Climate Services

Application-specific optimal model weighting of global climate models: A red tide example

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research paper
Keywords:	weighting; subset selection; multi-model ensemble; multi-objective optimization; decision-relevant metrics; climate models and Earth system models
Corresponding Author:	Ming Ye Florida State University UNITED STATES
First Author:	Ahmed S. Elshall
Order of Authors:	Ahmed S. Elshall
	Ming Ye
	Kranz Sven
	Harrington Julie
	Yang Xiaojuan
	Wan Yongshan
	Maltrud Mathew
Abstract:	Global climate models (GCMs) and Earth system models (ESMs) provide many climate services. The High Resolution Model Intercomparison Project (HighResMIP) of the Coupled Model Intercomparison Project Phase 6 (CMIP6) provide model runs that address regional phenomena. Ensemble members are formulated from these model runs. An ensemble member can be a model run, a single-model ensemble containing model runs of the same model with perturbed parameters, initialization, physics, and forcings, or ensemble of dependent models with each model having one or more model runs. Thus, developing a parsimonious ensemble requires multiple ensemble methods such as independent-model subset selection, prescreening-based subset selection, and model weighting. The manuscript focuses on application-specific optimal model weighting, with prescreening-based subset selection. As such, independent ensemble members are categorized, selected, and weighted based on their ability to reproduce physically-interpretable features of interest that are problem-specific. We discuss the strengths and caveats of optimal model weighting using a case study of red tide prediction in the Gulf of Mexico along the West Florida Shelf. Red tide is the common name of harmful algal blooms that occurs worldwide, causing adverse socioeconomic and environmental impacts. Our results show the importance of prescreening-based subset selection, as optimal model weighting can underplay robust ensemble members by optimizing error cancellation. Prescreening-based subset selection can also provide insights about the validity of the model weights. By illustrating the caveats of using non-representative models when optimal model weighting is used, the findings and discussion of this study are pertinent to many other climate services.
Suggested Reviewers:	Hai Pham, PhD Hai.Pham@dri.edu Expert on model weighting
	Jina Yin, PhD jnyin@hhu.edu.cn Expert on model weighting
	Christine Shoemaker, PhD shoemaker@nus.edu.sg Expert on optimization methods
Opposed Reviewers:	

September 10, 2021

Editorial Board
Climate Services

Dear Editorial Board:

Enclosed please find the manuscript titled “Application-specific optimal model weighting of global climate models: A red tide example.” The manuscript is being submitted for possible publication in Climate Services. The manuscript has not been submitted elsewhere. The coauthors are aware and approve of this submission.

The manuscript addresses an important topic in climate services that is decision-relevant metrics. We present a case study with environmental relevance. We consider the two approaches of optimal model weighting, and prescreening-based subset selection to improve ensemble predictions of Earth system models. Independent ensemble members are categorized, selected, and weighted based on their ability to reproduce physically-interpretable features of interest that are problem-specific. The message that we want to convey to the community is the importance of using prescreening-based subset selection with decision relevant metrics to identify non-representative models and understand their impact on ensemble prediction.

We appreciate your help with this submission.

Thank you very much in advance for your consideration.

Sincerely,

Ming Ye
Professor
Department of Earth, Ocean, and Atmospheric Science
Florida State University
mye@fsu.edu

Application-specific optimal model weighting of global climate models: A red tide example

Ahmed S. Elshall[1], Ming Ye[1*], Sven A. Kranz[1], Julie Harrington[2], Xiaojuan Yang[3], Yongshan Wan[4] , and Mathew Maltrud[5]

[1] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, United States of America
[2] Center for Economic Forecasting and Analysis, Florida State University, Tallahassee, FL, United States of America
[3] Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, United States of America
[4] Center for Environmental Measurement and Modeling, United States Environmental Protection Agency, Gulf Breeze, FL, United States of America
[5] Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, United States

*** Correspondence:**
Ming Ye
mye@fsu.edu

Submitted for Publication in Climate Services
September 2021

Abstract

Global climate models (GCMs) and Earth system models (ESMs) provide many climate services. The High Resolution Model Intercomparison Project (HighResMIP) of the Coupled Model Intercomparison Project Phase 6 (CMIP6) provide model runs that address regional phenomena. Ensemble members are formulated from these model runs. An ensemble member can be a model run, a single-model ensemble containing model runs of the same model with perturbed parameters, initialization, physics, and forcings, or ensemble of dependent models with each model having one or more model runs. Thus, developing a parsimonious ensemble requires multiple ensemble methods such as independent-model subset selection, prescreening-based subset selection, and model weighting. The manuscript focuses on application-specific optimal model weighting, with prescreening-based subset selection. As such, independent ensemble members are categorized, selected, and weighted based on their ability to reproduce physically-interpretable features of interest that are problem-specific. We discuss the strengths and caveats of optimal model weighting using a case study of red tide prediction in the Gulf of Mexico along the West Florida Shelf. Red tide is the common name of harmful algal blooms that occurs worldwide, causing adverse socioeconomic and environmental impacts. Our results show the importance of prescreening-based subset selection, as optimal model weighting can underplay robust ensemble members by optimizing error cancellation. Prescreening-based subset selection can also provide insights about the validity of the model weights. By illustrating the caveats of using non-representative models when optimal model weighting is used, the findings and discussion of this study are pertinent to many other climate services.

1 Introduction

Global climate models (GCMs) and Earth system models (ESMs) offer climate services that are important for societal decision making, across water, energy, food, health, environment, and many other sectors (White et al. 2017; van den Hurk et al. 2018; Eyring et al. 2019). The subtle difference between GCMs and ESMs is that ESMs have higher fidelity with respect processes inclusion and realism, GCMs through accounting for landuse, plant ecology, ocean biogeochemistry and ecology, atmospheric chemistry, and other processes. ESMs can be particularly useful for environmental, ecological, and coastal management (Payne et al. 2019; Jacox et al. 2020; Ward et al. 2020; Dixon et al. 2021; Tonelli et al. 2021). Hereafter, ESMs refer to both ESMs and GCMs. We provide a case study about using ESMs to address the problem of red tide, which is a common name of harmful algae blooms that occur in coastal regions worldwide due to high concentrations of marine microorganisms such as dinoflagellates, diatoms, and protozoans. Along the West Florida Shelf in the Gulf of Mexico, red tide occurs by the increase of the concentration of *Karenia brevis*, a toxic mixotrophic dinoflagellate. Red tide results in adverse impacts on fisheries by causing shellfish poisoning and massive fish kills, on public health by causing respiratory skin, and eye, irritation, on ecosystem services by harming sea turtles, marine mammals, and birds, on local communities by impacting tourism, and reactional activities due to the unpleasant scene and odor of red tide, and health concerns. This study focuses on Loop Current (LC), which is one of the main drivers of red tide in the West Florida Shelf (Weisberg et al. 2014; Maze et al. 2015; Perkins 2019). LC is a warm ocean current that penetrates and loops through the Gulf of Mexico until exiting the gulf to join the Gulf Stream. Several relations have been established between red tide and LC (Weisberg et al. 2014, 2019; Maze et al. 2015; Liu et al. 2016). The relation discussed in Maze et al. (2015) shows that the LC position, which can be inferred from sea surface heigh, can be a definitive predictor of a large red tide bloom possibility. Using the relation of Maze et al. (2015), we present an application-specific optimal model weighting method to improve the predictive performance of ESMs.

To improve and extract relevant information from ESMs, multiple techniques such as bias correction, downscaling, and ensemble methods are often employed. A commonly used ensemble method is model weighting, through assigning unequal weights to ensemble members (Sanderson et al. 2017; Lorenz et al. 2018; Herger et al. 2018; Merrifield et al. 2020; Brunner et al. 2020). An ensemble member can be a single model run (i.e., a model simulation or projection), many model runs of a single model with multiple realizations, or an ensemble of many dependent models each with a single or multiple model runs. Advancing methods for model weighting is thus needed to refine the most credible information on regional climate changes, impacts, and risks for stakeholders (Eyring et al. 2016). As there is no single best ESM, there is no universally best model weighting method, but a method may be useful given the criteria relevant for the application in question (Herger et al. 2018). Model democracy, which is the equal-weighting method, is the simplest model weighting method. Yet more advanced model weighting methods are needed depending on the model evaluation criteria.

