Framework for an Ocean-Connected Supermodel of the Earth System

François Counillon[1,2] ![ORCID]**, Noel Keenlyside[1,2]** ![ORCID]**, Shuo Wang[1], Marion Devilliers[3]** ![ORCID]**, Alok Gupta[4], Shunya Koseki[1]** ![ORCID]**, and Mao-Lin Shen[1]** ![ORCID]

[1]Bjerknes Centre for Climate Research, Geophysical Institute, University of Bergen, Bergen, Norway, [2]Nansen Environmental and Remote Sensing Center and Bjerknes Centre for Climate Research, Bergen, Norway, [3]Danish Meteorological Institute, Copenhagen, Denmark, [4]NORCE Norwegian Research Centre and Bjerknes Centre for Climate Research, Bergen, Norway

Abstract A supermodel connects different models interactively so that their systematic errors compensate and achieve a model with superior performance. It differs from the standard non-interactive multi-model ensembles (NI), which combines model outputs a-posteriori. Supermodels with Earth system models (ESMs) has not been developed because it is technically challenging to combine models with different state space. Here, we formulate the first supermodel framework for ESMs and use data assimilation to synchronise models. The ocean of three ESMs is synchronised every month by assimilating pseudo sea surface temperature (SST) observations generated by them on a common grid to handle discrepancies in grid and resolution. We compare the performance of two supermodel approaches to that of the NI. In the first (EW), the models are connected to the equal-weight multi-model mean, while in the second (SINGLE), they are connected to a single model. Both versions achieve synchronisation in the ocean and in the atmosphere, where the ocean drives the variability. The time variability of the supermodel multi-model mean SST is reduced compared to observations, most where synchronisation is not achieved and is lower-bounded by NI. The damping is larger in EW, for which variability in the individual models is also damped. Hence, under partial synchronisation, the unsynchronized variability gets damped in the multi-model average pseudo-observations, causing a deflation during the assimilation. The SST bias in individual models of EW is reduced compared to that of NI, and so is its multi-model mean in the synchronised regions. A trained supermodel remains to be tested.

Plain Language Summary Supermodeling is a novel approach in which different models are run interactively and are connected to achieve a model with superior performance. The method exploits model diversity and reduces model error by training the connection terms using observations. Structural differences between the models and the amount of data exchange among models are technical challenges and have limited the applicability of supermodeling to Earth system models (ESMs). Here, we show that data assimilation can handle these issues. A demonstration is done with three ESMs having different resolution and structural differences and using a limited amount of data exchange (monthly ocean surface temperature). Synchronisation is achieved in several key regions. There, the ocean surface temperature error is smaller than by taking the a-posteriori average of the non-interactive multi-model ensemble, an approach commonly used for handling Coupled Model Intercomparison Project simulations. It also modifies atmospheric teleconnections. Connecting the model via their weighted mean (one of the two approaches tested) causes a spurious deflation of variability. This study opens the application and training of supermodeling to ESMs.

1. Introduction

Climate models have been key tools for answering fundamental questions about our climate systems. However, large uncertainties exist with many key processes being parameterized and biases in several of the Earth system components (e.g., ocean, atmosphere, sea ice, and land) being larger than the projected changes in climate and larger than the variability being predicted (Palmer & Stevens, 2019). While models have improved through the successive generations of coordinated model intercomparison projects (CMIP)—version 6 being the latest (Eyring et al., 2016)—many of the large biases persist, for example,: the double-Intertropical Convergence Zone (ITCZ) problem (Tian & Dong, 2020); the tropical Atlantic bias (Richter et al., 2014) and the signal-to-noise paradox (Scaife & Smith, 2018). Although some simulations are being tested at a breakthrough resolution (Zhongming et al., 2020), we are still most likely decades away from being able to perform operational climate simulations

Writing – review & editing: Noel Keenlyside, Marion Devilliers

with models that can explicitly resolve the most important physical processes. In the meantime, one can test alternate methods to understand and mitigate these biases using the current generation of models.

The classical approach to mitigate model error is to take the multi-model average of independent model simulations (a-posteriori) so that errors in the different models cancel. This approach is standard for climate projection but can also improve predictions (Branicki & Majda, 2015). One can refine the post-processing by taking a weighted mean of the different runs—an approach referred to as a super ensemble (Krishnamurti et al., 2016). However, this approach has limitations as most models share the same deficiencies (e.g., double ITCZ, warm bias in the tropical Atlantic) and because linear post-processing does not necessarily correct non-linear responses, such as climate sensitivity.

Supermodeling builds on the interactive ensemble approach (Kirtman & Shukla, 2002) where multiple realizations of the same atmospheric general circulation model are simultaneously coupled to a single ocean general circulation model through averaging their air-sea fluxes. Supermodels couple simultaneously different models and take advantage of their diversity to compensate their errors (Duane et al., 2018). Models are connected as they run via their state variables or their tendencies. Models can either be connected to each other (e.g., Mirchev et al., 2012; Smith, 2001) or toward their weighted mean (Schevenhoven & Carrassi, 2021; Wiegerinck et al., 2013). During a training phase, the connection terms are optimized to formulate a new synchronised dynamical system that achieves enhanced performance. Supermodels rely on two important properties: first, it is possible to synchronise different models exchanging a few variables—an approach referred to as chaos synchronisation of non-linear dynamical systems (Duane & Tribbia, 2001; Pecora et al., 1997)—and second, model diversity can encompass the true behavior of the dynamical system.

Supermodels have been tested with various models and experimental designs of different complexity. At first, idealized framework experiments (or observing system simulation experiments, Halem & Dlouhy, 1984) were used. Theses are convenient because they avoid challenges associated with the real framework and because the truth (constructed from a model) is known. One can introduce model error (e.g., by perturbing parameter values or using a different model) and disclose only part of the true model state as observations (perfect or not). Supermodeling was successfully demonstrated for parametric model error forming a convex envelope around the truth and that with low dimensional dynamical systems (e.g., Lorenz 63, Lorentz 96, Rossler systems; see Mirchev et al., 2012; Van den Berge et al., 2011; Du & Smith, 2017), quasi-geostrophic atmospheric models (Schevenhoven & Selten, 2017; Wiegerinck & Selten, 2017), and the global atmosphere–ocean–land model of intermediate complexity SPEEDO (Schevenhoven & Selten, 2017; Selten et al., 2017). However, when the model error does not cancel out (i.e., parameters do not form a convex envelop around the truth), one can use negative weights (Schevenhoven & Carrassi, 2021; Schevenhoven et al., 2019), which raises new challenges. Furthermore, the supermodel can degrade performance on time scales different to those it was trained on (Wiegerinck & Selten, 2017). Nevertheless, the first demonstration of supermodeling with real data successfully mitigated the double ITCZ bias and improved the representation of the dynamics in the equatorial Pacific (Shen et al., 2016, 2017). These results were achieved with two versions of the ECHAM5 Atmospheric General Circulation Models (AGCM)—each of them using a different convection scheme—providing the weighted average fluxes to a single Oceanic General Circulation Model (OGCM), Max Planck Institute Ocean Model (MPIOM) (Shen et al., 2016). One may expect to enhance the performance of supermodeling by broadening model diversity that expands the convex envelop so that it may enclose the true Earth system—to the extent possible, recognizing that the real climate system is far more complicated than any numerical model.

Building a supermodel with fundamentally different ESMs raises new challenges. Models do not share the same model-state space, nor do they have comparable resolution and representativity (Janjić et al., 2018). For this purpose, Du and Smith (2017) suggested formulating pseudo-observations of the different models and assimilating them back into the different models. The approach was demonstrated with low-dimensional systems and in an idealized framework. Another challenge is practical limitations with feasible data volume exchanged among models and frequency of the synchronisation steps. Here, we aim to assess whether a supermodel can achieve synchronisation (a requirement of supermodeling) in a configuration where data exchange is sparse, synchronisation steps infrequent, and models have a radically different structure. We will apply part of the formalism proposed by Du and Smith (2017)—connecting models via assimilation of synthetic pseudo-observations generated from the multi-model ensemble. We will examine the impact of generating the pseudo-observations from a single model or constructing it from an equal-weighted mean. In future work, we will implement training and analyze the resulting supermodel's performance.

This paper is organized as follows. Section 2 presents the practical implementation of the supermodeling framework: the description of the individual ESMs, the synchronisation methodology and the data assimilation (DA)
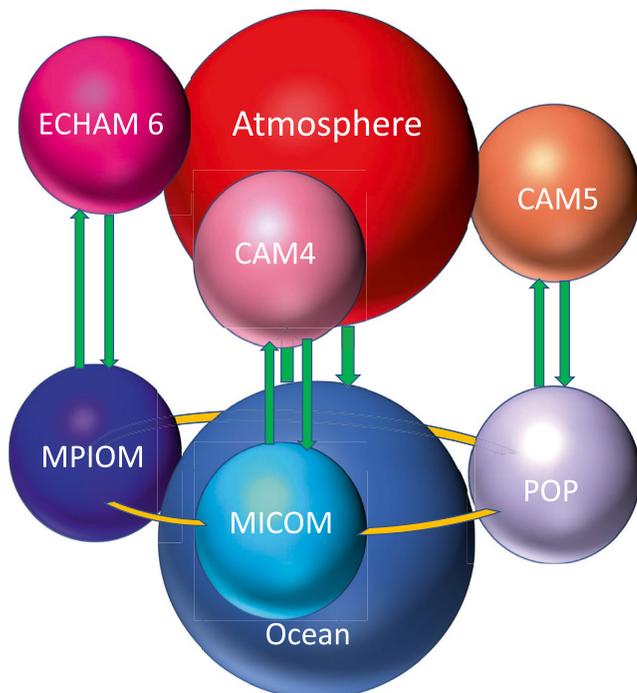
Figure 1. Schematics of the supermodel framework. The green arrows denote dynamical coupling in the individual models, and the yellow band denotes synchronisation in the different ocean models achieved using data assimilation.

method. Section 3 introduces the validation data sets and metrics, and Section 4 presents the result of two supermodeling frameworks, first globally with a focus on the damping of internal variability (ocean and atmosphere) and impact on the SST bias and second with a focus on the ENSO region and associated teleconnections.

