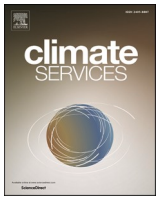
Climate Services 28 (2022) 100334

Contents lists available at ScienceDirect

Climate Services

journal homepage: www.elsevier.com/locate/cliser



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

Ahmed Elshall [a,b,c], Ming Ye [a,*], Sven A. Kranz [a], Julie Harrington [d], Xiaojuan Yang [e], Yongshan Wan [f], Mathew Maltrud [g]

[a] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA
[b] Department of Bioengineering, Civil Engineering and Environmental Engineering, U. A. Whitaker College of Engineering, Florida Gulf Coast University, Fort Myers, FL, USA
[c] The Water School, Florida Gulf Coast University, Fort Myers, FL, USA
[d] Center for Economic Forecasting and Analysis, Florida State University, Tallahassee, FL, USA
[e] Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[f] Center for Environmental Measurement and Modeling, United States Environmental Protection Agency, Gulf Breeze, FL, USA
[g] Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, USA

A B S T R A C T

Global climate models (GCMs) and Earth system models (ESMs) provide many climate services with environmental relevance. The High Resolution Model Intercomparison Project (HighResMIP) of the Coupled Model Intercomparison Project Phase 6 (CMIP6) provides model runs of GCMs and ESMs to address regional phenomena. Developing a parsimonious ensemble of CMIP6 requires multiple ensemble methods such as independent-model subset selection, prescreening-based subset selection, and model weighting. The work presented here 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 a common name of specific harmful algal blooms that occur worldwide, causing adverse socioeconomic and environmental impacts. Our results indicate the importance of prescreening-based subset selection as optimal model weighting can underplay robust ensemble members by optimizing error cancellation. Prescreening-based subset selection also provides 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.

Practical Implications

Coastal areas are frequently threatened by natural and human hazards such as massive harmful algae blooms (HABs). Red tides are a natural phenomenon caused by blooms (dense aggregations) of harmful microscopic algae in coastal areas worldwide. These events are influenced by a multitude of factors including oceanic, atmospheric, and land/river-based events. Here we use the term red tides for occurrences of large amounts of the toxic dinoflagellate *Karenia brevis*. Red tide events contribute generally to water quality degradation, and in the Gulf of Mexico these events have severe environmental and socioeconomic impacts on the State of Florida, USA. Earth system models (ESMs) present a unique opportunity for the regional environmental management of red tides as ESMs couple land, river, ocean, and atmospheric processes.

Projection of future trends of red tides is important to environmental management for planning and evaluating the short-term and long-term impacts and risks of red tides on the ecosystem health, social justice, and regional economy. The overarching goal of this research is to predict future trends of red tides under different Shared Socioeconomic Pathways (SSPs) of the Coupled Model Intercomparison Project Phase Six (CMIP6), which are scenarios of projected socioeconomic global changes up to year 2100 (with emission scenarios). These future projections of ESMs under SSPs scenarios can be used as data input for machine learning to predict long term trends in the occurrence of red tides (Elshall et al., 2021). This requires not only validating ESMs simulations with observational and reanalysis data to account for errors, but also using ensemble methods such as optimal model weighting to improve the predictive performance. The manuscript addresses an important topic in climate services that is regional

* Corresponding author.
E-mail addresses: aelshall@fgcu.edu (A. Elshall), mye@fsu.edu (M. Ye).

https://doi.org/10.1016/j.cliser.2022.100334
Received 9 January 2022; Received in revised form 8 August 2022; Accepted 12 October 2022
Available online 20 October 2022

and decision-relevant metrics in optimal model weighting. Our research method can be used to identify non-representative models, understand their impacts on ensemble prediction, and improve ensemble prediction. This is important for more accurate projection of red tides and corresponding socioeconomic impacts and mitigation efforts under different climate scenarios.

Data availability

The data and codes used are publically available as cited in the manuscript.

1. Introduction

The High-Resolution Model Intercomparison Project (HighResMIP, Haarsma et al., 2016) of the Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al., 2016) presents a new generation of high-resolution Earth system models (ESMs) with fine resolution and improved process representation focusing on regional phenomena. While global climate models (GCMs) mainly represent the physical atmospheric and oceanic processes, ESMs advance beyond GCMs by explicitly accounting for the interactions of the biogeochemical processes with the physical climate, and by simulating the interactions between the atmosphere, biosphere, cryosphere, geosphere, and hydrosphere. As ESMs account for atmospheric chemistry, ocean ecology and biogeochemistry, plant ecology, and land use, these models can provide many services at regional and seasonal scales that are important for a wide range of stakeholders. Hereafter, the term ESMs refer to both ESMs and GCMs for the convenience of discussion. Predictions of ESMs at the regional scale are useful for resource management and decision making in many sectors such as agriculture (Ceglar et al., 2018; Vajda and Hyvärinen, 2020), water resources (Mishra et al., 2019; Zhao et al., 2020), energy (Bett et al., 2017; De Felice et al., 2019; Lledo et al., 2019), health (Lowe et al., 2017), ecological and environmental management (Payne et al., 2019; Jacox et al., 2020; Dixon et al., 2021), coastal management (Ward et al., 2020), financial services (Fiedler et al., 2021), among many other applications as reviewed by White et al. (2017). While ESMs are key ingredients of many of these climate services, tailoring model results to real-world applications is a major challenge (van den Hurk et al., 2018). Focusing on improving predictive performance of ESMs using ensemble methods, we present a case study of red tides using the medium- and high-resolution ESMs of CMIP6.