Model weighting can be based on a single or combination of model evaluation criteria. One criterion is to assign model weights based on model performance. Performance-based model weighting methods include Bayesian model averaging, evaluation of probability density function, climate prediction index, upgraded reliability ensemble averaging, skill score of representing

annual cycle, and others as compared by several studies (Oh and Suh 2017; Zhang and Yan 2018; Wang et al. 2019). Performance-based model weighting methods consider the differences of model simulations to historical observation, and they differ in the metrics and algorithms used to determine model weights (Wang et al. 2019). For example, Oh and Suh (2017) compare three model weighting methods, which are weighted ensemble averaging based on root-mean-square error (RMSE) and correlation, the skill score of the representation of the annual cycle based on Taylor score (i.e., accounting for correlation coefficient, standard deviations, and centered RMSE), and multivariate linear regression that minimizes the RMSE of the ensemble prediction using least squares methods. In addition to model performance, model independence and convergence are two additional criteria. Multi-criteria-based model weighting methods extend beyond the model performance criterion to assign model weights. For example, the performance and interdependence skill method uses model bias to historical observation (performance criterion) and model distance to other ensemble members (interdependence criterion) to assign model weights (Knutti et al. 2017; Wang et al. 2019). Wang et al. (2019) assign model weights by using the reliability ensemble averaging method that considers both model bias to historical observation (performance criterion) and model similarity to other models in the future projection (convergence criterion). A fourth criterion for assigning model weights is intermodel comparison for observable climate and future climate (Räisänen and Ylhäisi 2012). As such, the closeness of two models in simulating observable climate and future climate is checked. For example, the Bayesian weighted averaging method of Xu et al. (2019) considers the model skills in reproducing historical observations and inter-model agreement in simulating future period to assign model weights.

This study complements an important aspect of model weighting by explicitly considering application-specific metrics. Given this additional criterion for model evaluation, the model performance is explicitly evaluated for its suitability for specific applications, apart from the regional and global predictive performance of the model. The evaluation includes process-based metrics and other relevant features, given a specific problem definition. Considering process-based emergent constraints is a promising way to focus evaluation on the observations most relevant to climate projections (Eyring et al. 2016). By using an optimal model weighting method, application-specific model weighting is accounted for in the objective function such that the ensemble is optimized given problem-specific and process-based features of the problem of interest. We use a multi-objective optimal ensemble method based on an objective function that defines the desired targets. For example, if the objective is to reduce regional bias, RMSE can be the objective function, and the output will be the lowest possible RMSE of the ensemble prediction and the observational product, giving possible combinations of the model weights of the ensemble members.

The proposed application-specific optimal model weighting method has several practical advantages. First, the flexibility in ensemble calibration by defining an adjustable objective function allows this method to be applied to a wide range of problems, with the meaning of "optimal" varying depending on the aim of the study (Herger et al. 2018). Second, an optimization method can readily account simultaneously for multi-objectives such as multiple variables of precipitation, sea surface temperature, and wind (Herger et al. 2019), and multiple metrics such as RMSE and spatial correlation in climate change information (Bhowmik and Sankarasubramanian 2021). Third, multi-objectives can account for metrics related to the

application of interest. For example, Wang et al. (2019) note that the process from climate variables to hydrological responses is nonlinear, and thus the assigned model weights based on performances of the climate simulations may not be correctly translated to hydrological responses. Thus, assigning model weights to the outputs of ESMs based on their ability to represent the climate variable of interest is more straightforward than accounting for other decision relevant metrics (e.g., occurrence or non-occurrence of large red tide bloom). In the remaining of the manuscript, Section2 presents the application-specific optimal model weighting method for the red tide case study, followed by the results with respect to model weights and predictive performance (Section 3). We discuss in Section 4 the pros and cons of model weighting, and conclude by summarizing our main findings and providing a research outlook in Section 5.

2 Method

2.1 Data

We select all the model runs of the Coupled Model Intercomparison Project Phase 6 (CMIP6) for both historical experiment (Eyring et al. 2016) and the hist-1950 experiment(Haarsma et al. 2016) with gridded monthly sea surface height above geoid (zos) and nominal resolution less than or equal 25 km. This resulted in 33 model runs (Table 1). The sea surface height above geoid has the variable name zos according to the Climate and Forecast (CF) metadata conventions. The two historical experiment and the hist-1950 experiment are from 1850-01 and 1950-01, respectively, to 2014-12. For analysis purpose, we additionally consider model runs with the standard resolution that are the E3SM-1-0 with variable ocean resolution of 30-60 km, and EC-Earth3P with nominal ocean resolution of about 100km. We account for model independence using institutional democracy (Leduc et al. 2016). For the same institution, we created further subsets given different grid resolutions, resulting in 11 independent model subsets (Table 1). Each independent model subset (IMS) constitutes an ensemble member. Each IMS contains one or more models, and each model has one or more model runs with perturbed realizations (r), initializations (i), physics (p), and forcings (f). For the reanalysis data of zos, we use the phy-001-030 global ocean eddy-resolving reanalysis product of the Copernicus Marine Environment Monitoring Service (CMEMS). This reanalysis product covers the altimetry from 1993 onward with approximatively 8 km horizontal resolution (Drévillon et al. 2018; Fernandez and Lellouche 2018).

Table1. Independent model subsets based on institutional democracy with ocean grid as a secondary criterion. An independent model subset (IMS) receives a score based on prescreening criteria (Section 2.3). The number of members (i.e., model runs) of each model can vary from one such as r1i1p1f1 of CESM1-CAM5-SE-HR to six such as r(1-6)i1p1f1 of ECMWF-IFS-HR.

IMS	Score	Institution	Country	Model (Reference)	Experiment ID	Members (Model Runs)	Ocean model resolution	Ocean grid
IMS01	1	NCAR	USA	CESM1-CAM5-SE-HR (Chang et al. 2020)	hist-1950	r1i1p1f1	0.1◦ (11 km) nominal resolution	POP2-HR
IMS02	2	CMCC	Italy	CMCC-CM2-HR4 (Cherchi et al. 2019)	hist-1950	r1i1p1f1	0.25◦ from the Equator degrading at the poles	ORCA025
				CMCC-CM2-VHR4 (Cherchi et al. 2019)	hist-1950	r1i1p1f1	0.25◦ from the Equator degrading at the poles	ORCA025
IMS03	1	CNRM-CERFACS	France	CNRM-CM6-1-HR (Voldoire et al. 2019)	hist-1950	r(1-3)i1p1f2	0.25◦ (27-28 km) nominal resolution	eORCA025
				CNRM-CM6-1-HR (Voldoire et al. 2019)	historical	r1i1p1f2	0.25◦ (27-28 km) nominal resolution	eORCA025
IMS04	0	DOE-E3SM-Project	USA	E3SM-1-0 (Golaz et al. 2019)	historical	r(1-5)i1p1f1	60 km in mid-latitudes and 30 km at the equator and poles	EC60to30
IMS05	0	EC-Earth-Consortium	Europe	EC-Earth3P (Haarsma et al. 2016)	hist-1950	r(1-3)i1p2f1	about 1◦ (110 km)	ORCA1
IMS06	2	EC-Earth-Consortium	Europe	EC-Earth3P-HR (Haarsma et al. 2016)	hist-1950	r(1-3)i1p2f1	about 0.25◦ (27-28 km)	ORCA025
IMS07	3	ECMWF	Europe	ECMWF-IFS-HR (Roberts et al. 2018)	hist-1950	r(1-6)i1p1f1	25 km nominal resolution	ORCA025
IMS08	3			ECMWF-IFS-MR (Roberts et al. 2018)	hist-1950	r(1-3)i1p1f1	25 km nominal resolution	ORCA025
IMS09	2	NOAA-GFDL	USA	GFDL-CM4 (Held et al. 2019)	historical	r1i1p1f1	0.25◦ (27-28 km) nominal resolution	tri-polar grid
				GFDL-ESM4 (Held et al. 2019)	historical	r(2-3)i1p1f1	0.25◦ (27-28 km) nominal resolution	tri-polar grid
IMS10	3	NERC	UK	HadGEM3-GC31-HH (Roberts et al. 2019)	hist-1950	r1i1p1f1	8 km nominal resolution	ORCA12
		MOHC-NERC	UK	HadGEM3-GC31-HM (Roberts et al. 2019)	hist-1950	r1i(1-3)p1f1	25 km nominal resolution	ORCA12
IMS11	3	MOHC	UK	HadGEM3-GC31-MM (Roberts et al. 2019)	hist-1950	r1i(1-3)p1f1	25 km nominal resolution	ORCA025
				HadGEM3-GC31-MM (Roberts et al. 2019)	historical	r(1-4)i1p1f3	25 km nominal resolution	ORCA025

We use the *Karenia brevis* cell count data of the harmful algal bloom database of the Fish and Wildlife Research Institute at the Florida Fish and the Wildlife Conservation Commission (FWRI 2020). According to Maze et al. (2015), a large red time bloom is defined as an event with the cell count exceeding 1×10^5 cells/L for ten or more successive days without a gap of more than five consecutive days, or 20% of the bloom length. Give the study period 1993-01 to 2014-12 with a six-month interval (i.e., a total of 44 intervals), we identified 15 intervals of large blooms, and 29 intervals with no bloom in the study area (Fig. 1).

2.2 Loop Current position and red tide blooms

LC is a warm ocean current that travels through the Gulf of Mexico. LC is an important factor that controls the occurrence of red tide (Perkins 2019). This occurs through varying deep ocean upwelling (Weisberg et al. 2014) and the retention time (Maze et al. 2015). In this study we focus on the retention time, and other relations (Weisberg et al. 2014, 2019; Liu et al. 2016) are warranted in future studies. The *Karenia brevis* is a slow growing dinoflagellate that requires an area with mixing slower than growth rate to form a bloom (Magaña and Villareal 2006). When LC in the north position (LC-N), this increases the retention allowing red tide blooms to form when other conditions are ideal (Maze et al. 2015). Accordingly, when the Loop Current is in the south position (LC-S) as shown in Fig. 1a, then there is no large bloom, while LC-N as shown in Fig.1b is a necessarily condition for red tide blooms to occur (Maze et al. 2015).