2. Supermodeling Framework

This section describes the practical implementation of the supermodel framework for ESMs using ocean connections. It combines the Norwegian Earth System Model (NorESM), the Community Earth System Model (CESM) and the Max Planck Institute Earth System Model (MPI-ESM) (Figure 1). The three ESMs are connected via their SST every month. We implement an assimilation system into each model individually based on the Ensemble Optimal Interpolation. The DA method updates the whole water column based on the synthetic SST pseudo-observations constructed from the multi-model ensemble mean or from a single model.

2.1. CESM

The CESM is a global, fully coupled model for climate simulations. We used the CESM Large Ensemble Project (LENS) version based on CESM1.1.2 (Kay et al., 2015), with all components at approximately 1° horizontal resolution. External forcing complies with CMIP5's historical experiment. The atmospheric component is the Community Atmosphere Model version 5, (CAM5, Hurrell et al., 2013) with 30 vertical levels and a finite-volume grid (f09, i.e., approximately 1°). The Community Land Model, version 4 (CLM4), is on the same grid as the atmosphere. The Parallel Ocean Program, version 2 (POP), is run with 60 vertical levels. The horizontal resolution of the ocean is approximately 1°, but it is enhanced in the meridional direction around the equator and both in zonal and meridional directions at high latitudes (g16 grid). The sea ice [Los Alamos Sea Ice Model (CICE) version 4] component model is on the same grid as the ocean model. We use historical forcing from 1920 to 2005 (Lamarque et al., 2010) and retrieve the initial conditions from the National Center for Atmospheric Research repository (b.e11.B20TRC5CNBDRD.f09_g16.001) in 1950. As POP2 only allows modification for one time level of the leapfrog scheme, we follow the approach of the data assimilation research testbed (DART, Anderson et al., 2009). It uses a forward Euler scheme for the first time step but reverts to the leapfrog scheme afterward. The barotropic velocities and surface pressure gradients are adjusted to preserve the ocean volume. The flag POPDART can activate this option in POP2.

2.2. NorESM

We use the medium-resolution NorESM1-ME (Tjiputra et al., 2013) that contributed to the CMIP5. It is based on the CESM version 1.0.3 (CESM1, Vertenstein et al., 2012). However, the atmospheric component (CAM4-OSLO) features an advanced aerosol chemistry scheme (Kirkevåg et al., 2013), and the ocean is an updated version of the isopycnal coordinates ocean model MICOM (Bentsen et al., 2012). The atmosphere and land components are configured on a finite-volume grid with a latitude and longitude resolution of $1.9 \times 2.5°$. The atmosphere component uses 26 hybrid sigma-pressure levels with a model top at approximately three hPa. The horizontal resolution of the ocean and sea-ice model is approximately 1°. The ocean uses 51 isopycnal layers and two layers representing the bulk mixed layer with time-evolving thicknesses and densities. The initial condition is taken from a historical simulation in 1980.

2.3. MPI-ESM

We use the MPI-ESM1-LR (Block & Mauritsen, 2013; Dunstone et al., 2018; Giorgetta et al., 2013) model that contributed to the CMIP5. The AGCM of MPI-ESM1-LR is the sixth generation European Centre Hamburg

Table 1
We Summarize the Key Characteristics of the Different Earth System Models Used

Model version	Ocean	Atmosphere	Reference
NorESM1-ME	MICOM(σ; 1°)	CAM4 (finite-volume, 2°)	Tjiputra et al. (2013)
CESM1.1.2	POP2 (z, 1°)	CAM5 (finite-volume, 1°)	Kay et al. (2015)
MPI-ESM1-LR	MPIOM (z,1.5°)	ECHAM6 (spectral, 2°)	Block and Mauritsen (2013)

Note. The first column details the model versions. The second column reports the name of the ocean models, their resolution and coordinate system in the vertical—that is, isopycnal (σ-coordinate) or geopotential depth (z-coordinate). The third column is the name of the atmospheric models and their discretization scheme. The last column provides a reference to the model version.

general circulation model (ECHAM6, Stevens et al., 2013), and the OGCM is the MPIOM (Jungclaus et al., 2013; Marsland et al., 2003). The land model (JSBACH, Reick et al., 2013; Schneck et al., 2013), which includes vegetation, and the marine biogeochemistry model (HAMOCC5, Ilyina et al., 2013) are considered as subsystems of ECHAM6 and MPIOM, respectively. ECHAM6 employs T63 spectral resolution (approximately 1.9° horizontal resolution) and 47 vertical levels, and MPIOM employs a rotated curvilinear grid with an approximate 1.5° horizontal resolution and 40 z-levels. The poles of MPIOM are moved to Greenland and the Weddell Sea. We start the simulation from a historical simulation in 1980.

2.4. Synchronisation Methodology

The three ESMs use different models, grids, coordinates, and resolutions (Table 1). NorESM and CESM may be more similar, but they use a fundamentally different ocean model (in geopotential depth coordinates for CESM and isopycnal coordinates for NorESM), and the atmospheric model in CESM is a more advanced version (CAM5 vs. CAM4) and has a higher resolution (1° vs. 2°). Consequently, one cannot simply use state replacement or nudging, as they will generate imbalances and may lead the models to crash. Data assimilation (DA) can estimate the best possible (most likely) state based on observations, a dynamical model, and their uncertainties. It is designed to preserve the dynamical consistency of the individual models and optimally handle observation uncertainty (Carrassi et al., 2018).

A limitation of DA methods for this application is that they are (with few exceptions, e.g., Nerger et al., 2020; S. Zhang et al., 2007) working offline—meaning that the model is stopped, the state written on disk, DA applied on the files and the model restarted. With such large systems as the ESMs, the time required for initializing the model and writing the input/output is burdensome (see, e.g., Karspeck et al., 2018), limiting the feasible frequency of the synchronisation. Similarly, one needs to limit the number of variables that will be synchronised to keep the cost of the DA-step low.

As a first attempt before developing a more advanced connected supermodel of ESMs, we try to synchronise the three models through their SST at a monthly frequency. SST is sufficient to constrain the variability in many regions of the earth system, particularly in the tropics (Shukla, 1998; Wang et al., 2019; Zhu et al., 2017). It is observed over a long period with a good level of accuracy, enabling the possibility to effectively train our supermodel and validate it for an independent period. With monthly synchronisation, the additional computational cost of the monthly assimilation remains small.

We test two supermodeling approaches that differ in their formulation of the pseudo-observations. The first scheme belongs to the category of state-constrained weighted supermodels (Schevenhoven & Carrassi, 2021). The three models are integrated forward for 1 month, and their SSTs are interpolated to a common 1° grid. The pseudo-observations are the equal-weighted mean of these outputs (referred to as EW in the following). Weights should be trained using observations to optimize model performance, but this optimisation is out of the scope of the paper. The pseudo-observations are then assimilated into the individual models using the EnOI method (Section 2.5), and the models are then restarted for the next cycle. The first synchronisation step in the two supermodel frameworks started on the first of February 1980 for practical reasons. We do not synchronise models under the union of all three models sea ice mask.

In the second scheme, the workflow is similar, but the pseudo-observations are formulated from a single model (hereafter referred to as SINGLE). This approach is, for instance, used in the cross-pollination in time method

(Du & Smith, 2017; Smith, 2001). Another objective of this experiment is that it should not be affected by variability damping—for example, experienced with interactive ensemble (Kirtman & Shukla, 2002; Kirtman et al., 2004; W. Zhang & Kirtman, 2019)—and can serve as a benchmark for the EW. It can serve to assess the potential synchronisation that can be achieved with SST and monthly synchronisation steps. We have selected arguably CESM that has higher resolution and provides overall the best performance. We interpolate the CESM pseudo-observations onto the common grid to have a comparable interpolation error between the two schemes. Formulating the pseudo-observations in the native CESM grid enhances the performance slightly but does not change any conclusion unless reported.

The performances of the two supermodel approaches are compared to a non-interactive multi-model ensemble (hereafter referred to as NI) which start from the same initial condition than the two supermodels, but do not stop for the synchronisation steps. Thus, NI provides a benchmark for assessing the degree of synchronisation in the supermodels. All models start from a historical simulation in 1980 using CMIP5 historical forcing, and RCP8.5 is used for 2006 (Taylor et al., 2012) when our simulations stopped.

The system runs on 11 nodes (1408 CPU) on the Norwegian high performance computer Betzy (a BullSequana XH2000) and can achieve about 10 model-year per day. CESM runs on eight nodes and runs 1 month in approximately 7 mins, NorESM runs on two nodes and performs the 1-month simulation in approximately 7 mins, MPI-ESM uses one node and performs a 1-month simulation in approximately four and half minutes. The model integration accounts for approximately 7 mins, while the DA step accounts for approximately 2 mins. The assimilation is currently performed sequentially (i.e., one model after the other), and this step could have been reduced to 40 s if parallelized.