Red tide is a common name of harmful algae blooms that occur worldwide, and is caused by toxic dinoflagellates such as *Karenia brevis*. Red tides contribute to water quality degradation worldwide, resulting in many undesirable effects. For example, the occurrence of red tides in the Gulf of Mexico has severe environmental and socioeconomic impacts on the State of Florida, USA. These impacts affect fishery (e.g., massive fish kills and shellfish poisoning), ecosystem health and services (e.g., harming birds, marine mammals, and sea turtles), local community and tourism industry (e.g., unpleasant odor and scenery), public health (e.g., skin, eye, and respiratory irritation), and other sectors as reviewed by Zohdi and Abbaspour (2019). The initiation, growth, maintenance, and termination stages of red tides in the Gulf of Mexico have many driving factors including regional warm ocean currents, local and deep-ocean upwelling, river flow, sediment transport, submarine groundwater discharge, nutrients from multiple sources (e.g., river, groundwater, ocean, atmospheric deposition and biology), African Sahara dust, tropical cyclones, and wind-direction (Brand and Compton, 2007; Heil et al., 2014; Weisberg et al., 2014; Maze et al., 2015). An example an important physical driver that controls the occurrence of red tides is the the Loop Current, which is a warm ocean current that penetrates through the Gulf of Mexico (Weisberg et al., 2014; Maze et al., 2015; Perkins, 2019). Maze et al. (2015) show that the Loop Current sets necessary condition for a large red tide blooms to occur, and point out that the Loop Current can be "the first definitive predictor of bloom possibility". The development of management models such as machine learning models for regional environmental management of red tides using global climate models (Elshall et al., 2021) requires the validation of ESMs simulations with observational and reanalysis data to account for errors. The development also requires the use of ensemble methods to improve model predictive performance. These are important for more accurate projections of red tides and corresponding socioeconomic impacts and mitigation efforts under different climate scenarios. Using the Loop Current for red tide bloom prediction as a case study, 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). Advanced methods for model weighting are 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 method of model weighting, 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 tailored model weighting methods are needed depending on a set of model evaluation criteria.

Model weighting can be based on a single or combination of model evaluation criteria. Pioneering work on model weighting (Doblas-Reyes et al., 2005; Raftery et al., 2005; Tebaldi et al., 2005; Tebaldi and Knutti, 2007) gave impetus for subsequent work on 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 observations, 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 coefficients, standard deviations, and centered RMSE), and multivariate linear regression that minimizes the RMSE of the ensemble prediction using least squares regression methods. Multi-criteria-based model weighting methods extend beyond the model performance criterion to assign model weights. In addition to model performance, model independence and convergence are two additional criteria. 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 a reliability ensemble averaging method that considers both model bias to historical observation (performance criterion) and model similarity to other models in future projections (convergence criterion). A fourth criterion for assigning model weights is inter-model comparison for observable climate and future climate (Räisänen and Ylhäisi, 2012). For this, 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 rather than generic

model assessment of ESMs that may be irrespective of the application. 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, the 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 method for application-specific optimal model weighting has several practical advantages. First, the flexibility in ensemble calibration by defining an adjustable objective function allows this method to be applicable 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 simultaneously account for multiple objectives such as multiple variables of precipitation, sea surface temperature, and wind (Herger et al., 2019), and for 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. In other words, assigning model weights to the outputs of ESMs based on their ability to represent the climate variable of interest (e.g., Loop Current) is more straightforward than accounting for other decision relevant metrics (e.g., occurrence or non-occurrence of large red tide blooms), yet accounting for both can be desirable. In the remainder of the manuscript, Section 2 presents the application-specific optimal model weighting method for the red tide case study. This is followed by the presentation of the model weights and predictive performance results (Section 3). We discuss in Section 4 the advantages and disadvantages of model weighting, and conclude by summarizing our main findings and providing a research outlook.

2. Method

2.1. Data

We select all the model runs of the Coupled Model Intercomparison Project Phase 6 (CMIP6) for both the historical experiment (Eyring et al., 2016) and the hist-1950 experiment (Haarsma et al., 2016) of the HighResMIP with gridded monthly sea surface height above geoid and nominal resolution less than or equal 25 km. This resulted in a total of 33 model runs (Table 1). The sea surface height above geoid is called zos according to the climate and forecast metadata conventions. The historical experiment and the hist-1950 experiment are from years Jan-1850 and Jan-1950, respectively, to Dec-2014. For analysis purposes, we also consider model runs with the standard resolution. These are E3SM-1–0 with variable ocean resolution of 30–60 km, and EC-Earth3P with nominal ocean resolution of about 100 km (Table 1).

Model independence was accounted for by using institutional democracy (Leduc et al., 2016) and ocean grid resolution as a secondary

Table 1
Independent model subsets based on institutional democracy with the 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

criterion. Institutional democracy is only the first step for defining model independence, and additional practical and theoretical considerations can be employed as needed (Leduc et al., 2016; Annan and Hargreaves, 2017; Boé, 2018). While it is reasonable to assume that members of the same model that differ in resolution are dependent (Boé, 2018; Lorenz et al., 2018; Merrifield et al., 2020; Brunner et al., 2020), determining where to draw the line between independent and dependent models is difficult (Merrifield et al., 2020). For example, when considering a temperature variable, Leduc et al. (2016) showed that higher model resolution can result in independent models at certain geographical regions. In our case-study about red tides, ocean grid resolution can be critical for the processes of interest. It has been shown that the Loop Current cannot be simulated appropriately by E3SM with the standard resolution (Golaz et al., 2019) that has ocean and sea ice grid resolution of 60 km in the mid-latitudes and 30 km at the equator and poles (Caldwell et al., 2019). However, when considering a higher ocean grid resolution that can better resolve mesoscale eddies (Caldwell et al., 2019; Hoch et al., 2020) the Loop Current is better represented. Thus, institutional democracy alone is insufficient, and we need to account for ocean grid resolution as a secondary criterion for defining model independence as ocean resolution affects the regional phenomena of interest (Elshall, 2020).

Accordingly, for the same institution, we create further subsets given different grid resolutions, resulting in 11 ensemble members (Table 1). Each independent model subset (IMS) constitutes an ensemble member. Each IMS contains one model run or several model runs with different indices of realizations (r), initializations (i), physics (p), and forcings (f). We assume each ensemble member to be an IMS. For example, IMS01 has only one model run "r1i1p1f1", and IMS11 has seven model runs, three with initialization r1i(1–3)p1f1 and four with realizations r(1–4) i1p1f3 as listed in Table 1. Generally, the choices with respect to model independence criteria are partly subjective and likely not perfect, such that with an in-depth knowledge of the differences between codes, different choices would be possible (Boé, 2018). The Supplementary Material contains additional information about the model independence assumptions used in this study. However, comparing different methods to account for model independence is beyond the scope of this work. 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).