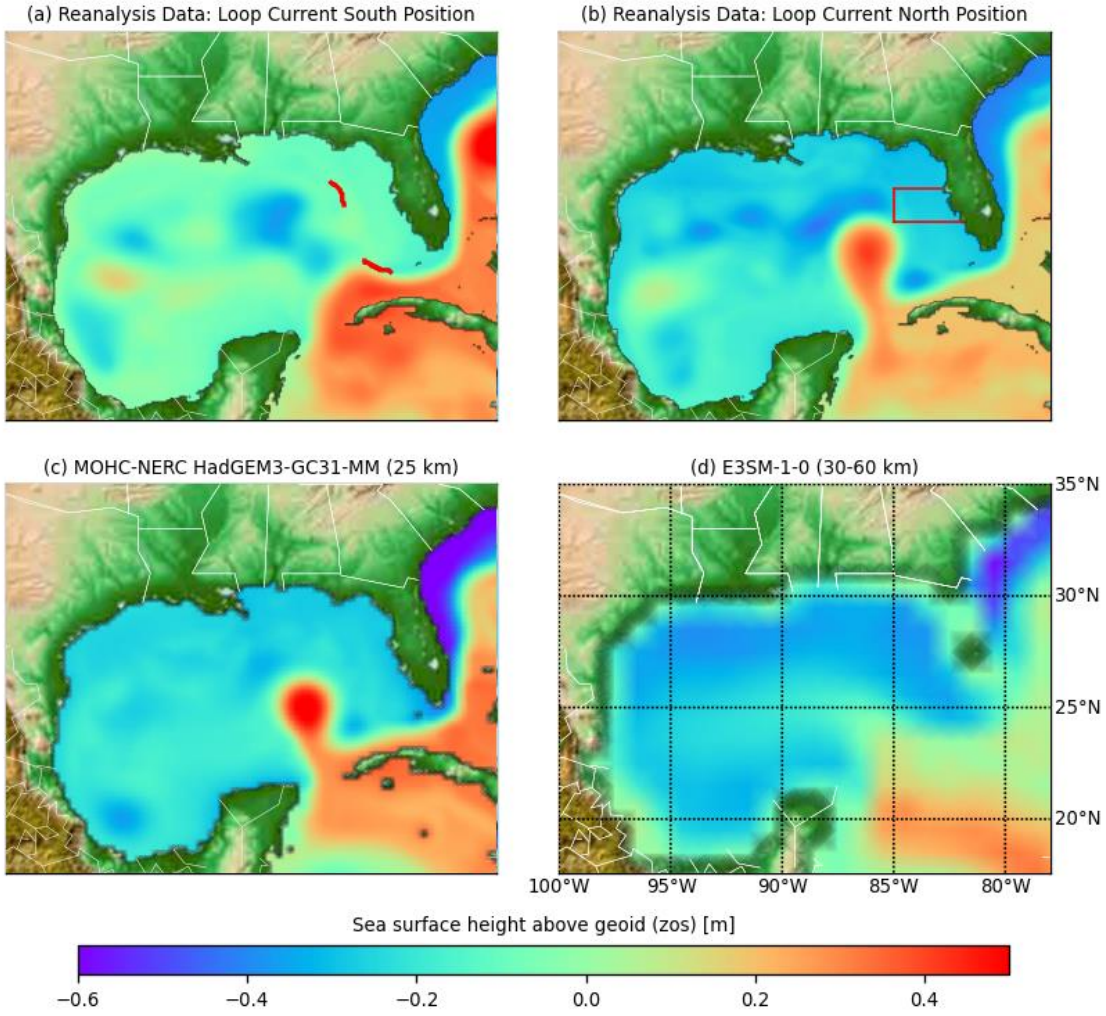

Figure 1. Snapshots of sea surface height above geoid (zos) showing (a) the Loop Current in the south position (LS-C) in 2010-03 for reanalysis data, and the Loop Current in the north position (LC-N) in 2010-06 for (b) reanalysis data, (c) a high-resolution ESM, and (d) a standard-resolution ESM. Two red segments along the 300m isobath in (a) are used to determine Loop Current position (i.e., LC-N and LC-S). The red box of (b) shows the study area, where red tide blooms are considered by this study and Maze et al. (2015).

The LC position is detected from sea surface height variability. Following the method of Maze et al. (2015) the zos anomaly between the north and south segments along the 300 m isobath (Fig. 1a) can be used as proxy for LC position such that positive and negative difference represents LC-N and LC-S, respectively. The zos anomaly per interval t can be estimated as

$$h_n = \max_{h_t}\left(\Delta_m \left[E_l \left[\sum_{k=1}^{K} w_k E_m(h_{j,k,l,m,t,n} \mid M_k) \right] \right] \right) \quad (1)$$

In this equation, we first take the expectation $E_m(.)$ for all model runs with index m in each ensemble member M_k, and then data is averaged for all ensemble members with index $k \in [1, K]$ where w_k is the weight of each ensemble member M_k. Subsequently, the exception $E_l(.)$ is taken for all data points with index l along each of the north and south segments, respectively. Afterward, we take the difference $\Delta_m(.)$ between the data of the two segments. Finally, for each of the 6-month interval the maximum zos anomaly $\max_{h_t}(.)$ is selected resulting in zos anomaly per interval $n \in [1, N]$, with $N = 44$ given the study period 1993-2015 and a 6-month interval length. Since we are not interested in the value of h_n per se but the sign difference between the north and south segments, we express Eq. 1 as an indicator function for LC-S

$$H_{LC-S}(h_n) = \begin{cases} 1, & h_n < 0 \\ 0, & h_n \geq 0 \end{cases} \quad (2)$$

such that $H_{LC-S}(h_t) = 1$ indicates a LC-S interval. We use Eqs. 1-2 to process CMIP6 and reanalysis data, which are hereafter donated by h_n and $h_{n,obs}$, respectively. We can further define the *oscillating event frequency*

$$x_0 = \frac{\sum_{n=1}^{N} H_{LC-S}(h_n)}{N} \quad (3)$$

as the ratio of the LC-S intervals to the total number of intervals T. The reanalysis data products of Maze et al. (2015) and this study result in $x_{0,obs} = 0.267$ and 0.273 (Fig.2a), respectively. The slight difference is not unexpected because the study period and the reanalysis data product used in this study are different from that of Maze et al. (2015). We compare $x_{0,obs} = 0.273$ of the reanalysis data with the model simulations in the results section.

2.3 Ensemble methods

The number (*K*) of model weights of each ensemble differs depending on the number of included ensemble members. Each independent model subset (IMS) listed in Table 1 is an ensemble member, as we account for model independence prior to model weighting (Eq.1). We consider two ensemble methods that are prescreening-based subset selection, and model weighting. Prior to model weighting, we include and exclude members from the ensemble based on prescreening-based subset selection criteria. These criteria are evolving such that each ensemble member receives a score from zero to three. Ensemble members that cannot simulate LC-N, as shown in Fig. 1d for example, receive a score zero (e.g., Fig. 2b). Ensemble members that can simulate LC-N, but without sign fluctuation to indicate both the LC-N and LC-S according to

Eq. 1, receive score one (e.g., Fig. 2c). The ensemble member receives a score of two if it can reproduce both LC-N and LC-S according to Eq.1 with the frequency of LC-N being smaller than that of LC-S (e.g., Fig. 3d). The ensemble member receives a score of three, if it can reproduce both LC-N and LC-S according to Eq.1 with the frequency of LC-N being greater than that of LC-S (e.g., Fig. 3e). Higher LC-N frequency is a more realistic condition, given reanalysis data with $x_{0,obs} = 0.273$ (Fig.2a).