2.5. Ensemble Optimal Interpolation

The Ensemble Optimal Interpolation (EnOI, Evensen, 2003; Oke et al., 2002) is a computationally cheap sequential DA method derived from the Ensemble Kalman Filter (EnKF, Evensen, 2003).

The EnOI provides multivariate updates based on the model's historical covariance. Formulating the covariance from the same model ensures the preservation of linear quantities (such as geostrophic balance) and limits initialization shock (Counillon & Bertino, 2009). The covariance matrix is constructed based on an ensemble of N model snapshots (each of dimension n, where n is the state dimension):

$$\mathbf{X}^s = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathcal{R}^{n \times N}. \tag{1}$$

The static ensemble anomaly $\mathbf{A}^s$ is calculated so that $\mathbf{A}^s = \mathbf{X}^s - \overline{\mathbf{X}^s}\mathbb{1}^{\mathbf{T}}$, with $\overline{\mathbf{X}^s}$ being the static ensemble mean and $\mathbb{1}_m = [1, 1, \ldots, 1] \in \mathcal{R}^{1 \times N}$.

To correct a forecast $\mathbf{x}^f$, using the observation vector $\mathbf{d}$, one can estimate a new analysis state $\mathbf{x}^a$ as follows:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^f). \tag{2}$$

where $\mathbf{H}$ is the observation operator that relates the prognostic model state variables to the measurements. The Kalman Gain $\mathbf{K}$, is computed as follows:

$$\mathbf{K} = \mathbf{A}^s \mathbf{A}^{s\mathrm{T}} \mathbf{H}^{\mathrm{T}} (\mathbf{H}\mathbf{A}^s \mathbf{A}^{s\mathrm{T}} \mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{3}$$

For each model, the static ensemble is composed of the monthly snapshot outputs from a stable pre-industrial control run. Sampling the model states from a pre-industrial condition was preferred over sampling it from a historical run because composing the static ensemble from a model with transient forcing can introduce spurious correlation. The monthly static ensemble comprises 72 members for MPI-ESM, 80 for NorESM, and 80 for CESM. We use a local framework analysis (Evensen, 2003) with a radius of 235 km without tapering—this radius ensures at least one observation per grid cell. We limit the local observations to one (retaining only the nearest). We update all prognostic state variables in the vertical. The update is done in the model's native coordinates—that is, in geopotential depth for MPI-ESM and CESM and isopycnal coordinates for NorESM. We use the k-factor formulation (Sakov et al., 2012) which artificially inflates the observation error if the assimilation pushes the update beyond twice the ensemble spread. We set the pseudo-observation error for both supermodel approaches
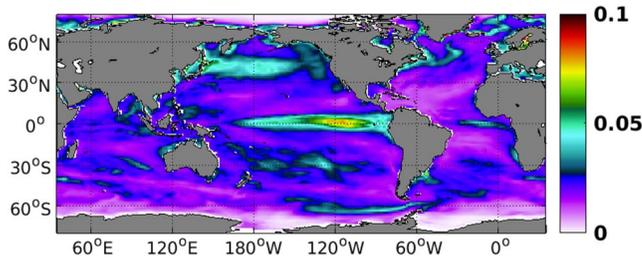
Figure 2. Pseudo-observation error standard deviation (in °C) used for assimilation.

as the pointwise de-seasoned (with the mean seasonal cycle removed) time standard deviation of the three models divided by 30 (see Figure 2). We have played with the scaling factor (from 3 to 100), but the results were not very sensitive to this choice (not shown) due to the k-factor formulation.

3. Validation Data Sets and Metrics

For validating the simulations, we use SST from the NOAA OISST V2 (Reynolds et al., 2002) analysis data set. We use the monthly averaged product available on a 1° grid, which extends back to 1982. We assess performance by comparing the climatological difference between the models and the observation and calculating grid cell area-weighted root mean square error (RMSE). For estimating correlation between SST and precipitation, we use in addition the Global Precipitation Climatology Project monthly precipitation (GPCP) Version 2.3 data set (Adler et al., 2003).

Internal variability can be suppressed when combining models, as for example, through the averaging of multi-instance fluxes in the interactive ensemble framework (Kirtman & Shukla, 2002; Kirtman et al., 2004; W. Zhang & Kirtman, 2019). Hence, we introduce two metrics to investigate how internal variability is affected by the connection. The parameter δ relates to how synchronisation is achieved by the model and it is the ratio between the time standard deviation of the multi-model mean and the time average of the inter-model standard deviation. If we denote a model quantity (e.g., SST) at a given grid cell and calendar month $\mathbf{x}_i^j$ with the superscript j referring to the model indices ($j \in [1, N_s]$, with $N_s = 3$) and the subscript i referring to the year ($i \in [1, N_y]$; with $N_y = 26$). We can decompose the model variable into:

$$\mathbf{x}_i^j = \mathbf{x}_i^s + \mathbf{a}_i^j, \tag{4}$$

where $\mathbf{x}_i^s$ is the multimodel mean at a given year i and $\mathbf{a}_i^j$ is the anomaly from the mean. The time variance of the model j is:

$$\sigma_j^2 = \sigma_s^2 + \sigma_{a,j}^2, \tag{5}$$

with σ_s^2 is the time variance of the multimodel mean and $\sigma_{a,j}^2$ is the variance of the anomaly from the mean.

$$\delta^2 = \frac{\sigma_s^2}{\frac{1}{N_s} \sum_{j=1}^{N_s} \sigma_{a,j}^2} \tag{6}$$

δ is computed for every grid cell and every calendar month. If all models are independent and identically distributed (iid) variables (with, $\sigma_j^2 = \sigma^2, \forall j$), we know from the Bienaymé identity that $\sigma_s^2 = \sigma^2/N_s$. Using this relation and Equation 5, we can show that δ^2 for iid is:

$$\delta_{iid}^2 = \frac{\sigma_s^2}{\frac{1}{N_s} \sum_{j=1}^{N_s} \left(\sigma^2 - \sigma_s^2 \right)} = \frac{\sigma_s^2}{N_s \sigma_s^2 - \sigma_s^2} = \frac{1}{N_s - 1}, \tag{7}$$

giving $\delta_{iid} = 0.7$ with three models. If δ is larger than 0.7, some synchronisation is achieved. For example, a value of three indicates that the time standard deviation of the multi-model mean is three times larger than the inter-model standard deviation. Values lower than 0.7 can occur when the bias of the individual models is larger than the time variability of the multi-model mean (i.e., models are strongly attracted to their bias and have little variability). In the manuscript we analyze the yearly mean value of the monthly δ estimate. This is preferable to computing the δ from de-seasoned time series, which can overestimate synchronisation for quantity with a strong seasonal variability—for example, precipitation in the tropic where it rains only during few calendar months.

A second metric, called λ, quantifies the damping of the multi-model mean; that is, the ratio between the multi-model mean time variability (denominator) and the mean of the variability in the unconnected individual models (numerator; Equation 8). In a perfect synchronisation regime, there is no damping (Duane & Tribbia, 2001). Assuming that the time variability of each individual models is equal, an equal-weight multi-model mean would have comparable variability to the mean variability of the individual models (λ should be one). If only partial
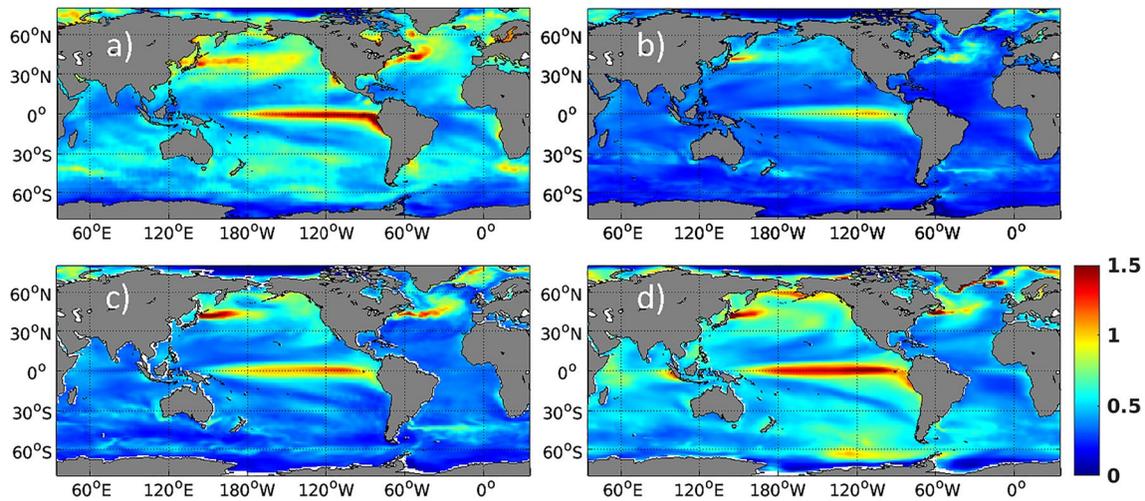
Figure 3. De-seasoned time standard deviation of sea surface temperature in the NOAA OISST2 observations (a), the multi-model average of NI (b), EW (c) and SINGLE (d).

synchronisation is achieved, the value gets larger than one. For example, λ equal to 2 means that the standard deviation of the multi-model mean is half that of the unconnected model. For the iid case the damping will be $\sqrt{N_s}$, meaning about 1.8 with 3 models. As for δ we present the yearly average of the monthly estimates.

$$\lambda^2 = \frac{\frac{1}{N_s} \sum_{i=1}^{N_s} \sigma_j^2}{\sigma_s^2} \tag{8}$$

4. Supermodel Results

We compare the performance of EW, SINGLE and NI (see Section 2.4) over the period 1982–2006. In Figure 3, we show the de-seasoned time standard deviation of the multi-model mean of NI, EW and SINGLE and the OISSTV2 observations. The time variability in the NI multi-model mean provides a benchmark for synchronisation, as it is a result of averaging largely independent variability in three simulations. As a result, the NI standard deviation of the multi-model mean is $1/\sqrt{3}$ that of the standard deviation of the individual models, and it is substantially lower than in the observations. In EW, it is also reduced—albeit less than in the NI—, while SINGLE has nearly comparable amplitude to the observations.