We use the *Karenia brevis* cell count of the harmful algal bloom database of the Fish and Wildlife Research Institute at the Florida Fish

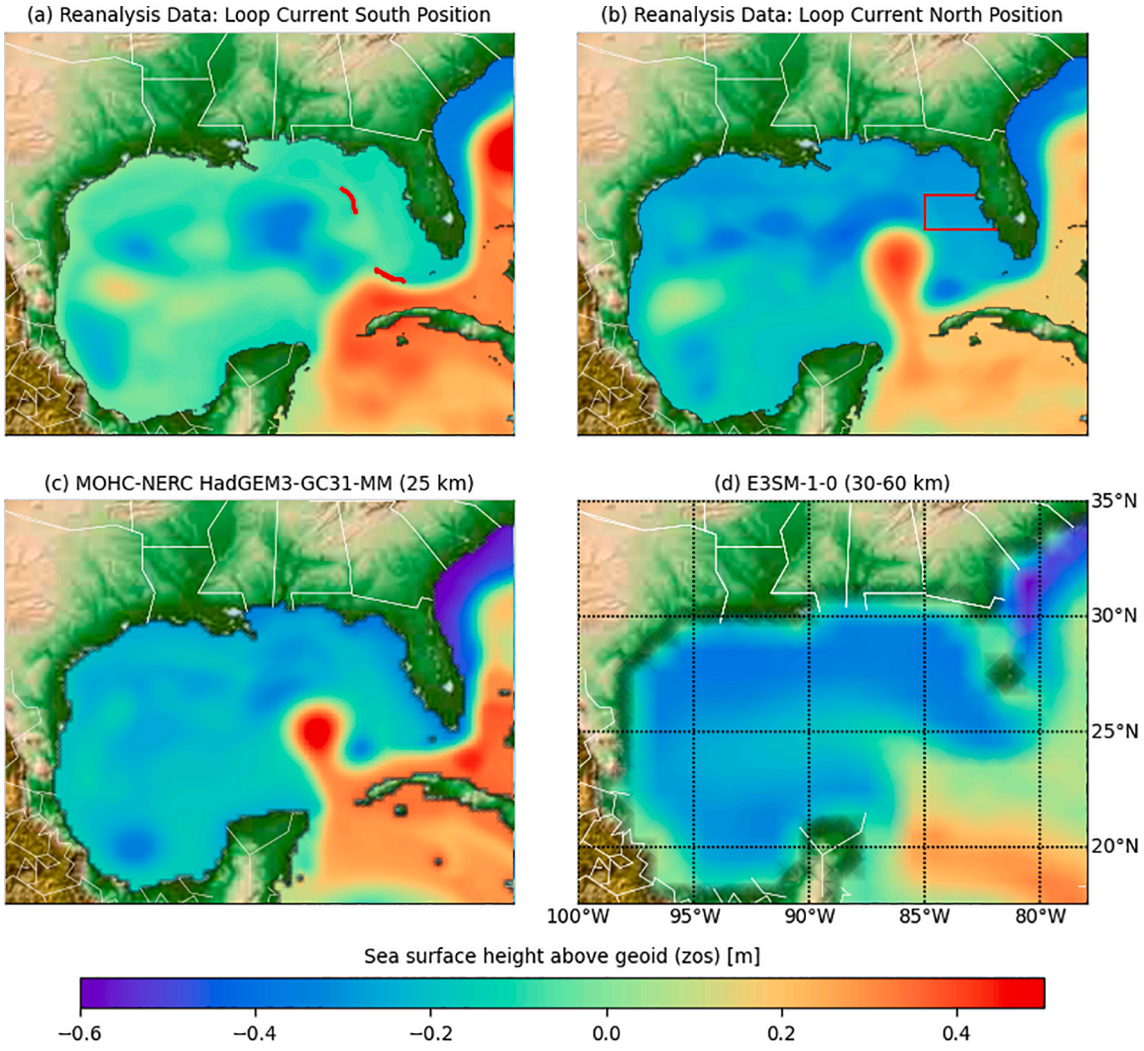


Fig. 1. Snapshots of sea surface height above geoid (zos) showing (a) the Loop Current (LC) in the south position (LS-C) in 2010–03 for reanalysis data, and the LC in the north position (LC-N) in 2010–06 for (b) reanalysis data, (c) a high-resolution ESM, and (d) a standard-resolution ESM. Two red segments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

and the Wildlife Conservation Commission (FWRI, 2020). We define the bloom severity of *Karenia brevis* according to the definition implemented by Maze et al. (2015). A large red tide 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 for years Jan-1993 to Dec-2014 with a six-month interval (i.e., a total of 44 intervals), we identify 15 intervals of large blooms and 29 intervals with no bloom in the study area (Fig. 1).

2.2. The Loop Current position and presence of red tide blooms

The Loop Current (LC) is a warm ocean current that travels through the Gulf of Mexico. The LC is an important factor that controls the occurrence of red tides (Perkins, 2019) by altering the upwelling intensity and position of deep ocean water along the West Florida Shelf (Weisberg et al., 2014) and the retention time of the waters within the Guld of Mexico(Maze et al., 2015). In this study we focus on the retention time. Other relations (Weisberg et al., 2014; 2019; Liu et al., 2016) are warranted in future studies. *Karenia brevis* is a slow growing dinoflagellate. To form large blooms, it is required that its retention within a certain region is high. Specifically, the growth rate of this species needs to be higher than the rate of advection out of the region (Magaña and Villareal, 2006). When the LC is in the southern position (LC-S), as shown in Fig. 1a, the retention time does not allow large blooms to occur. Whereas, when the LC is in the northern position (LC-N), as shown in Fig. 1b, the retention time is enhanced, allowing large red tide blooms to form when other conditions are ideal. LC-N is a necessary condition for large red tide blooms to occur (Maze et al., 2015).