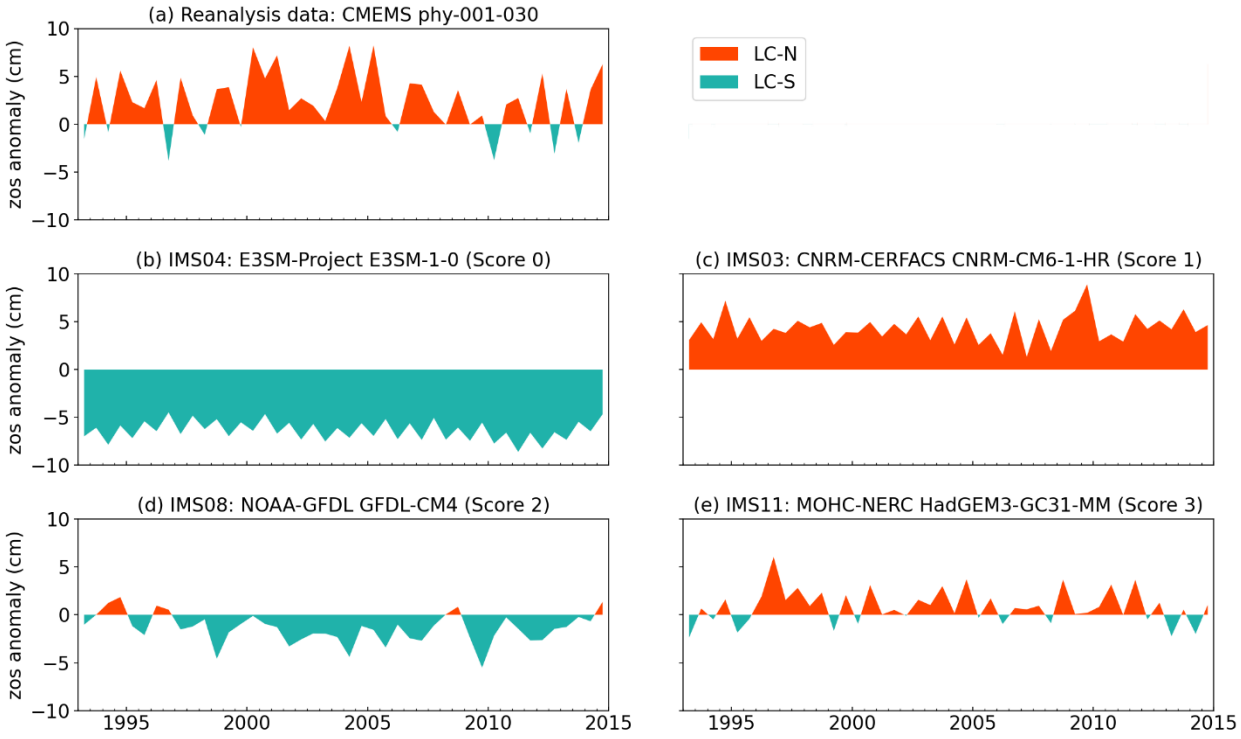

Figure 2. Surface height above geoid (zos) anomaly according to Eq.1 of (a) reanalysis data, and (b-e) enesmble members. The title of the reanalysis data shows the data provider name, and product ID. The title of ensemble member shows ensemble member number that is the number of each independent model subset (IMS): modeling group name, model name(s), and ensmble member score.

Given the defined prescreening-based subset selection criteria, we consider four ensemble compositions for the case of weighted-average multi-model ensemble (WME). For example, the WME3210 with $K = 11$ includes all ensemble members with a score from three to zero, while WME3XXX with $K = 4$ includes only the top performing ensemble members with only a score of three. We know from prior information (Caldwell et al. 2019; Hoch et al. 2020) that standard-resolution ESMs are generally incapable of simulating LC; see for example Fig. 1d. This is mainly because the standard resolution grids (e.g., Fig. 1d) cannot resolve the mesoscale eddies and boundary currents, and require global parametrization, unlike the high resolution eddy-permitting grids (e.g., Fig1c). Thus, we consider WME3210 and WME321X to evaluate the combined impacts of prior information and model weighting. We consider WME32XX and WME3XXX with different ensemble size of $K = 7$ and $K = 4$, respectively, to study the combined impact of subset selection and model weighting. To study the impacts of model weighting, we consider the case of simple-average multi-model ensemble (SME) using equal model weights that

lends SME3210, SME321X, SME32XX and SME3XXX, given the same ensemble composition criteria. SME321X with $K=9$ is the reference ensemble that only considers prior information without any prescreening-based subset selection and model weighting.

2.4 Optimal model weighting

Each ensemble member has a model weight. The model weights w_k in Eq. 1 satisfy

$$\sum_{k=1}^{K} w_k = 1 \tag{4}$$

and

$$w_k = 1/K \tag{5}$$

for equal model weighting. For unequal model weighting, w_k can be estimated using an optimization algorithm through minimizing an objective function with multiple objectives. In this study, the objective of the optimization problem is to estimate the model weights w_k in Eq. 1 that minimize the objective function f such that

$$\min_{w_k} f = \min_{w_k}\left[\prod_{i=1}^{5}(x_i+1)^{c_i}\right] \tag{6}$$

with five minimization objectives x_i such having an objective-weighting constant c_i. We formulate the objective function as (x_i+1) so that the product term $\prod_{i=1}^{5}(x_i+1)^{c_i}$ will not be zero if any objective x_i is fully achieved that is $x_i=0$. Accounting for multiple objectives can be achieved through Pareto-optimal solutions (Herger et al. 2019) or objective-weighting constants c_i as in this study. As such, each objective is assigned an objective-weighting constant c_i representing the importance of the objective relative to other objectives. The first minimization objective x_1 is the oscillating event count error

$$x_1 = \left|\sum_{n=1}^{N} H_{LC-S}(h_n) - \sum_{n=1}^{N} H_{LC-S}(h_{n,obs})\right| \tag{7}$$

between model simulation $H_{LC-S}(h_n)$ and reanalysis data $H_{LC-S}(h_{n,obs})$. The second minimization objective x_2 is the LC position temporal match error

$$x_2 = \frac{N - \sum_{n=1}^{N}(h_{n,obs}<0 \wedge h_n<0) - \sum_{n=1}^{N}(h_{n,obs}\geq 0 \wedge h_n \geq 0)}{N} \tag{8}$$

where $\sum_{n=1}^{N}(h_{n,obs}\geq 0 \wedge h_n \geq 0)$ and $\sum_{n=1}^{N}(h_{n,obs}<0 \wedge h_n<0)$ are the temporal match counts of model simulations and reanalysis data for LC-N and LC-S, respectively. The logical conjunction $\wedge$ gives a value of one when the statement $(h_{n,obs}<0 \wedge h_n<0)$ is true when $h_{n,obs}<0$ and $h_n<0$ are both true, otherwise gives a value of zero. No temporal match is generally expected between simulations of ESMs, and re-analysis data. The term temporal match as used in this manuscript refers to a pseudotemporal correspondence that captures the general pattern of a dynamic

process of the LC position given the heuristic relation (Eq. 1) with a coarse-temporal-resolution of six months. The third objective x_3 is the LC-S temporal match error

$$x_3 = \frac{\sum_{n=1}^{N} H_{LC-S}(h_{n,obs}) - \sum_{t=1}^{T} (h_{n,obs} < 0 \wedge h_n < 0)}{\sum_{n=1}^{N} H_{LC-S}(h_{n,obs})} \quad (9)$$

The fourth objective x_4 is the red tide bloom error

$$x_4 = \frac{\sum_{n=1}^{N} (h_n < 0 \wedge H(z_n) = 1)}{N_{bloom}} \quad (10)$$

which represents the false negative prediction of red tide bloom. This is the ratio of the number of LC-S coinciding with large bloom to the number of large-bloom N_{bloom} , such that $H(z_n)$ is an indicator function with one and zero for large bloom and no bloom, respectively. The fifth objective x_5 is the RMSE between model simulation and reanalysis data

$$x_5 = \sqrt{\frac{\sum_{n=1}^{N} (h_n - h_{n,obs})^2}{N}} \quad (11)$$

With respect to objective-weighting constants c_i , we set c_i = 1, assuming that all objectives are equally important.

A common practice to solve Eq.6 subject to Eqs.7-11 is to use an optimization algorithm such as genetic algorithm (Bhowmik and Sankarasubramanian 2021), mathematical programming solver (Herger et al. 2018), and Simple Cull algorithm (Herger et al. 2019). We minimize the objective function (Eq. 6) using the covariance matrix adaptation evolution strategy (CMA-ES, Hansen and Ostermeier 2001; Hansen et al. 2003) that has robust performance in terms of search capacity. CMA-ES randomly generates an initial population (i.e., iteration). A population is composed of a number, λ , of solutions, and a solution in this context is a set of model weights with size K . Each solution in the population is evaluated in terms of its fitness f that is the objective function value (Eq. 6). The population keeps evolving to reach the optimal solution, which is the smallest f value, with a maximum of 200 iterations. Increasing the population size improves the search capacity (Elshall et al. 2015), and we use a population size of $\lambda = 100K$, where K is the number of model weights. For each ensemble, we conducted 10 repeat optimization runs with random initial solutions. For all the repeated optimization runs we obtained well-posed solutions such that no multimodality is observed, and the model weights are generally consistent. For each ensemble we select the solution with the smallest f value.

2.5 Evaluation metrics

We use several metrics to evaluate the ensemble performance. To evaluate the performance of each individual ensemble, we use metrics x_0 - x_5 . Metric x_0 is defined by Eq.3, and x_1 - x_5 by Eqs.7-11. We use AIC and BIC scores to compare different ensembles by accounting for both ensemble performance and ensemble size, which is the number of ensemble members with each ensemble member having model weight. In this context, the number of model weights for each ensemble is equivalent to the number of model parameters in an inverse modeling context and

to the number of decision variables in simulation optimization context. The AIC and BIC scores are calculated as (Akaike 1974)

$$x_{6,AIC} = 2K - 2\ln(\hat{L}) \tag{12}$$

and (Bhat and Kumar 2010)

$$x_{7,BIC} = K\ln(N) - 2\ln(\hat{L}) \tag{13}$$

respectively, where $N = 44$ is the data size, K is the number of model weights w_k, and $\ln(\hat{L})$ is the natural logarithm of the likelihood function. As such, $\ln(\hat{L})$ can be equivalent the mean-square error

$$MSE = \frac{\sum_{t=1}^{T}(h_t - h_{t,obs})^2}{T} \tag{14}$$

such that minimizing the $\ln(MSE)$ is equivalent to maximizing the $\ln(\hat{L})$ of the data (Akaike 1974). The defined metrics x_0 - x_7 are specifically designed to judge the predictive performance of these ESMs with respect to the targets of a specific application, and are not meant to judge the predictive performance of these ESMs regionally and globally for general purposes. Judging the predictive performance of these ESMs with respect to regional and global simulation of zos or any other variable, is beyond the scope of this work. Fig. 4 in provides a summary of the methods presented in this section.