In Figure 4, we are analyzing the properties of each system in achieving synchronisation and causing damping of multi-model mean time variability. We present individual 2-dimensional maps of the yearly mean of the monthly δ and λ (metrics introduced in Section 3) and the 2-dimensional probability density function (PDF) of δ versus λ. An ideal supermodel will have the highest probability centered on the dashed brown line ($\lambda = 1$) and on the right-hand side of the red line (i.e., $\delta > 0.7$).

We can see that for NI (Figures 4a, 4d, and 4g), δ and λ values are centered near the expected value of an iid case. There is barely any synchronisation, and models are only connected via their historical CMIP5 forcing, while the interannual variability overpowers the climate change signal for the period 1980–2005. Note that such damping can be avoided by averaging the variance of the individual NI models.

In the EW supermodel (Figures 4b, 4e, and 4h), there is some synchronisation ($\delta > 0.7$) and the damping λ is mostly around 1.5. The synchronisation and the damping are improved compared to NI, and the maximum likelihood moves toward the optimal. Synchronisation is maximum in the equatorial Pacific, reaching a value of 3.5 and is above 1 in the tropical Pacific, part of the North Pacific and part of the North Atlantic (entrance of the Nordic Seas). In those regions, λ lowers toward 1 (albeit remaining above). However, δ remains below or close to 0.7 in several regions: for example, in the equatorial Atlantic and the eastern boundary upwelling system and most of the Southern Oceans. No synchronisation is achieved there, and internal variability of the multi-model mean is damped and of comparable amplitude to NI mean. The shape of δ in the equatorial Pacific shows maxima on
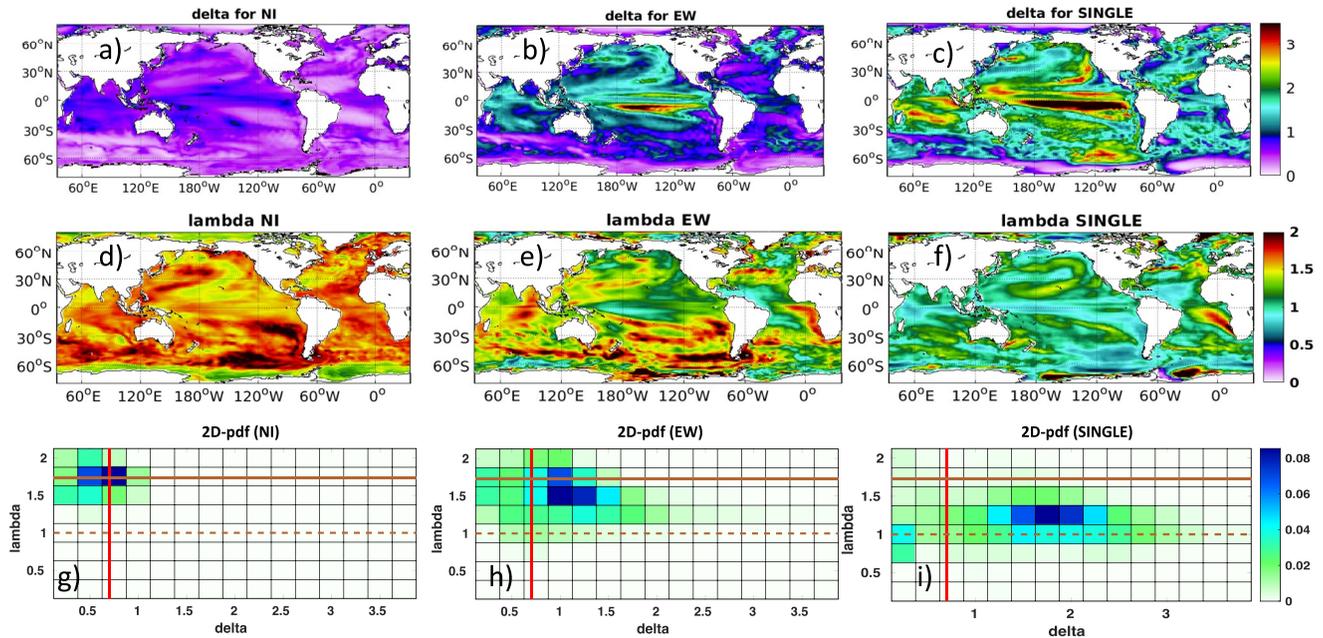
Figure 4. The first row is the synchronisation metric δ with NI mean (a), EW (b) and SINGLE (c). The second row (d–f) is the same for the synchronisation metric λ, and the last row (g–i) is the 2-dimensional probability density function of δ versus λ in the different systems. The red and brown solid lines highlight the expected values for an iid case ($\lambda = \sqrt{N_s}$; $\delta = \sqrt{\frac{1}{N_s - 1}}$). The dashed brown line highlights the perfect synchronisation case ($\lambda = 1$; $\delta = \infty$).

either side of the equator but achieves only moderate synchronisation at the equator. Hence, the fast Kelvin waves are driven by wind bursts in the equatorial Pacific, while atmospheric variability is poorly constrained in our ocean-constrained system. Slower Rossby waves control the thermocline variability off the equator, and the effect of winds is weaker and slower. Ocean DA with monthly synchronisation steps can control such a process better.

The pattern of the value of δ with SINGLE (Figures 4c, 4f, and 4i) resembles that of EW, but the values are considerably larger. Synchronisation δ is mostly greater than 1.5, and the damping λ is reduced (approaching one and rarely exceeding 1.5). Actually, the CESM model (to which all models are connected) has a lower time standard deviation than MPI-ESM and NorESM, while the denominator in λ is the average of the time standard deviation of the three models. It partly explains the slight damping remaining ($\lambda > 1$). Synchronisation is now achieved in most parts of the Southern Ocean, the Atlantic Ocean, and the Indian Ocean. In the equatorial Atlantic and the eastern boundary upwelling system, synchronisation is still not achieved, and the damping is substantial. This region is notorious for being very challenging for models (Richter et al., 2014) and a considerable fraction of the bias relates to atmospheric origin (e.g., wind bias, Koseki et al., 2018).

The analysis above is carried for SST that is directly constrained by assimilation, and we now analyze synchronisation for other quantities: sea surface height (SSH), sea level pressure, zonal surface stress and precipitation (Figure 5). As for SST, no synchronisation is found for NI (Figures 5a, 5d, 5g, and 5j).

The SSH variations relates to the thickness of the upper layer and thermocline-depth variations, which are mostly wind driven and not strongly forced by the common external forcing. Some synchronisation is found for both EW and SINGLE (Figures 5b and 5c), which can be related the multivariate properties of our assimilation scheme—the SST observation corrects the full water column and as such the steric sea level. As a matter of fact, the region where synchronisation is achieved coincides very well with region where assimilation of SST reduces error in SSH (see e.g., Figure 14 in, Wang et al., 2019).

Atmospheric quantities are not directly corrected by assimilation but can be synchronised through forcing by the underlying SST. This dynamical adjustment occurs in between the monthly synchronisation cycle. Sea level pressure (SLP), zonal surface stress and precipitation show some synchronisation in EW (Figures 5e, 5h, and 5k) and SINGLE (Figures 5f, 5i, and 5l), as a response to the constrained ocean variability. The synchronisation for these three variables is greatest over the Pacific and Indian Ocean, but extends to the tropical
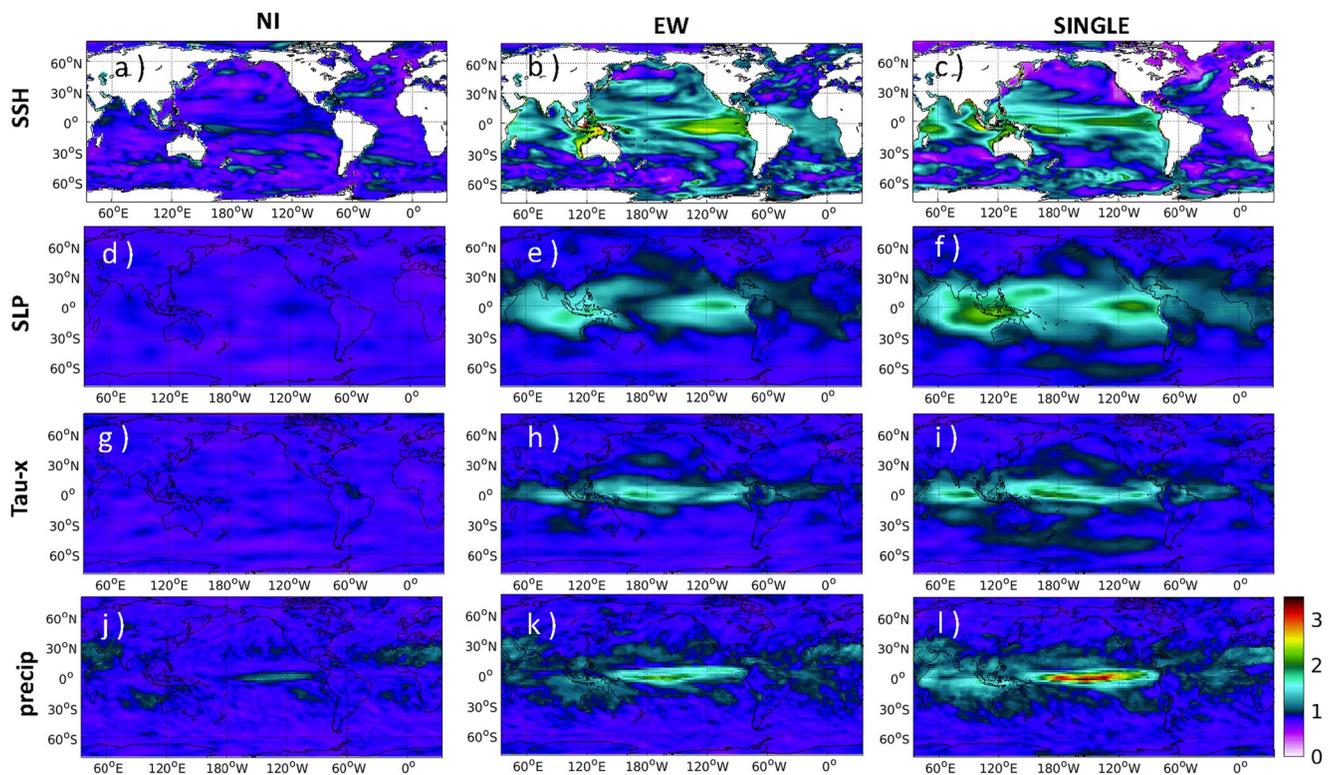
Figure 5. The first row is the synchronisation metric δ with NI mean (a), EW (b) and SINGLE (c) for sea surface height, the second row (d–f) is for sea level pressure, the following row (g–i) for zonal surface stress and last row (j–l) for precipitation.