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

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

In this equation, we first take the expectation $E_j(.)$ for all model runs with index j in each ensemble memberM_k, and then the $E_j(.)$ data are averaged for all ensemble members with index $k \in [1, K]$ where w_k is the weight of each ensemble memberM_k. Subsequently, the expectation $E_l(.)$ 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 $\max_{h_n}(.)$ is selected resulting in zos anomaly per interval$t \in [1, T]$, with $T = 44$, for the study period for years 1993–2014 and a 6-month interval length. Since we are not interested in the value of h_t per se but the sign difference between the north and south segments, we express Equation (1) as an indicator function for the south segment (LC-S):

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

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

$$x_0 = \frac{\sum_{t=1}^{T} H_{LC-S}(h_t)}{T} \quad (3)$$

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 surprising, because the study period and the reanalysis data products

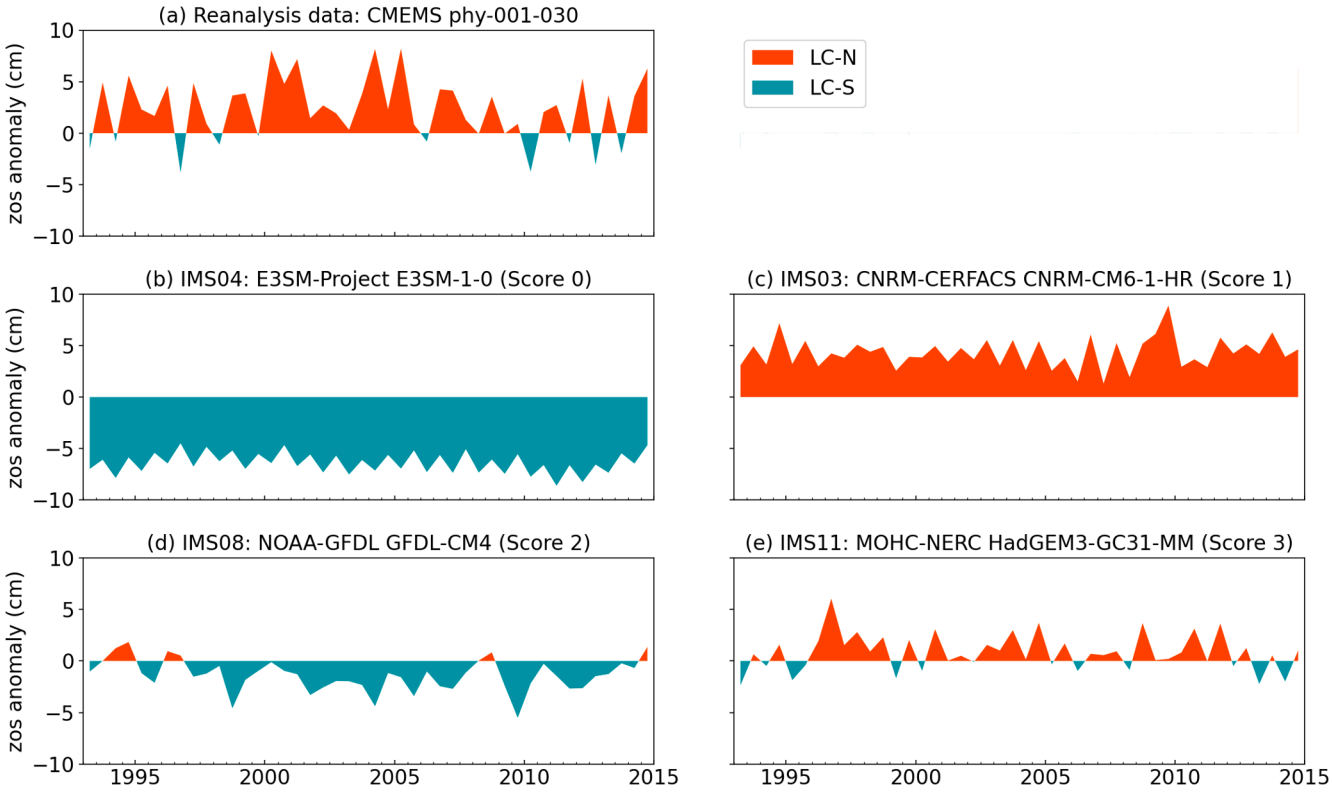


Fig. 2. Surface height above geoid (zos) anomaly according to Equation (1): (a) reanalysis data; (b-e) enesmble members. The title of the reanalysis data shows the data provider name, and product ID. The title of ensemble member shows ensemble member number that is the number of each independent model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

used in this study are different from those 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 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 as shown by Equation (1). We consider the following two ensemble methods namely a 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 (Elshall et al., 2022). These criteria are evolving such that each ensemble member receives a score from zero to three. Ensemble members that cannot simulate LC-N, as that shown in Fig. 1d for example, receive a score zero (e.g., Fig. 2b). Ensemble members that can simulate LC-N, but without sign fluctuations to indicate both the LC-N and LC-S according to Equation (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 Equation (1) with the frequency of LC-N being smaller than that of LC-S (e.g., Fig. 2d). The ensemble member receives a score of three, if it can reproduce both LC-N and LC-S according to Equation (1) with the frequency of LC-N being greater than that of LC-S (e.g., Fig. 2e). Higher LC-N frequency is a more realistic condition, considering that $x_{0,obs} = 0.273$ for reanalysis data (Fig. 2a). The score is calculated for each ensemble. For example, IMS4 and IMS5 have a score of zero because they cannot simulate LC-N (their simulation results are similar to Fig. 2b), while IMS7, IMS8, IMS10 and IMS11 have a score of three because they can simulate both LC-N and LC-S with higher LC-N frequency (their simulation results are similar to Fig. 2e).