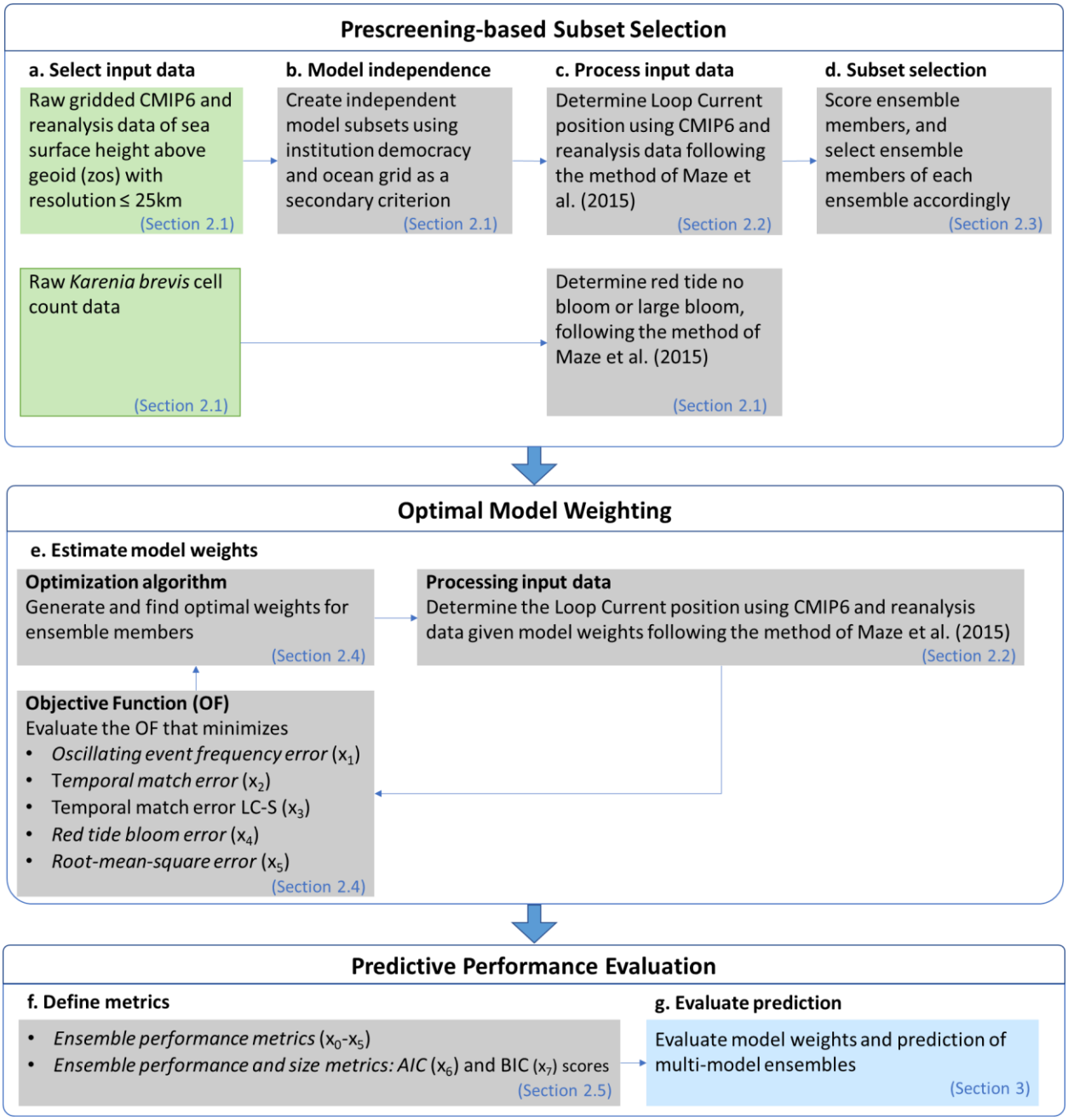

Figure 3. Method overview. See Jupyter notebooks of Elshall (2021) for details.

3 Results

3.1 Model weights

We study the impacts of model weighting given the four cases of including high- and standard-resolution model runs (WME3210), high-resolution model runs (WME321X), and high-resolution model runs given prescreening information (WME32XX and WME3XXX). Fig. 4 shows the optimal model weights of each ensemble member. Three remarks can be drawn from Fig. 4. First, for the ensemble WME3210, the IMS03 with score one and IMS05 with score zero, which overestimate LC-N and underestimate LC-N, respectively, did not receive zero model weights

despite their low scores. This might imply that model weighting optimized the error cancellation of these two members. Second, one of the best four ensemble members with score three (i.e., IMS10) receives less than 1% weight given WME3210. This is also the case for IMS07 given WME321X. This may imply that including unsuitable members in the ensemble (i.e., the standard-resolution members IMS04 and IMS05, or members not presenting LC oscillation IMS01 and IMS03) can result in flawed model weights. This again might be attributed to model weighting that optimizes the error cancellation of these members, resulting in underplaying robust models. Theses first two remarks suggest the importance of the prescreening when optimal model weighting is used. Even if subset selection is not employed (i.e., WME321X), prescreening helps to evaluate the model weighting method and results. Third, with respect to WME32XX, members with prescreening score of three generally receive higher model weights than members with prescreening score of two. This is generally desirable since these members have a better performance with respect to the application of interest. Thus, this implies that these members maintain important ensemble characteristics.

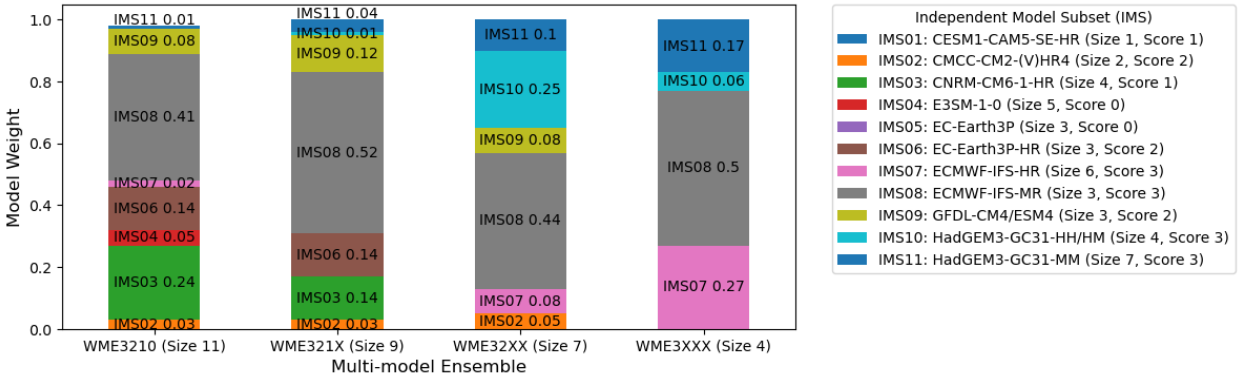

Figure 4. Model weights of ensemble members (i.e., independent model subsets) for each weighted multi-model ensemble (WME). The legend shows the number of each ensemble member, model name(s), size of the ensemble member, and score of the ensemble member. The size of the ensemble member refers to the total number of model runs per ensemble member. The size of multi-model ensemble refers to the number of ensemble members per multi-model ensemble. Ensemble members with model weights less than 1% are not shown.

3.2 Predictive performance

We evaluate the ensemble predictive performance using metrics x_0 - x_5. Table 2 presents the raw data that are used to calculate x_0 - x_5. Table 2 shows that the four weighted ensembles have relatively similar predictive performance. The ensembles have LC-S frequency x_0 of 0.227 (versus 0.273 and 0.227 for the reanalysis data and reference ensemble SME321X, respectively), which corresponds to an oscillating event count error x_2 of two. The ensembles have temporal match error x_2 of 18% expect for WME3XXX that has an error of 23%, versus 36% for the reference ensemble. Model weighting has additionally reduced the temporal match error LC-S x_3 for all the ensembles to 42% expect for WME3XXX to 50%, versus 75% for the reference ensemble. The ensembles have red tide bloom error x_4 of 7%, versus 25% for the reference ensemble. The

ensemble RMSE x_5 is generally inversely proportion to the ensemble size with the exception of WME32XX.

Table 2. Raw data of Loop Current at North (LC-N) and South (LC-S) positions and their relation to the occurrence of no bloom and large blooms shown for the reanalysis data, reference ensemble SME3210X with simple-average multi-model ensemble (SME), and four ensembles with weighted-average multi-model ensemble (WME). The corresponding performance metrics (x_0 - x_5) and fitness f value (Eq.6) for each ensemble.