Atlantic and to the higher latitudes. This is consistent with stronger ocean-atmosphere interaction in the tropics, with SST influencing SLP and precipitation directly, which drives changes in surface winds (Back & Bretherton, 2009; Gill, 1980; Lindzen & Nigam, 1987). Synchronisation across the tropics is also facilitated by interbasin interactions (Cai et al., 2019; Keenlyside et al., 2020), while teleconnections from the tropics can explain some weak synchronisation in the extra-tropics (Bjerknes, 1969; Hoskins & Karoly, 1981; Wallace & Gutzler, 1981); where underlying SST are less synchronised and have relatively weak impact on the atmosphere that is not well simulated with relatively coarse horizontal resolution models (Kirtman et al., 2012; Kushnir et al., 2002; Smirnov et al., 2015).

The above analysis focused on the time standard deviation of the multi-model mean. We now analyze the variability in the individual models. In Figure 6, we show the quantile-quantile plot of de-seasoned SST of the CESM from the NI simulation (i.e., the unconnected CESM historical run) versus that of CESM in the EW and SINGLE simulations. Focusing on CESM allows us to discard the discrepancies of variability between models—CESM has weaker temporal variability than the two other models (i.e., $\sigma^2_{CESM} < \sigma^2_{NorESM}$ and σ^2_{MPIESM}). In EW, the time variability of CESM is reduced. The other models (MPI-ESM and NorESM) of EW show comparable damping (not shown). Contrarily, SINGLE does not cause deflation, and the regression line yields a slight overestimation. It may relate to the fact that assimilating pseudo-observations in a common grid adds energy to the system (because of the imbalance). A perfect fit is obtained if pseudo-observations are formulated in the CESM native grid (not shown).

In Figure 7, we analyze the SST bias of the individual models compared to OISSTV2 observations. In the NI system (Figures 7a–7c), CESM has a lower bias than MPI-ESM and NorESM. NorESM has a pronounced cold bias. In the EW supermodel (Figures 7d–7f), the bias of CESM and NorESM is reduced compared to that of NI, but remains identical in MPI-ESM. Achieving such a result is highly promising, considering that weights have not been trained. In SINGLE (Figures 7g–7i), the bias structure in all models resembles that of CESM in NI. It yields a bias reduction for MPI-ESM and NorESM but increases the bias in CESM compared to NI. The interpolation to the common grid causes the degradation (not shown).
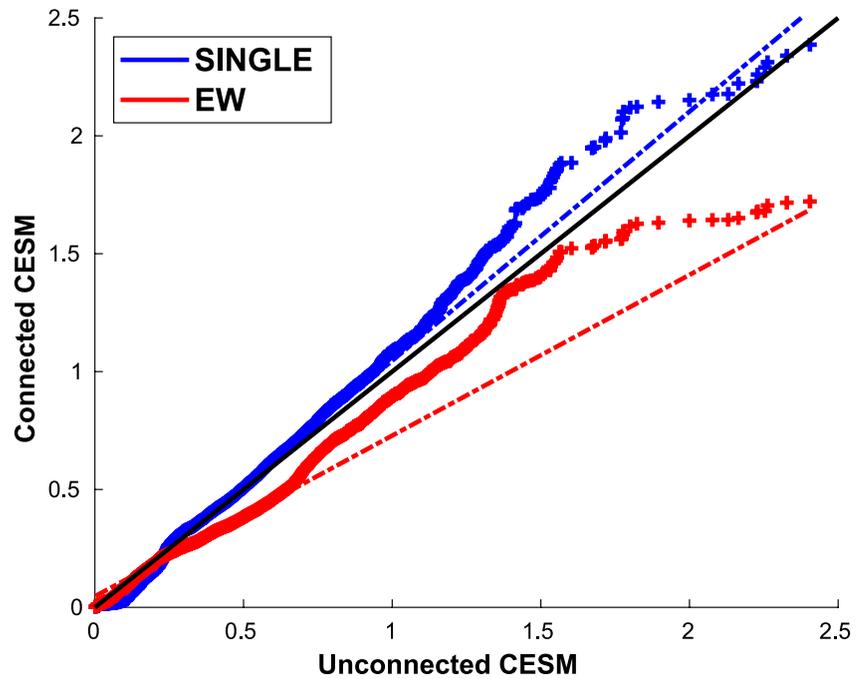
Figure 6. Quantile-quantile plot of the pointwise de-seasoned time standard deviation of a historical unconnected Community Earth System Model (CESM) simulation (CESM in NI multi-model) versus CESM in the connected multi-model EW (red) and SINGLE (blue). The dashed color line show the regression line and the solid black line shows the perfect regression line.

The bias of the multi-model mean of EW and NI (see Figures 8a and 8b) are comparable in pattern and amplitude (NI has a slightly lower area-weighted RMSE). Still, some discrepancies exist. Figure 9 shows the absolute difference between the two. The bias of EW reduces in the tropics, except in the equatorial Atlantic and the eastern boundary upwelling system. Improvement tend to coincides with regions where synchronisations is achieved (i.e., $\delta > 1$ and $\lambda < 1.5$; Figure 4b). In the equatorial Pacific (5°S–5°N and 150°W–90°W), the error reduces
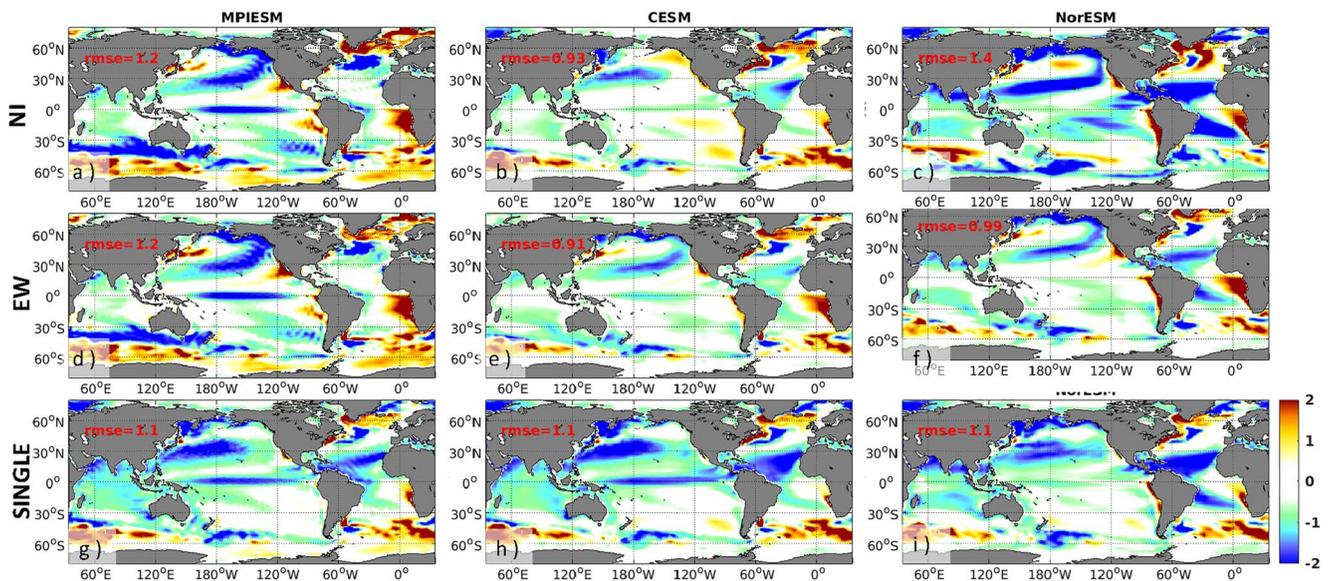
Figure 7. The first row is the sea surface temperature climatological bias computed over 1982–2006 with the NI ensemble for Max Planck Institute Earth System Model (a), Community Earth System Model (b) and Norwegian Earth System Model (c); the second row with the EW (d–f) and the last row with SINGLE (g–i). The quantity in red reports the global spatial root mean square error normalized by grid cell area.
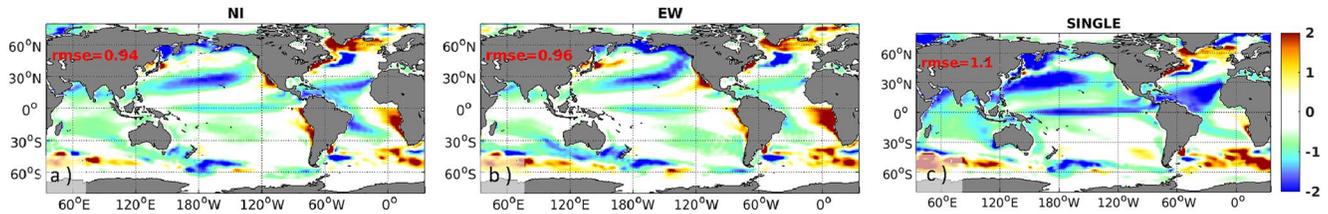
Figure 8. Climatological sea surface temperature bias computed over 1982–2006 for the multi-model mean of NI ensemble (a), EW (b) and SINGLE (c). The quantity in red reports the global spatial root mean square error normalized by grid cell area.