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, WME3210 with $K = 11$ (K being the number of model weights of each ensemble such that each IMS has one weight) includes all ensemble members with a score from three to zero; WME321X with $K = 9$ includes all ensemble members with a score from three to one; WME32XX with $K = 7$ includes all ensemble members with a score from three to two; and WME3XXX with $K = 4$ includes only the top performing ensemble members with only a score of three. For example, WME3XXX has four ensemble members with a total of 20 members such that IMS07, IMS08, IMS10, and IMS11 have 6,3, 4, and 7 members, respectively. We know from previous studies (Caldwell et al., 2019; Hoch et al., 2020) that, unlike the high resolution eddy-permitting grids (e.g., Fig. 1c), 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. 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 impacts 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. This leads to ensemble SME3210, SME321X, SME32XX and SME3XXX for 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 Equation (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 Equation (1) that minimizes 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 each having an objective-weighting constantc_i. We constrain 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 resulting in $x_i = 0$. Accounting for multiple objectives can be achieved through Pareto-optimal solutions (Herger et al., 2019) or objective-weighting constantsc_i. 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:

$$x_1 = \left| \sum_{t=1}^{T} H_{LC-S}(h_t) - \sum_{t=1}^{T} H_{LC-S}(h_{t,obs}) \right| \tag{7}$$

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

$$x_2 = \frac{T - \sum_{T=1}^{T}(h_{t,obs} < 0 \wedge h_t < 0) - \sum_{t=1}^{T}(h_{t,obs} \geqslant 0 \wedge h_t \geqslant 0)}{T} \tag{8}$$

where $\sum_{t=1}^{T}(h_{t,obs} \geqslant 0 \wedge h_t \geqslant 0)$ and $\sum_{t=1}^{T}(h_{t,obs} < 0 \wedge h_t < 0)$ 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 $(h_{t,obs} < 0 \wedge h_t < 0)$ is true if both $h_{t,obs} < 0$ and $h_t < 0$ are true, otherwise, a value of zero is given. The simulations of HighResMIP are generally free-running, and thus no temporal match is expected between simulations of ESMs and re-analysis data. The term temporal match used in this manuscript refers to a pseudo-temporal correspondence that captures the general pattern of a dynamic process of the LC position given the heuristic relation (Equation (1)) with a coarse-temporal-resolution of six months. The third objective x_3 is the LC-S temporal match error:

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

The fourth objective x_4 is the red tide bloom error occurrence:

$$x_4 = \frac{\sum_{t=1}^{T}(h_t < 0 \wedge H(z_t) = 1)}{T_{bloom}} \tag{10}$$

which represents the false negative prediction of red tide blooms. This is the ratio of the number of LC-S coinciding with large bloom to the number T_{bloom} of large blooms, such that $H(z_t)$ 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:

$$x_5 = \sqrt{\frac{\sum_{n=1}^{N}(h_t - h_{t,obs})^2}{T}} \tag{11}$$

With respect to objective-weighting constantsc_i, we set $c_i = 1$, assuming that all objectives are of equal importance.

A common practice to solve Equation (6) subject to Equations 7–11 is to use an optimization algorithm such as a 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 (Equation (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. 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 in Equation (6). The population keeps evolving to reach the optimal solution, which is the smallest f value, with a user-specified maximum iterations (i.e., 200 in this study). 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 conduct 10 repeat optimization runs with random initial solutions. For all the repeated optimization runs we obtain 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

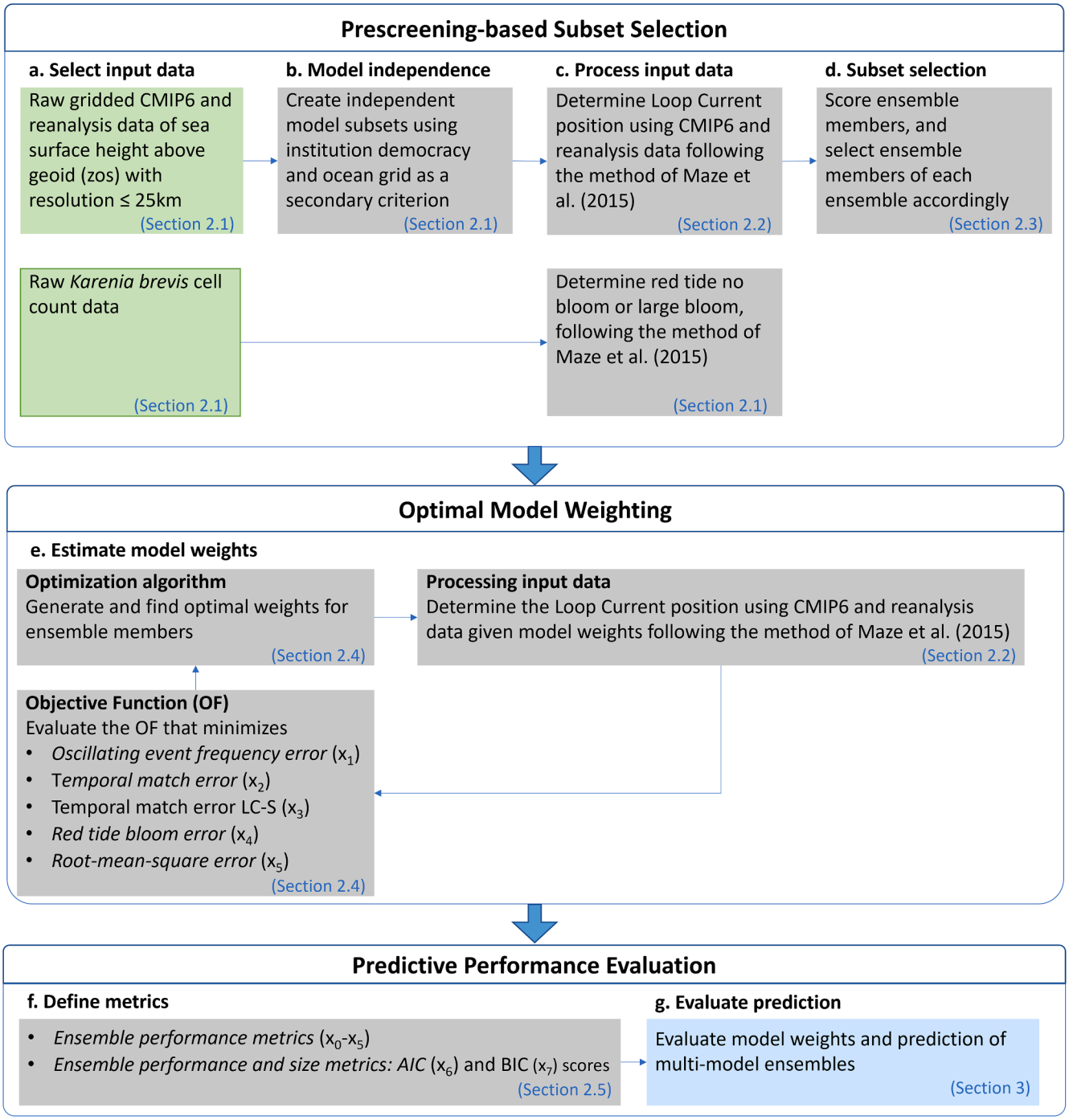


Fig. 3. Method overview. More details of the method are referred to the Jupyter notebooks of Elshall (2021).