	Model runs	Count		Count LC-N		Count LC-S		Temporal Match			Performance Metrics						f
		LC-N	LC-S	No-Bloom	Large-Bloom	No-Bloom	Large-Bloom	LC-N	LC-S	Total	x_0	x_1	x_2	x_3	x_4	x_5 (RMSE)	
Reanalysis	1	32	12	17	15	12	0	32	12	44	0.273	0	0.00	0.00	0.00	0	1
SME321X	33	34	10	22	12	7	3	25	3	28	0.227	2	0.36	0.75	0.25	3.71	42.1
WME3210	41	34	10	20	14	9	1	29	7	36	0.227	2	0.18	0.42	0.07	3.56	24.5
WME321X	33	34	10	20	14	9	1	29	7	36	0.227	2	0.18	0.42	0.07	3.59	24.7
WME32XX	28	34	10	20	14	9	1	29	7	36	0.227	2	0.18	0.42	0.07	3.69	25.2
WME3XXX	20	34	10	20	14	9	1	28	6	34	0.227	2	0.23	0.50	0.07	3.67	27.6

From a model weighting perspective, both predictive performance and ensemble size are evaluated. WME32XX has same predictive performance as WME321X and WME3210 given x_0 - x_4 and very similar predictive performance given x_5. WME32XX ($K = 7$) has smaller ensemble size than WME321X ($K = 9$) and WME3210 ($K = 11$). Accordingly, WME32XX is a better ensemble than WME321X and WME3210 from a model weighting perspective. For WME32XX and WME3XXX, while WME32XX has slightly better predictive performance, WME3XXX has smaller ensemble size with only four model weights ($K = 4$). To evaluate these two ensembles, we use the AIC and BIC scores (Eqs. 11 and 12) that combine the complexity of the ensemble (i.e., the number of model weights) with the performance of the ensemble into a single score. Smaller AIC and BIC scores indicate a better ensemble, given the criteria of complexity and performance. We calculate the AIC and BIC of the four ensembles. The AIC (BIC) scores are 16.92 (36.5), 12.89 (28.9), 8.78 (21.3), and 2.8 (9.9) for WME3210, WME321X, WME32XX and WME3XXX, respectively.

The estimated AIC and BIC scores are as expected, in that WME3XXX and WME3210 are the best and worst performing ensembles from a model selection perspective. In a summary, the prescreening-based subset-selection step improves the model weighting, resulting in reduction of the number of decision variables, while maintaining similar (i.e., WME32XX) or relatively similar (i.e., WME3XXX) predictive performance. Although the most parsimonious ensemble (i.e., WME3XXX) might not necessarily produce the best predictive performance, it is still favorable from a model selection perspective by balancing the ensemble performance and complexity.

The predictive performance of the simple-average and weighted-average multi-model ensembles are shown in Fig. 5. Ensembles based on prior information (e.g., SME321X) correspond better to reanalysis data than SME3210 without prior information. Similarly, ensembles based on prescreening information (i.e., SME32XX and SME3XXX) are better than the reference ensemble SME321X. Also, ensembles with model weighting have generally good correspondence with respect to reanalysis data irrespective of prior and prescreening information. Yet ensembles with

prescreening information and model weighting (i.e., WME32XX and WME3XXX) have the best correspondence with reanalysis data.

Figure 5. Temporal match of large bloom and no bloom with Loop Current positions given by (a) reanalysis data, and simulations of four multi-model ensembles with (b-e) simple-average multi-model ensemble (SME), and (f-i) weighted-average multi-model ensemble (WME). Positive and negative bars indicate Loop Current North (LC-N) and Loop Current South (LC-S), respectively.

4 Discussion

The effects of different ensemble composition criteria are summarized in Fig. 6 with respect to key metrics. Prior information appears to be an important criterion that should be considered, as SME3210 has the worst predictive performance than the other ensembles do. Subset selection seems to relatively improve the predictive performance, suggesting that it can be used to in place of model weighting, or prior to model weighting. When subset selection is used, prior to model weighting (WME3XX) or without model weighting (SME3XX), this results in the most robust ensemble from a model selection perspective. These results suggest four key points. First, while Yun et al. (2017) propose a process-based subset selection as an alternative approach to model weighting, we show that considering such process-based information can yield parsimonious

ensemble with good predictive performance. Parsimonious ensemble is favorable especially with model weighting, as several studies indicate that predictive performance improves from model diversity rather than from larger ensemble (DelSole et al., 2014; Manzanas, 2020).

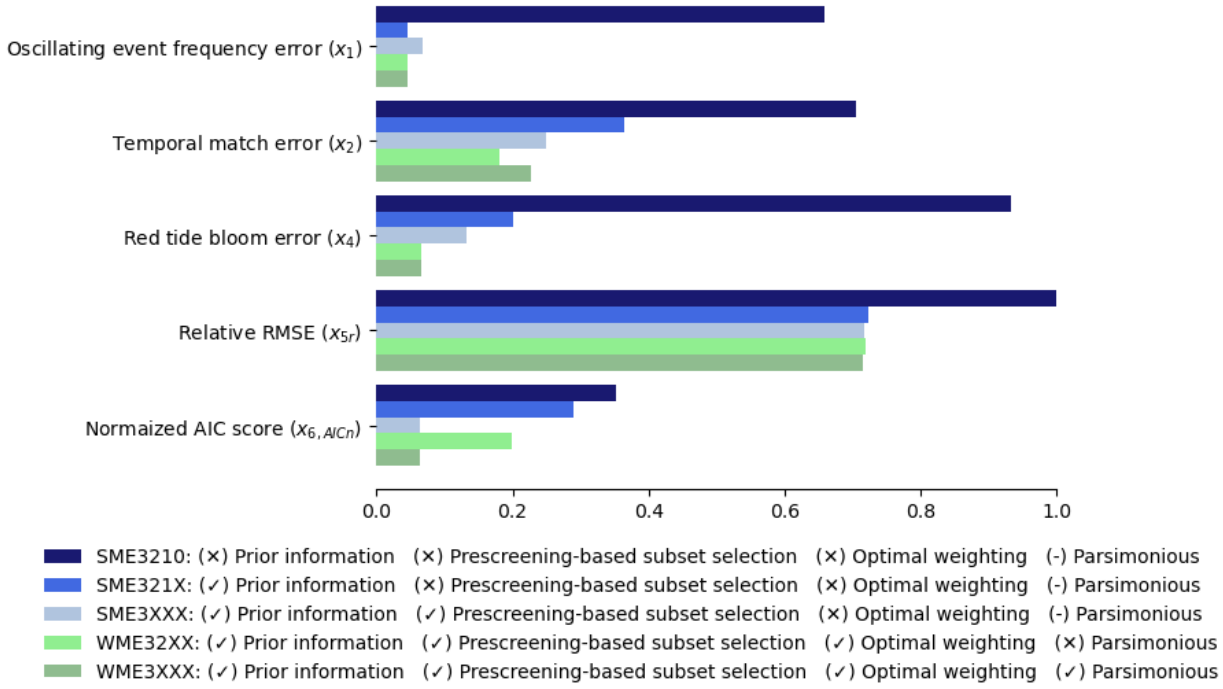

Figure 6. Predictive performance, and AIC score accounting for performance and complexity, given different ensemble composition criteria of prior information, prescreening-based subset section, optimal model weighting, and parsimony. To scale the data to the graph we calculate the relative RMSE(x_{5r}) that is the RMSE of each of ensemble divided by the maximum RMSE of the four ensembles, and the normalized AIC score ($x_{6,AICn}$) through dividing the AIC score of each ensemble ($x_{6,AIC}$) by the number of data points $N = 44$.

Second, our study reveals a caveat of optimal model weighting. Models with poor performance showing both over and underestimation can get higher model weights to maximize error cancellation, which can be undesirable. For example, Li et al. (2021) note that comparable total Antarctic Sea ice area can result from large positive bias neutralizing the strong negative bias, yet with an inaccurate physical process. Additionally, we show that optimal model weighting can further underplay robust climate models, highlighting the importance of ensemble process-based prescreening and subset selection prior to model weighing. As it has been argued that model uncertainty can be decreased by giving more weight to models that are more skillful and realistic for a specific process or application (Lorenz et al. 2018), we use binary model weights in the way similar to that of Herger et al. (2018) in which models are included or excluded.