from 0.91 to 0.67°. A bias reduction compared to the a-posteriori average results from the non-linear response of the climate system to model bias. Contrarily, performance degrades where synchronisation is poor (Antarctic Circumpolar current, Equatorial Atlantic and eastern boundary upwelling system). There, ocean synchronisation is quickly lost, and noise adds energy to the system. In the north Pacific, synchronisation is achieved, but the bias is degraded. The reason is unclear, but we propose several hypotheses. First, variability there during winter is driven by storm systems mixing the ocean. The EW experiment without atmospheric synchronisation effectively damps the impact of fluxes in the ocean, which can cause a bias. Second, the Pacific Decadal Oscillation (PDO) over 1982–2006 is almost solely positive, while the interannual variability of our models is not synchronised with the observations. The bias may relate to a modulation of the PDO. Third, we did not apply synchronisation under sea ice, which may cause some artifacts near the ice edge.

The bias in the SINGLE multi-model mean (see Figure 8c) has a comparable spatial pattern than in the CESM model of NI, but is larger because of the interpolation error in the pseudo-observation grid. It does not outperform the multi-model mean of NI.

We further analyze the results in the ENSO region, which stands out as one of the regions where our ocean-connected supermodel reaches good synchronisation and reduced bias. The time series of the Niño 3.4 index (SST anomaly in the region, 5°S–5°N and 120°W–170°W) is presented in Figure 10. In the NI (Figure 10), the individual models produce large ENSO variability that is not synchronised. The ensemble mean shows one prominent peak when a strong El Niño event occurs (by chance) in phase in 2003 in all models. The probability of such a coincidental occurrence scale with the power of the number of models. In our example (with three models and El Niño occurring typically every 4 years), such an event will happen every 64 years. Consequently, the NI methods will underestimate extreme events, particularly with more models. Contrarily, in EW and SINGLE, all models are evolving in phases.

To analyze further the representation of ENSO events, we compare the PDF of the Niño 3.4 index for 1982–2006 in Figure 11. The tropical Pacific variability is notoriously asymmetric, with the magnitude of SST anomalies over the eastern equatorial Pacific being more prominent during the warm phase than during the cold phase (e.g., T. Zhang et al., 2017). By taking the mean of unconnected and independent models, one expects the PDF to show less variability (being steeper around 0) and becoming more Gaussian (getting less skewed). The latter is because the average of skewed distributions will converge toward a Gaussian distribution (central limit theorem). As expected, there is damping of variability in the NI model mean, while in EW and SINGLE, the standard deviation

Figure 9. Difference of the absolute climatological sea surface temperature bias (EW-NI) of the multi-model mean over 1982–2006. Negative values indicate that the bias in EW is smaller than in NI.

reasonably matches the observed PDF. All multi-model means show a more skewed PDF than in the observations. However, the period 1982–2006 was quite anomalous in the observations, as no significant ENSO events occurred beyond 1998. In comparison, the standard deviation and skewness computed for 1950–2005 were respectively 0.87, and 0.89 (T. Zhang et al., 2017). We do not see the reduction of skewness expected in NI, but we think it is artificially high here because of the coincidental El Niño event, referred to above. Obtaining more statistically robust results requires much longer runs and thus we limit our analysis to these basic metrics (Wittenberg, 2009).

To assess the ability of the supermodel framework to reproduce ENSO teleconnections we have analyzed the correlation of Niño3.4 SST index with global SST, precipitation, and SLP. Overall, the teleconnections of the individual models are very similar to that of the multi-model mean (not shown).

Figure 10. Niño 3.4 time series of the three models in the NI ensemble (left), the EW (middle) and the SINGLE (right). The black line shows the multi-model ensemble mean.

To first order all model configurations agree reasonably well with the observed patterns. To illustrate this we only show results for precipitation for boreal summer (Figure 12). There are subtle differences among the NI, EW, and SINGLE simulations. During June–September, teleconnections with Africa are changing from a positive correlation in NI to a negative correlation in EW, as in the observations. In some regions EW degrades the teleconnection pattern, such as in June-July-August-September over Australia where SINGLE performs best. These results are nevertheless encouraging. They show that a connected supermodel can reasonably simulate atmospheric teleconnection patterns and that the pattern could be potentially improved through training.

5. Conclusions and Future Perspectives

This paper investigates critical characteristics for developing the first supermodel of ESMs with ocean synchronisation. Synchronisation with models having different grids and structures is handled by assimilating synthetic pseudo-observations of SST constructed from the multi-model every month. We show that such a framework can achieve partial synchronisation in the ocean component and in the atmospheric component in regions where the ocean drives the climate variability—where the impact of atmospheric fluxes is limited in relation to the oceanic timescales.

We compared two methodologies for constructing the pseudo-observations, either from a single model or an equal-weight mean. The latter tends to reduce model bias in the synchronised region compared to the unconnected version. The variability of the multi-model mean in the two connected models is smaller than in the observations. This variability is damped less in synchronised regions, while it converges to that of the non-interactive multi-model mean in the unsynchronized regions. The damping is more pronounced in the equal-weight version than in the version connected to a single model, as in the former, the variability of all of the individual models is also reduced. The multi-model averaging of unsynchronized variability (e.g., driven by chaotic atmospheric variability) reduces variability in the assimilated pseudo-observations causing a deflation when updating the model snapshots.
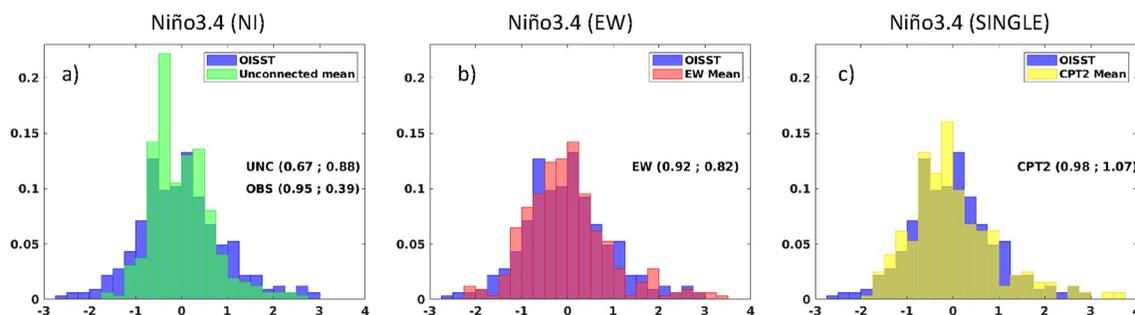
Figure 11. Histogram of Niño 3.4 over 1982–2006. The blue bars are for OISSTV2; the green bars are for the multi-model mean of NI (a), the red for EW (b) and the yellow for SINGLE (c). The standard deviation and skewness of the distribution are reported in parenthesis for observations and the multi-model mean.
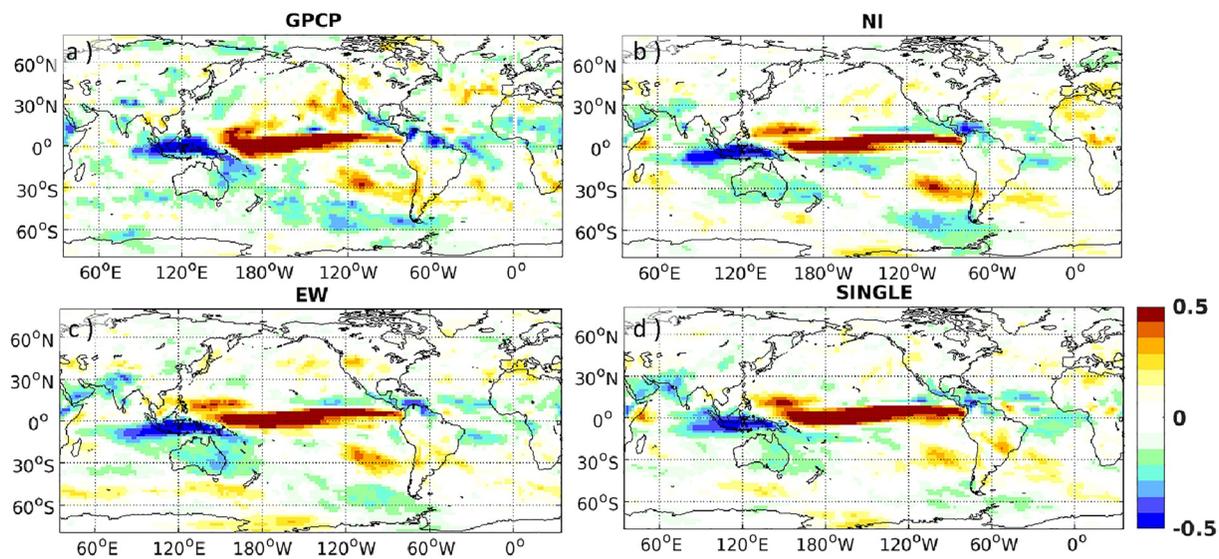
Figure 12. Averaged of detrended correlation of Niño3.4 indices with precipitation for June–September in the GCPC observations (a) multi-model mean of NI ensemble (b) EW (c) and SINGLE (d).