x_0-x_5. Metric x_0 is defined by Equation (3), and x_1-x_5 by Equations 7–11. We use Akaike Information criterion (AIC) and Bayesian Information Criterion (BIC) to compare different ensembles by accounting for both ensemble performance and ensemble size. The ensemble size is the number of model weights of each ensemble that are 11, 9, 7 and 4 for WME3210, WME321X, WME32XX, and WME3XXX, respectively. 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):

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

and (Bhat and Kumar, 2010):

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

respectively, where $N = 44$ is the data size, K is the number of model weightsw_k, and $\ln(\widehat{L})$ is the natural logarithm of the likelihood function. $\ln(\widehat{L})$ can be equivalent to the mean-square error (MSE):

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

such that minimizing ln(MSE) is equivalent to maximizing $\ln(\widehat{L})$ of the data (Akaike, 1974). AIC and BIC combine the complexity of the ensemble (i.e., the number of model weights) and the performance of the ensemble into a single score. Smaller AIC and BIC scores indicate a better ensemble. 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 evaluate the predictive performance of these ESMs regionally and globally for general purposes. Assessing the predictive performance of these ESMs with respect to regional and global simulations of zos or any other variable, is beyond the scope of this work. Fig. 3 provides a summary of the methods presented in this section.

3. Results

3.1. Model weights

We investigated the impacts of model weighting given four cases of high- and standard-resolution model runs (WME3210), high-resolution model runs (WME321X), and high-resolution model runs with prescreening information (WME32XX and WME3XXX). Fig. 4 shows the optimal model weights of each ensemble member. Three remarks can be drawn from Fig. 4. 1) For ensemble WME3210, IMS03 has a score of one due to overestimating LC-N, and IMS05 has a score of zero due to underestimating LC-N, respectively. However, they did not receive zero model weights despite their low scores. This might imply that model weighting optimizes the error cancellation of the two members. 2) One of the best four ensemble members with a score of three (i.e., IMS10 of WME3210) receives less than 1 % weight. This is also the case for IMS07 of WME321X. This may imply that including unsuitable members in the ensemble (i.e., the standard-resolution members IMS04 and IMS05, or members IMS01 and IMS03 not presenting LC oscillation) can result in flawed model weights. This also might be attributed to model weighting that optimizes the error cancellation of these members, resulting in underplaying robust models. These first two remarks suggest the importance of the prescreening when optimal model weighting is used. Even when subset selection is not employed (i.e., WME321X), prescreening helps evaluate the model weighting method and results. 3) With respect to WME32XX, its members with a prescreening score of three generally receive higher model weights than members with a 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.

3.2. Predictive performance

We evaluated 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 a LC-S frequency x_0 of 0.227 (versus 0.273 and 0.227 for the reanalysis data and the reference ensemble SME321X, respectively), which corresponds to an oscillating event count errorx_2, of two. With respect to the temporal match, it is generally not expected between the simulations of ESMs and reanalysis data, yet with the absence of large drift, pseudotemporal relation might be possible. This secondary evaluation criterion can provide additional insights on the frequency and trend of red tide. The ensembles have temporal match error x_2 of 18 % except for WME3XXX that has an error of 23 %, versus 36 % for the reference ensemble. Model weighting also reduces the temporal match error LC-S x_3 for all the ensembles to 42 % (except 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 proportional to the ensemble size with the exception of WME32XX.

From a model weighting perspective, both predictive performance and ensemble size are evaluated. WME32XX has the same predictive performance as WME321X and WME3210, in terms of x_0-x_4. The three ensembles have very similar predictive performance in terms ofx_5. WME32XX ($K = 7$) has smaller ensemble size than WME321X ($K = 9$)

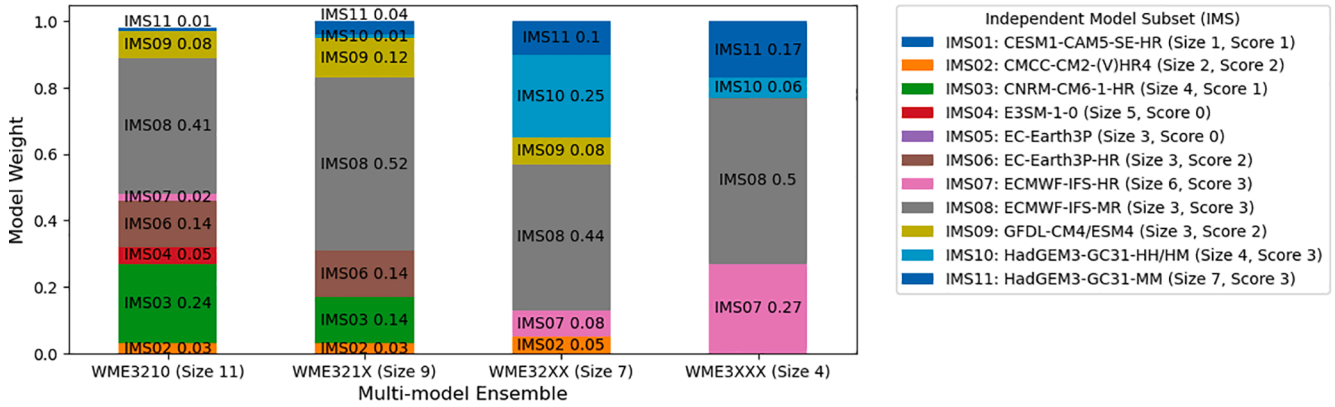


Fig. 4. Model weights of ensemble members: Data present independent model subsets for each weighted multi-model ensemble (WME). The legend indicates the number of each ensemble member, model name(s), size of the ensemble member, and score of the ensemble member. The size of the ensemble member refers 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.