Third, we show that subset selection alone can be an effective mean to improve the predictive performance in case model weighting is undesirable. Since giving equal weight to each available model projection can be suboptimal, advanced methods for model weighting are needed (Eyring et al. 2019). This suggests the importance of accounting for model independence such that each model run is not given an equal weight. Yet when each ensemble member is an ensemble of

dependent models, whether unequal model weighting methods improve upon equal model weighting methods is a question that needs careful consideration. For example, even if the skills of the competing methods are equal, one method will, by chance, prove superior in a given sample (DelSole et al. 2013). In addition, model weighting can underplay the critical model, which has the largest effect on the solution reliability, because its goodness-of-fit is less than other models (Elshall et al., 2020). Even worse is that the model could get a higher weight because the observational errors better coincide with the model errors (Haughton et al. 2015). In addition, while non-equal model weighting can improve the uncertainty quantifications, it will not necessarily result in improved description of mean climate states, yet will add another level of uncertainty (Christensen et al. 2010). Furthermore, the efficacy of using model weights derived on a historical reference period for future projection can be questionable. In other words, the construction of model weights using twentieth century observed data with expectation that the advantages afforded by model weighting will persist throughout the twentieth first century is hard to justify (Haughton et al. 2015). Given these and similar remarks, Weigel et al. (2010) suggest that for many applications equal model weighting may be the safer and more transparent way to combine models. Here we do not argue for equal or unequal model weighting of multi-model ensemble, but rather show that in case unequal model weighting is undesirable, prescreening-based subset selection can be used to improve the predictive performance. This remark was also indicated with respect to other methods for improving predictive performance. For example, Wang et al. (2019) show that when bias correction is applied, unequal model weighting do not bring significant differences in the multi-model ensemble mean and uncertainty of hydrological impacts. As bias-corrected climate simulations become rather close to observations, Wang et al. (2019) suggest that using bias correction and equal model weighting is viable and sufficient for their study purpose. In addition, DelSole et al. (2013) suggest that, for the forecast of temperature and precipitation, methods of unequal model weighting may be of value only over a relatively small fraction of the globe, suggesting that strategies for screening models prior to combining them would seem to be an important step. The same argument applies for subset selection given this case study, especially when model weighting does not result in significant improvement (e.g., SME3XXX and WME3XXX).

Finally, the application specific nature of the problem is an important factor to keep in mind. There is no universally best method. For example, Ross and Najjar (2019) evaluate six model selection methods with respect to performance, and the sensitivity of the results to the number of chosen model, showing that methods and models used should be carefully chosen, and obtained results should be interpreted with caution. Similarly, with respect model weighting, Herger et al. (2018) note that as in any calibration exercise, the final ensemble is sensitive to the metric, observational product, and pre-processing steps used. Likewise, with respect to accounting for model independence, Abramowitz et al. (2019) state that the sensitivity of model weighting and subset selection to metric, variable, observational estimate, location, time, and spatial scale, and calibration time period emphasizes that model dependence is application-specific, and not a general property of an ensemble. Also with respect to bias correction, the results of Hemri et al. (2020) underpin the importance of processing raw ensemble forecasts differently depending on the final forecast product needed. These remarks suggest that the application-specific nature of the problem should not be overlooked, and accordingly the

application specific ensemble methods such the two presented in this study (i.e., prescreening-base subset selection and application-specific optimal weighting) can be useful.

5 Conclusions

This study discusses the application-specific optimal model weighting of ESMs using a red tide example. Three key points can be concluded:

- First, while optimal model weighting can potentially improve predictive performance, at least one caveats need to be considered. Including non-representative models with both over and underestimation can result in error cancellation. Whether to include or exclude these non-representative models from the ensemble is a point that requires further investigation through studying model projection. However, this study clearly shows that including these non-representative models can underplay the model weights of robust models when optimal ensemble weight approach is used.
- Second, excluding all non-representative models result in the most parsimonious ensemble accounting for both ensemble size and performance.
- Third, we further show the importance of prescreening-based subset selection, which screens and select ensemble members based on their ability to reproduce certain key features. We conclude that prescreening-based subset selection is viable option that can either substitute model weighting, or be used prior to model weighing. Prescreening-based subset selection does not only help to develop a parsimonious ensemble, but also provides insights about the validity of the model weights.

These few insights provided by this study adds to the literature of application-specific optimal model weighting of ESMs. The analysis in this study is limited to historical period, and our outlook is to consider optimal model weighting with accounting for model convergence criterion given future climate projection.

Data Availability Statement

Data and codes that support the findings of this study are openly available (Elshall 2021).

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Author Contributions

MY, SK, JH, XY, and YW: Motivation, and framing for the project. MY, SK, JH, XY,YW, and AE: Conceptualization, and methodology. AE: Data curation, software, investigation, visualization, and writing the original draft. AE, MY, SK, JH, XY, YW, and MM: Validation, and manuscript

reviewing and editing. MY: Resources, supervision, and project administration. MY, SK, JH, XY, and YW: Funding acquisition.

Funding

This work is funded by NSF Award #1939994.

Acknowledgments

We thank Emily Lizotte in the Department of Earth, Ocean, and Atmospheric Science (EOAS) at Florida State University (FSU) for contacting the Florida Fish and Wildlife Conservation Commission (FWC) to obtain the *Karenia brevis* data. We thank FWC for data provision. We are grateful to Maria J. Olascoaga in the Department of Ocean Sciences at University of Miami for our communication regarding *Karenia brevis* data analysis. We thank Sally Gorrie, Emily Lizotte, Mike Stukel, and Jing Yang in EOAS at FSU for their fruitful discussion and suggestions on the project. We dedicate this paper to the memory of Stephen Kish the former professor in EOAS at FSU, who assisted with the motivation and framing for the project.

References

Abramowitz G, Herger N, Gutmann E, et al (2019) ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth System Dynamics 10:91–105. https://doi.org/10.5194/esd-10-91-2019

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723. https://doi.org/10.1109/TAC.1974.1100705

Bhat H, Kumar N (2010) On the Derivation of the Bayesian Information Criterion

Bhowmik RD, Sankarasubramanian A (2021) A performance-based multi-model combination approach to reduce uncertainty in seasonal temperature change projections. International Journal of Climatology 41:E2615–E2632. https://doi.org/10.1002/joc.6870

Brunner L, Pendergrass AG, Lehner F, et al (2020) Reduced global warming from CMIP6 projections when weighting models by performance and independence. Earth System Dynamics 11:995–1012. https://doi.org/10.5194/esd-11-995-2020

Caldwell PM, Mametjanov A, Tang Q, et al (2019) The DOE E3SM Coupled Model Version 1: Description and Results at High Resolution. Journal of Advances in Modeling Earth Systems 11:4095–4146. https://doi.org/10.1029/2019MS001870

Chang P, Zhang S, Danabasoglu G, et al (2020) An Unprecedented Set of High-Resolution Earth System Simulations for Understanding Multiscale Interactions in Climate Variability and Change. Journal of Advances in Modeling Earth Systems 12:e2020MS002298. https://doi.org/10.1029/2020MS002298

Cherchi A, Fogli PG, Lovato T, et al (2019) Global Mean Climate and Main Patterns of Variability in the CMCC-CM2 Coupled Model. Journal of Advances in Modeling Earth Systems 11:185–209. https://doi.org/10.1029/2018MS001369

559 Christensen JH, Kjellström E, Giorgi F, et al (2010) Weight assignment in regional climate models. Climate
560 Research 44:179–194. https://doi.org/10.3354/cr00916

561 DelSole T, Yang X, Tippett MK (2013) Is unequal weighting significantly better than equal weighting for
562 multi-model forecasting? Quarterly Journal of the Royal Meteorological Society 139:176–183.
563 https://doi.org/10.1002/qj.1961

564 Dixon AM, Forster PM, Beger M (2021) Coral conservation requires ecological climate-change
565 vulnerability assessments. Frontiers in Ecology and the Environment n/a:
566 https://doi.org/10.1002/fee.2312

567 Drévillon M, Régnier C, Lellouche J-M, et al (2018) QUALITY INFORMATION DOCUMENT For Global
568 Ocean Reanalysis Products GLOBAL-REANALYSIS-PHY-001-030. 48

569 Elshall AS (2021) Python and MATLAB codes for application-specific optimal model weighting of GCMs
570 with a red tide example (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5499459

571 Elshall AS, Pham HV, Tsai FT-C, et al (2015) Parallel Inverse Modeling and Uncertainty Quantification for
572 Computationally Demanding Groundwater-Flow Models Using Covariance Matrix Adaptation.
573 Journal of Hydrologic Engineering 20:04014087. https://doi.org/10.1061/(ASCE)HE.1943-
574 5584.0001126

575 Eyring V, Bony S, Meehl GA, et al (2016) Overview of the Coupled Model Intercomparison Project Phase
576 6 (CMIP6) experimental design and organization. Geoscientific Model Development 9:1937–
577 1958. https://doi.org/10.5194/gmd-9-1937-2016

578 Eyring V, Cox PM, Flato GM, et al (2019) Taking climate model evaluation to the next level. Nature Clim
579 Change 9:102–110. https://doi.org/10.1038/s41558-018-0355-y

580 Fernandez E, Lellouche JM (2018) PRODUCT USER MANUAL For the Global Ocean Physical Reanalysis
581 product GLOBAL_REANALYSIS_ PHY_001_030. 15

582 FWRI (2020) HAB Monitoring Database. In: Florida Fish And Wildlife Conservation Commission.
583 http://myfwc.com/research/redtide/monitoring/database/. Accessed 23 Dec 2020

584 Golaz J-C, Caldwell PM, Roekel LPV, et al (2019) The DOE E3SM Coupled Model Version 1: Overview and
585 Evaluation at Standard Resolution. Journal of Advances in Modeling Earth Systems 11:2089–
586 2129. https://doi.org/10.1029/2018MS001603