The system presented only uses a minimal amount of data exchange (monthly SST). Increasing the frequency of the synchronisation steps and assimilating more pseudo-observations will enhance synchronisation and may help to reduce the damping in the weighted mean supermodel framework. Ongoing works include a complete ocean pseudo-observations network (sea surface elevation, 3D hydrography), increasing the frequency of the ocean connection (to weekly synchronisation step) and complementing the system with synchronisation of other components of the ESM (ocean and sea ice component).

We will investigate in a future study how model error can reduce with the training of the weights. It is unclear which of the two framework presented here will, at term, achieve the best performance. Weighted-mean super-models have drawbacks in a partial synchronisation regime (variability damping) not faced with a connection to a single model. However, a weighted supermodel can provide locally a more accurate fit to the observation than a version where models are connected to a single model becaus —in the latter, skill is bounded by the best model. However, the single model to which we connect can vary spatially and for different variables (Du & Smith, 2017; Schevenhoven & Carrassi, 2021; Smith, 2001). Furthermore, if one can afford an ensemble of supermodels (with several members for each model), models could be synchronised from a randomly drawn single member/model every time, so that the frequency of the optimal weight is satisfied. Such a scheme resembles the cross-pollination in time, envisioned in Du and Smith (2017).

We will explore other alternatives to reduce the spurious damping in the weighted mean supermodel version by (a) adding back the reduced atmospheric driven variability term in the pseudo-observations or (b) isolating the part of variability that can be synchronised from the model snapshots (observation operator). In particular, machine learning techniques have emerged as powerful tools to carry process-identification (e.g., Sonnewald et al., 2019).

Data Availability Statement

References

Adler, R.-F., Huffman, G.-J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., et al. (2003). The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, *4*(6), 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:tvgpcp>2.0.co;2

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, *90*(9), 1283–1296. https://doi.org/10.1175/2009bams2618.1

Back, L. E., & Bretherton, C. S. (2009). On the relationship between SST gradients, boundary layer winds, and convergence over the tropical oceans. *Journal of Climate*, *22*(15), 4182–4196. https://doi.org/10.1175/2009jcli2392.1

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, Ø., et al. (2012). The Norwegian Earth system model, NorESM1-M—Part 1: Description and basic evaluation. *Geoscientific Model Development Discussions*, *5*, 2843–2931. https://doi.org/10.5194/gmdd-5-2843-2012

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review*, *97*(3), 163–172. https://doi.org/10.1175/1520-0493(1969)097<0163:atftep>2.3.co;2

Block, K., & Mauritsen, T. (2013). Forcing and feedback in the MPI-ESM-LR coupled model under abruptly quadrupled CO_2. *Journal of Advances in Modeling Earth Systems*, *5*(4), 676–691. https://doi.org/10.1002/jame.20041

Branicki, M., & Majda, A.-J. (2015). An information-theoretic framework for improving imperfect dynamical predictions via multi-model ensemble forecasts. *Journal of Nonlinear Science*, *25*(3), 489–538. https://doi.org/10.1007/s00332-015-9233-1

Cai, W., Wu, L., Lengaigne, M., Li, T., McGregor, S., Kug, J.-S., et al. (2019). Pantropical climate interactions. *Science*, *363*(6430). https://doi.org/10.1126/science.aav4236

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535. https://doi.org/10.1002/wcc.535

Counillon, F., & Bertino, L. (2009). Ensemble optimal interpolation: Multivariate properties in the Gulf of Mexico. *Tellus*, *61*(2), 296–308. https://doi.org/10.3402/tellusa.v61i2.15549

Du, H., & Smith, L.-A. (2017). Multi-model cross-pollination in time. *Physica D: Nonlinear Phenomena*, *353*, 31–38. https://doi.org/10.1016/j.physd.2017.06.001

Duane, G.-S., & Tribbia, J.-J. (2001). Synchronized chaos in geophysical fluid dynamics. *Physical Review Letters*, *86*(19), 4298–4301. https://doi.org/10.1103/physrevlett.86.4298

Duane, G.-S., Wiegerinck, W., Selten, F., Shen, M.-L., & Keenlyside, N. (2018). Supermodeling: Synchronization of alternative dynamical models of a single objective process. In *Advances in nonlinear geosciences* (pp. 101–121). Springer.

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Fereday, D., O'Reilly, C., et al. (2018). Skilful seasonal predictions of summer European rainfall. *Geophysical Research Letters*, *45*(7), 3246–3254. https://doi.org/10.1002/2017gl076337

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Eyring, V., Bony, S., Meehl, G.-A., Senior, C.-A., Stevens, B., Stouffer, R.-J., & Taylor, K.-E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Gill, A. E. (1980). Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society*, *106*(449), 447–462. https://doi.org/10.1002/qj.49710644905

Giorgetta, M.-A., Jungclaus, J., Reick, C.-H., Legutke, S., Bader, J., Böttinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project Phase 5. *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597. https://doi.org/10.1002/jame.20038

Halem, M., & Dlouhy, R. (1984). *Observing system simulation experiments related to space-borne lidar wind profiling. Part 1: Forecast impacts of highly idealized observing systems*. Res. Rev.

Hoskins, B. J., & Karoly, D. J. (1981). The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of the Atmospheric Sciences*, *38*(6), 1179–1196. https://doi.org/10.1175/1520-0469(1981)038<1179:tslroa>2.0.co;2

Hurrell, J., Holland, M., Gent, P., Ghan, S., Kay, J., Kushner, P., et al. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*(9), 1339–1360. https://doi.org/10.1175/bams-d-12-00121.1

Ilyina, T., Six, K.-D., Segschneider, J., Maier-Reimer, E., Li, H., & Núñez-Riboni, I. (2013). Global ocean biogeochemistry model hamocc: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modeling Earth Systems*, *5*(2), 287–315. https://doi.org/10.1029/2012ms000178

Janjić, T., Bormann, N., Bocquet, M., Carton, J., Cohn, S., Dance, S.-L., et al. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1257–1278. https://doi.org/10.1002/qj.3130

Jungclaus, J., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., et al. (2013). Characteristics of the ocean simulations in the max Planck institute ocean model (MPIOM) the ocean component of the MPI-Earth system model. *Journal of Advances in Modeling Earth Systems*, *5*(2), 422–446. https://doi.org/10.1002/jame.20023

Karspeck, A.-R., Danabasoglu, G., Anderson, J., Karol, S., Collins, N., Vertenstein, M., et al. (2018). A global coupled ensemble data assimilation system using the community Earth system model and the data assimilation research testbed. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2404–2430. https://doi.org/10.1002/qj.3308

Kay, J.-E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The community Earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/bams-d-13-00255.1

Keenlyside, N., Kosaka, Y., Vigaud, N., Robertson, A., Wang, Y., Dommenget, D., et al. (2020). Basin interactions and predictability. *Interacting Climates of Ocean Basins*, 258–292. https://doi.org/10.1017/9781108610995.009

Kirkevåg, A., Iversen, T., Seland, Ø., Hoose, C., Kristjánsson, J., Struthers, H., et al. (2013). Aerosol–climate interactions in the Norwegian Earth system model–NorESM1-M. *Geoscientific Model Development*, *6*(1), 207–244. https://doi.org/10.5194/gmd-6-207-2013

Kirtman, B. P., Bitz, C., Bryan, F., Collins, W., Dennis, J., Hearn, N., et al. (2012). Impact of ocean model resolution on ccsm climate simulations. *Climate Dynamics*, *39*(6), 1303–1328. https://doi.org/10.1007/s00382-012-1500-3

Kirtman, B.-P., Pegion, K., & Kinter, S.-M. (2004). *Internal atmospheric dynamics and tropical indo-pacific climate variability* (Vol. 165). Center for Ocean-Land-Atmosphere Studies.