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 (Equation6) for each ensemble.

	Model runs	Count		Count LC-N No-Bloom		Count LC-S No-Bloom		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

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 of the ensembles using Equations (11) and (12), respectively. 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 respectively the best and worst performing ensembles from a model selection perspective. In summary, the prescreening-based subset-selection step improves the model weighting, resulting in the 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

Fig. 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.

the best predictive performance, it is still favorable from a model selection perspective by balancing the ensemble performance and complexity. Yet from a practical perspective the smallest ensemble is not necessarily the best choice (particularly if the ensemble is too small) because this can lead to potential loss in projection accuracy (Weigel et al., 2010).

The predictive performance of the simple-average and weighted-average multi-model ensembles are shown in Fig. 5. The ensembles based on prior information (e.g., SME321X) correspond better to reanalysis data than SME3210 without prior information. Similarly, the ensembles based on prescreening information (i.e., SME32XX and SME3XXX) are better than the reference ensemble SME321X. In addition, the ensembles with model weighting have generally good correspondence with respect to the reanalysis data, irrespective of prior and prescreening information. However, ensembles with prescreening information and model weighting (i.e., WME32XX and WME3XXX) have the best correspondence with reanalysis data.

4. Discussion

With respect to key metrics, the effects of different ensemble composition criteria are summarized in Fig. 6. 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 either 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).

Second, our study reveals a caveat of optimal model weighting. Models with poor performance showing both overestimation and underestimation can receive higher model weights due to error cancellation. This is undesirable. Li et al. (2021) present a similar study in which good model simulations are obtained due to neutralizing large positive bias by large negative bias. The large biases indicate inaccurate representations of physical processes. In addition, 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 weighting. We use binary model weights in the way similar to that of Herger et al. (2018) in which models are either included or excluded. It has been argued that model uncertainty can be reduced by giving more weight to models that are more skillful and realistic for a specific process or application (Lorenz et al., 2018).

Third, we show that subset selection alone can be an effective way to improve the predictive performance in case that 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. However, it is still an open question whether unequal model weighting methods improve upon equal model weighting methods. 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 if the goodness-of-fit of the critical model is less than other models (Elshall et al., 2020b). The critical model is the model that has the largest effect on the solution reliability but might not necessarily have the largest weight (Kourakos and Mantoglou, 2008; Elshall et al., 2020a; b). Even

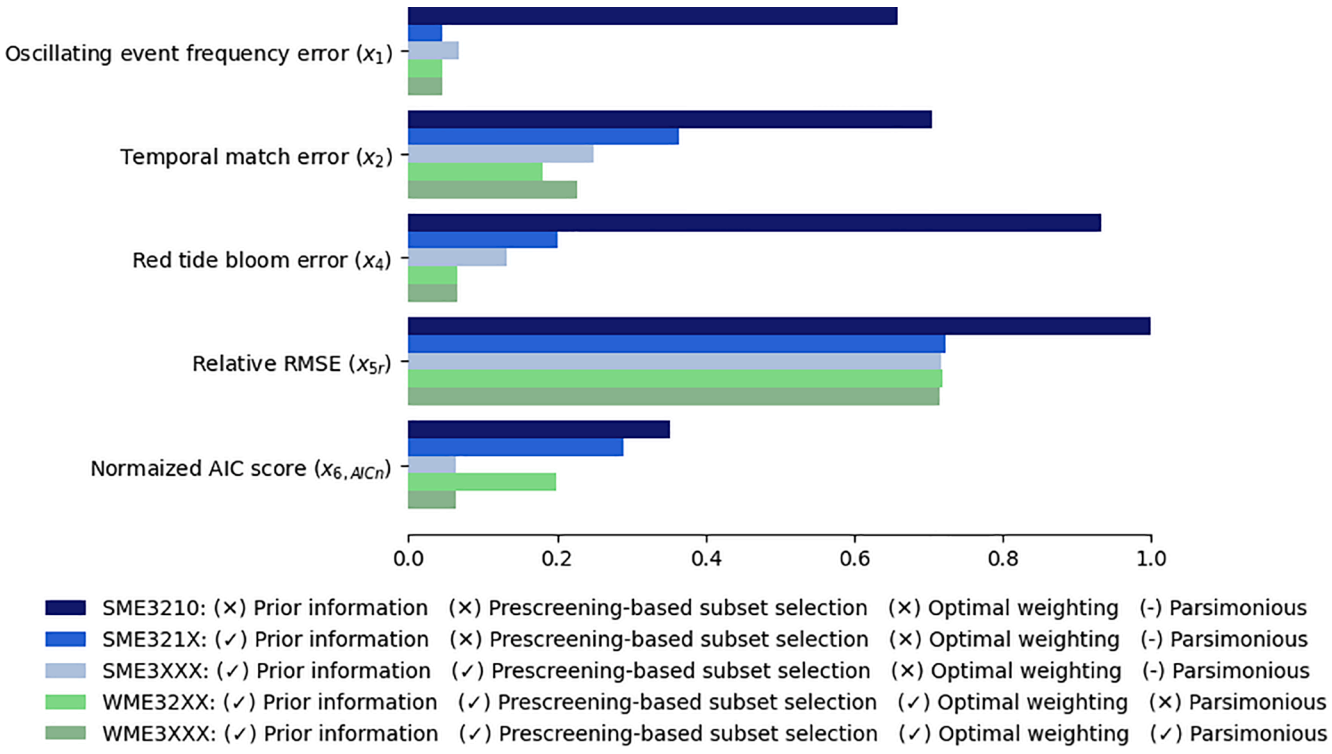


Fig. 6. Predictive performance, and AIC score accounting for performance and complexity, given different ensemble composition criteria of prior information, prescreening-based subset section, optimal model weighting, and parsimony. To scale the data to the graph we calculate the relative RMSE(x_{5r}); 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$N = 44$..

more detrimental is that a model could have a larger 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 does not necessarily result in an improved description of mean system 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, it is hard to justify that the construction of model weights using observed data of the twentieth century will persist throughout the twenty-first century (Haughton et al., 2015). Given these and similar remarks, Weigel et al. (2010) suggest that for many applications using 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 the 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 is also indicated by other methods for improving predictive performance. For example, Wang et al. (2019) show that when bias correction is applied, unequal model weighting does 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) show 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 in this case study, especially when model weighting does not result in significant improvement (e.g., SME3XXX and WME3XXX).

