/2002JD002670
- Reifen C, Toumi R (2009) Climate projections: past performance no guarantee of future skill? *Geophys Res Lett* 36:L13704. doi:10.1029/2009GL038082
- Santer BD, Taylor KE, Gleckler PJ, Bonfils C, Barnett TP, Pierce DW, Wigley TML, Mears C, Wentz FJ, Brüggemann W, Gillett NP, Klein SA, Solomon S, Stott PA, Wehner MF (2009) Incorporating model quality information in climate change detection and attribution studies. *Proc Natl Acad Sci USA* 106:14778–14783
- Smith TM, Reynolds RW, Peterson TC, Lawrimor J (2008) Improvements to NOAA’s historical merged land–ocean surface temperature analysis (1880–2006). *J Clim* 21:2283–2296
- Stockdale TN (1997) Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon Weather Rev* 125:809–818
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res* 106(D7):7183–7192. doi:10.1029/2000JD900719
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc* 93:485–498
- Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans R Soc Lond A* 365:2053–2075

- Washington WM, Meehl GA (1984) Seasonal cycle experiment on the climate sensitivity due to a doubling of CO₂ with an atmospheric general circulation model coupled to a simple mixed-layer ocean model. *J Geophys Res* 89(D6):9475–9503. doi:[10.1029/JD089iD06p09475](https://doi.org/10.1029/JD089iD06p09475)
- Xavier PK, Duvel J-P, Braconnot P, Doblus-Reyes FJ (2010) An evaluation metric for intraseasonal variability and its application to CMIP3 twentieth-century simulations. *J Clim* 23:3497–3508
- Zhou T, Wu B, Wang B (2009) How well do atmospheric general circulation models capture the leading modes of the interannual variability of the Asian–Australian monsoon? *J Clim* 22: 1159–1173