Kirtman, B.-P., & Shukla, J. (2002). Interactive coupled ensemble: A new coupling strategy for cgcms. *Geophysical Research Letters*, *29*(10), 5–1. https://doi.org/10.1029/2002gl014834

Koseki, S., Keenlyside, N., Demissie, T., Toniazzo, T., Counillon, F., Bethke, I., et al. (2018). Causes of the large warm bias in the Angola–benguela frontal zone in the Norwegian Earth system model. *Climate Dynamics*, *50*(11), 4651–4670. https://doi.org/10.1007/s00382-017-3896-2

Krishnamurti, T., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T., & Ross, R. (2016). A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Reviews of Geophysics*, *54*(2), 336–377. https://doi.org/10.1002/2015rg000513

Kushnir, Y., Robinson, W., Bladé, I., Hall, N., Peng, S., & Sutton, R. (2002). Atmospheric gcm response to extratropical SST anomalies: Synthesis and evaluation. *Journal of Climate*, *15*(16), 2233–2256. https://doi.org/10.1175/1520-0442(2002)015<2233:agrtes>2.0.co;2

Lamarque, J.-F., Bond, T.-C., Eyring, V., Granier, C., Heil, A., Klimont, Z., et al. (2010). Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: Methodology and application. *Atmospheric Chemistry and Physics*, *10*(15), 7017–7039. https://doi.org/10.5194/acp-10-7017-2010

Lindzen, R. S., & Nigam, S. (1987). On the role of sea surface temperature gradients in forcing low-level winds and convergence in the tropics. *Journal of the Atmospheric Sciences*, *44*(17), 2418–2436. https://doi.org/10.1175/1520-0469(1987)044<2418:otross>2.0.co;2

Marsland, S. J., Haak, H., Jungclaus, J.-H., Latif, M., & Röske, F. (2003). The max-planck-institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modelling*, *5*(2), 91–127. https://doi.org/10.1016/s1463-5003(02)00015-x

Mirchev, M., Duane, G.-S., Tang, W. K., & Kocarev, L. (2012). Improved modeling by coupling imperfect models. *Communications in Nonlinear Science and Numerical Simulation*, *17*(7), 2741–2751. https://doi.org/10.1016/j.cnsns.2011.11.003

Nerger, L., Tang, Q., & Mu, L. (2020). Efficient ensemble data assimilation for coupled models with the parallel data assimilation framework: Example of AWI-CM (AWI-CM-PDAF 1.0). *Geoscientific Model Development*, *13*(9), 4305–4321. https://doi.org/10.5194/gmd-13-4305-2020

Oke, P.-R., Allen, J.-S., Miller, R.-N., Egbert, G.-D., & Kosro, P.-M. (2002). Assimilation of surface velocity data into a primitive equation coastal ocean model. *Journal of Geophysical Research*, *107*(C9), 5–1. https://doi.org/10.1029/2000jc000511

Palmer, T., & Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24390–24395. https://doi.org/10.1073/pnas.1906691116

Pecora, L.-M., Carroll, T.-L., Johnson, G.-A., Mar, D.-J., & Heagy, J.-F. (1997). Fundamentals of synchronization in chaotic systems, concepts, and applications. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *7*(4), 520–543. https://doi.org/10.1063/1.166278

Reick, C., Raddatz, T., Brovkin, V., & Gayler, V. (2013). Representation of natural and anthropogenic land cover change in MPI-ESM. *Journal of Advances in Modeling Earth Systems*, *5*(3), 459–482. https://doi.org/10.1002/jame.20022

Reynolds, R.-W., Rayner, N.-A., Smith, T.-M., Stokes, D.-C., & Wang, W. (2002). An improved in situ and satellite SST analysis for climate. *Journal of Climate*, *15*(13), 1609–1625. https://doi.org/10.1175/1520-0442(2002)015<1609:aiisas>2.0.co;2

Richter, I., Behera, S.-K., Doi, T., Taguchi, B., Masumoto, Y., & Xie, S.-P. (2014). What controls equatorial atlantic winds in boreal spring? *Climate Dynamics*, *43*(11), 3091–3104. https://doi.org/10.1007/s00382-014-2170-0

Sakov, P., Counillon, F., Bertino, L., Lisæter, K., Oke, P., & Korablev, A. (2012). TOPAZ4: An ocean-sea ice data assimilation system for the North Atlantic and Arctic. *Ocean Science*, *8*(4), 633–656. https://doi.org/10.5194/os-8-633-2012

Scaife, A.-A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *Npj Climate and Atmospheric Science*, *1*(1), 1–8. https://doi.org/10.1038/s41612-018-0038-4

Schevenhoven, F., & Carrassi, A. (2021). Training a supermodel with noisy and sparse observations: A case study with cpt and the synch rule on speedo-v. 1. *Geoscientific Model Development Discussions*, 1–23.

Schevenhoven, F., & Selten, F. (2017). An efficient training scheme for supermodels. *Earth System Dynamics*, *8*(2), 429–438. https://doi.org/10.5194/esd-8-429-2017

Schevenhoven, F., Selten, F., Carrassi, A., & Keenlyside, N. (2019). Improving weather and climate predictions by training of supermodels. *Earth System Dynamics*, *10*(4), 789–807. https://doi.org/10.5194/esd-10-789-2019

Schneck, R., Reick, C.-H., & Raddatz, T. (2013). Land contribution to natural CO_2 variability on time scales of centuries. *Journal of Advances in Modeling Earth Systems*, *5*(2), 354–365. https://doi.org/10.1002/jame.20029

Selten, F.-M., Schevenhoven, F.-J., & Duane, G.-S. (2017). Simulating climate with a synchronization-based supermodel. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(12), 126903. https://doi.org/10.1063/1.4990721

Shen, M.-L., Keenlyside, N., Bhatt, B.-C., & Duane, G.-S. (2017). Role of atmosphere-ocean interactions in supermodeling the tropical Pacific climate. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(12), 126704. https://doi.org/10.1063/1.4990713

Shen, M.-L., Keenlyside, N., Selten, F., Wiegerinck, W., & Duane, G. S. (2016). Dynamically combining climate models to? Supermodel? The tropical Pacific. *Geophysical Research Letters*, *43*(1), 359–366. https://doi.org/10.1002/2015gl066562

Shukla, J. (1998). Predictability in the midst of chaos: A scientific basis for climate forecasting. *Science*, *282*(5389), 728–731. https://doi.org/10.1126/science.282.5389.728

Smirnov, D., Newman, M., Alexander, M. A., Kwon, Y.-O., & Frankignoul, C. (2015). Investigating the local atmospheric response to a realistic shift in the oyashio sea surface temperature front. *Journal of Climate*, *28*(3), 1126–1147. https://doi.org/10.1175/jcli-d-14-00285.1

Smith, L.-A. (2001). Disentangling uncertainty and error: On the predictability of nonlinear systems. In *Nonlinear dynamics and statistics* (pp. 31–64). Springer.

Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, *6*(5), 784–794. https://doi.org/10.1029/2018ea000519

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M Earth system model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Tian, B., & Dong, X. (2020). The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophysical Research Letters*, *47*(8), e2020GL087232. https://doi.org/10.1029/2020gl087232

Tjiputra, J., Roelandt, C., Bentsen, M., Lawrence, D., Lorentzen, T., Schwinger, J., et al. (2013). Evaluation of the carbon cycle components in the Norwegian Earth system model (NorESM). *Geoscientific Model Development*, *6*(2), 301–325. https://doi.org/10.5194/gmd-6-301-2013

Van den Berge, L.-A., Selten, F.-M., Wiegerinck, W., & Duane, G.-S. (2011). A multi-model ensemble method that combines imperfect models through learning. *Earth System Dynamics*, *2*(1), 161–177. https://doi.org/10.5194/esd-2-161-2011

Vertenstein, M., Craig, T., Middleton, A., Feddema, D., & Fischer, C. (2012). CESM1.0.3 user guide. UCAR doc (Vol. 12). Retrieved from https://www2.cesm.ucar.edu/models/cesm1.0/cesm/cesm_doc_1_0_6/ug.pdf

Wallace, J. M., & Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, *109*(4), 784–812. https://doi.org/10.1175/1520-0493(1981)109<0784:titghf>2.0.co;2

Wang, Y., Counillon, F., Keenlyside, N., Svendsen, L., Gleixner, S., Kimmritz, M., et al. (2019). Seasonal predictions initialized by assimilating sea surface temperature observations with the enkf. *Climate Dynamics*, *53*(9–10), 5777–5797. https://doi.org/10.1007/s00382-019-04897-9

Wiegerinck, W., Burgers, W., & Selten, F. (2013). On the limit of large couplings and weighted averaged dynamics. In *Consensus and synchronization in complex networks* (pp. 257–275). Springer.

Wiegerinck, W., & Selten, F. (2017). Attractor learning in synchronized chaotic systems in the presence of unresolved scales. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(12), 126901. https://doi.org/10.1063/1.4990660

Wittenberg, A. T. (2009). Are historical records sufficient to constrain enso simulations? *Geophysical Research Letters*, *36*(12), L12702. https://doi.org/10.1029/2009gl038710

Zhang, S., Harrison, M., Rosati, A., & Wittenberg, A. (2007). System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Monthly Weather Review*, *135*(10), 3541–3564. https://doi.org/10.1175/mwr3466.1

Zhang, T., Shao, X., & Li, S. (2017). Impacts of atmospheric processes on ENSO asymmetry: A comparison between cesm1 and ccsm4. *Journal of Climate*, *30*(23), 9743–9762. https://doi.org/10.1175/jcli-d-17-0360.1

Zhang, W., & Kirtman, B.-P. (2019). Estimates of decadal climate predictability from an interactive ensemble model. *Geophysical Research Letters*, *46*(6), 3387–3397. https://doi.org/10.1029/2018gl081307

Zhongming, Z., Linong, L., Wangqiang, Z., & Wei, L. (2020). Europe is building a "digital twin" of Earth to revolutionize climate forecasts. *Science*.

Zhu, J., Kumar, A., Lee, H.-C., & Wang, H. (2017). Seasonal predictions using a simple ocean initialization scheme. *Climate Dynamics*, *49*(11), 3989–4007. https://doi.org/10.1007/s00382-017-3556-6