587 Haarsma RJ, Roberts MJ, Vidale PL, et al (2016) High Resolution Model Intercomparison Project
588 (HighResMIP v1.0) for CMIP6. Geoscientific Model Development 9:4185–4208.
589 https://doi.org/10.5194/gmd-9-4185-2016

590 Hansen N, Müller SD, Koumoutsakos P (2003) Reducing the time complexity of the derandomized
591 evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary Computation 11:1–
592 18. https://doi.org/10.1162/106365603321828970

593 Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies.
594 Evolutionary computation 9:159–195

595 Haughton N, Abramowitz G, Pitman A, Phipps SJ (2015) Weighting climate model ensembles for mean
596 and variance estimates. Clim Dyn 45:3169–3181. https://doi.org/10.1007/s00382-015-2531-3

597 Held IM, Guo H, Adcroft A, et al (2019) Structure and Performance of GFDL's CM4.0 Climate Model.
598 Journal of Advances in Modeling Earth Systems 11:3691–3727.
599 https://doi.org/10.1029/2019MS001829

600 Hemri S, Bhend J, Liniger MA, et al (2020) How to create an operational multi-model of seasonal
601 forecasts? Clim Dyn 55:1141–1157. https://doi.org/10.1007/s00382-020-05314-2

602 Herger N, Abramowitz G, Knutti R, et al (2018) Selecting a climate model subset to optimise key
603 ensemble properties. Earth System Dynamics 9:135–151. https://doi.org/10.5194/esd-9-135-
604 2018

605 Herger N, Abramowitz G, Sherwood S, et al (2019) Ensemble optimisation, multiple constraints and
606 overconfidence: a case study with future Australian precipitation change. Clim Dyn 53:1581–
607 1596. https://doi.org/10.1007/s00382-019-04690-8

608 Hoch KE, Petersen MR, Brus SR, et al (2020) MPAS-Ocean Simulation Quality for Variable-Resolution
609 North American Coastal Meshes. Journal of Advances in Modeling Earth Systems
610 12:e2019MS001848. https://doi.org/10.1029/2019MS001848

611 Jacox MG, Alexander MA, Siedlecki S, et al (2020) Seasonal-to-interannual prediction of North American
612 coastal marine ecosystems: Forecast methods, mechanisms of predictability, and priority
613 developments. Progress in Oceanography 183:102307.
614 https://doi.org/10.1016/j.pocean.2020.102307

615 Knutti R, Sedláček J, Sanderson BM, et al (2017) A climate model projection weighting scheme
616 accounting for performance and interdependence. Geophysical Research Letters 44:1909–1918.
617 https://doi.org/10.1002/2016GL072012

618 Leduc M, Laprise R, Elía R de, Šeparović L (2016) Is Institutional Democracy a Good Proxy for Model
619 Independence? Journal of Climate 29:8301–8316. https://doi.org/10.1175/JCLI-D-15-0761.1

620 Li S, Huang G, Li X, et al (2021) An Assessment of the Antarctic Sea Ice Mass Budget Simulation in CMIP6
621 Historical Experiment. Front Earth Sci 9:649743. https://doi.org/10.3389/feart.2021.649743

622 Liu Y, Weisberg RH, Lenes JM, et al (2016) Offshore forcing on the "pressure point" of the West Florida
623 Shelf: Anomalous upwelling and its influence on harmful algal blooms. Journal of Geophysical
624 Research: Oceans 121:5501–5515. https://doi.org/10.1002/2016JC011938

625 Lorenz R, Herger N, Sedláček J, et al (2018) Prospects and Caveats of Weighting Climate Models for
626 Summer Maximum Temperature Projections Over North America. Journal of Geophysical
627 Research: Atmospheres 123:4509–4526. https://doi.org/10.1029/2017JD027992

628 Magaña HA, Villareal TA (2006) The effect of environmental factors on the growth rate of Karenia brevis
629 (Davis) G. Hansen and Moestrup. Harmful Algae 5:192–198.
630 https://doi.org/10.1016/j.hal.2005.07.003

631 Maze G, Olascoaga MJ, Brand L (2015) Historical analysis of environmental conditions during Florida Red
632 Tide. Harmful Algae 50:1–7. https://doi.org/10.1016/j.hal.2015.10.003

633 Merrifield AL, Brunner L, Lorenz R, et al (2020) An investigation of weighting schemes suitable for
634 incorporating large ensembles into multi-model ensembles. Earth System Dynamics 11:807–834.
635 https://doi.org/10.5194/esd-11-807-2020

636 Oh S-G, Suh M-S (2017) Comparison of projection skills of deterministic ensemble methods using
637 pseudo-simulation data generated from multivariate Gaussian distribution. Theor Appl Climatol
638 129:243–262. https://doi.org/10.1007/s00704-016-1782-1

639 Payne MR, Hobday AJ, MacKenzie BR, Tommasi D (2019) Editorial: Seasonal-to-Decadal Prediction of
640 Marine Ecosystems: Opportunities, Approaches, and Applications. Front Mar Sci 6:.
641 https://doi.org/10.3389/fmars.2019.00100

642 Perkins S (2019) Inner Workings: Ramping up the fight against Florida's red tides. PNAS 116:6510–6512.
643 https://doi.org/10.1073/pnas.1902219116

644 Räisänen J, Ylhäisi JS (2012) Can model weighting improve probabilistic projections of climate change?
645 Clim Dyn 39:1981–1998. https://doi.org/10.1007/s00382-011-1217-8

646 Roberts CD, Senan R, Molteni F, et al (2018) Climate model configurations of the ECMWF Integrated
647 Forecasting System (ECMWF-IFS cycle 43r1) for HighResMIP. Geoscientific Model Development
648 11:3681–3712. https://doi.org/10.5194/gmd-11-3681-2018

649 Roberts MJ, Baker A, Blockley EW, et al (2019) Description of the resolution hierarchy of the global
650 coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments. Geoscientific
651 Model Development 12:4999–5028. https://doi.org/10.5194/gmd-12-4999-2019

652 Ross AC, Najjar RG (2019) Evaluation of methods for selecting climate models to simulate future
653 hydrological change. Climatic Change 157:407–428. https://doi.org/10.1007/s10584-019-02512-
654 8

655 Sanderson BM, Wehner M, Knutti R (2017) Skill and independence weighting for multi-model
656 assessments. Geoscientific Model Development 10:2379–2395. https://doi.org/10.5194/gmd-
657 10-2379-2017

658 Tonelli M, Signori CN, Bendia A, et al (2021) Climate Projections for the Southern Ocean Reveal Impacts
659 in the Marine Microbial Communities Following Increases in Sea Surface Temperature. Front
660 Mar Sci 8:636226. https://doi.org/10.3389/fmars.2021.636226

661 van den Hurk B, Hewitt C, Jacob D, et al (2018) The match between climate services demands and Earth
662 System Models supplies. Climate Services 12:59–63.
663 https://doi.org/10.1016/j.cliser.2018.11.002

664 Voldoire A, Saint-Martin D, Sénési S, et al (2019) Evaluation of CMIP6 DECK Experiments With CNRM-
665 CM6-1. Journal of Advances in Modeling Earth Systems 11:2177–2213.
666 https://doi.org/10.1029/2019MS001683

Wang H-M, Chen J, Xu C-Y, et al (2019) Does the weighting of climate simulations result in a better quantification of hydrological impacts? Hydrology and Earth System Sciences 23:4033–4050. https://doi.org/10.5194/hess-23-4033-2019

Ward ND, Megonigal JP, Bond-Lamberty B, et al (2020) Representing the function and sensitivity of coastal interfaces in Earth system models. Nat Commun 11:2458. https://doi.org/10.1038/s41467-020-16236-2

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of Model Weighting in Multimodel Climate Projections. Journal of Climate 23:4175–4191. https://doi.org/10.1175/2010JCLI3594.1

Weisberg RH, Liu Y, Lembke C, et al (2019) The Coastal Ocean Circulation Influence on the 2018 West Florida Shelf K. brevis Red Tide Bloom. Journal of Geophysical Research: Oceans 124:2501–2512. https://doi.org/10.1029/2018JC014887

Weisberg RH, Zheng L, Liu Y, et al (2014) Why no red tide was observed on the West Florida Continental Shelf in 2010. Harmful Algae 38:119–126. https://doi.org/10.1016/j.hal.2014.04.010

White CJ, Carlsen H, Robertson AW, et al (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol Appl 24:315–325. https://doi.org/10.1002/met.1654

Xu D, Ivanov VY, Kim J, Fatichi S (2019) On the use of observations in assessment of multi-model climate ensemble. Stoch Environ Res Risk Assess 33:1923–1937. https://doi.org/10.1007/s00477-018-1621-2

Zhang X, Yan X (2018) Criteria to evaluate the validity of multi-model ensemble methods. International Journal of Climatology 38:3432–3438. https://doi.org/10.1002/joc.5486

Declaration of competing interest

All authors declare no competing interest.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Funding

This work is funded by NSF Award #1939994.