Finally, the application-specific nature of model selection and/or averaging should not be overlooked, as there are no universally-best methods. 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. Their study shows that methods and models used should be carefully chosen, and that obtained results should be interpreted with caution. Similarly, with respect to 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 a number of factors (e.g., metric, variable, observational estimate, location, time, spatial scale, and calibration time period) emphasizes that model dependence is application-specific, and not a general property of an ensemble. 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 the application-specific ensemble methods, such as the two presented in this study (i.e., prescreening-base subset selection and application-specific optimal weighting) can be useful. Lastly, prescreening-based subset-selection entails an extreme form of weighting such that models that are not suitable for this application or variable, are discarded. Yet this does not mean that the discarded models do not have robust performance with respect to other variables related to red tides (e.g., wind speed and direction, and sea-surface temperature) and other aspects of global or regional climates. In contrast to the generic evaluation of ESMs irrespective of the applications, subset selection can include region-, application-, and sector-specific metrics depending on the modeling proposes.

5. Conclusions

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

1. While optimal model weighting can potentially improve predictive performance, at least one caveat needs to be considered. Including non-representative models with both overestimation 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 projections. However, this study clearly shows that, when the optimal ensemble weight approach is used, including these non-representative models can underplay the model weights of robust models.
2. Excluding all non-representative models results in the most parsimonious ensemble accounting for both ensemble size and performance.
3. Prescreening-based subset selection, which screens and selects ensemble members based on their ability to reproduce certain key features, is a viable option that can either substitute model weighting, or be used prior to model weighting. Prescreening-based subset selection does not only help to develop a parsimonious ensemble, but also provides insights about the validity of the model weights.

The insights provided by this study add to the literature of application-specific optimal model weighting of ESMs. The analysis in this study is limited to the historical period. Model weighting can be based not only on historical performance, but also on the spread and convergence of future projections. Exploring optimal model weighting with respect to the trade-off between historical and future performance, is warranted in a future study.

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.

Funding

This work is funded by NSF Award #1939994.

CRediT authorship contribution statement

AE, MY, SK, JH, XY, YW, and MM: Conceptualization. **AE and MY:** Data curation and Formal analysis. **MY, SK, and JH:** Funding acquisition. **AE, MY, SK, JH, XY, YW, and MM:** Investigation. **AE, and MY:** Methodology. MY: Project administration, Resources, and Supervision. **AE:** Software, and Visualization. **AE, MY, SK, JH, XY, YW, and MM:** Validation. **AE:** Writing – original draft. **AE, MY, SK, JH, XY, YW, and MM:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and codes used are publically available as cited in the manuscript.

Acknowledgments

We thank two anonymous reviewers for their thoughtful comments that helped to improve the manuscript. 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 also thank the FWC for their provision of the data. We are grateful to Maria J. Olascoaga in the Department of Ocean Sciences at the University of Miami for our communication regarding the data analysis *of Karenia brevis* data. We wish to also thank Sally Gorrie, Emily Lizotte, Mike Stukel, and Jing Yang in EOAS at FSU for their fruitful discussions and suggestions relating to this project. We dedicate this paper to the memory of Stephen Kish, former professor in EOAS at FSU, who was inspirational in the research and planning for this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cliser.2022.100334.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., Schmidt, G.A., 2019. ESD reviews: model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth Syst. Dyn. 10 (1), 91–105.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Annan, J.D., Hargreaves, J.C., 2017. On the meaning of independence in climate science. Earth Syst. Dyn. 8, 211–224. https://doi.org/10.5194/esd-8-211-2017.

Bett, P.E., Thornton, H.E., Lockwood, J.F., Scaife, A.A., Golding, N., Hewitt, C., Zhu, R., Zhang, P., Li, C., 2017. Skill and reliability of seasonal forecasts for the Chinese energy sector. J Appl Meteorol Climatol 56 (11), 3099–3114.

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

Bhowmik, R.D., Sankarasubramanian, A., 2021. A performance-based multi-model combination approach to reduce uncertainty in seasonal temperature change projections. Int. J. Climatol. 41, E2615–E2632. https://doi.org/10.1002/joc.6870.

Boé, J., 2018. Interdependency in multimodel climate projections: component replication and result similarity. Geophys. Res. Lett. 45, 2771–2779. https://doi.org/10.1002/2017GL076829.

Brand, L.E., Compton, A., 2007. Long-term increase in Karenia brevis abundance along the Southwest Florida Coast. Harmful Algae 6, 232–252. https://doi.org/10.1016/j.hal.2006.08.005.

Brunner, L., Pendergrass, A.G., Lehner, F., Merrifield, A.L., Lorenz, R., Knutti, R., 2020. Reduced global warming from CMIP6 projections when weighting models by performance and independence. Earth Syst. Dyn. 11 (4), 995–1012.

Caldwell, P.M., Mametjanov, A., Tang, Q.i., Van Roekel, L.P., Golaz, J.-C., Lin, W., Bader, D.C., Keen, N.D., Feng, Y., Jacob, R., Maltrud, M.E., Roberts, A.F., Taylor, M. A., Veneziani, M., Wang, H., Wolfe, J.D., Balaguru, K., Cameron-Smith, P., Dong, L. u., Klein, S.A., Leung, L.R., Li, H.-Y., Li, Q., Liu, X., Neale, R.B., Pinheiro, M., Qian, Y., Ullrich, P.A., Xie, S., Yang, Y., Zhang, Y., Zhang, K., Zhou, T., 2019. The DOE E3SM coupled model version 1: description and results at high resolution. J. Adv. Model. Earth Syst. 11 (12), 4095–4146.

Ceglar, A., Toreti, A., Prodhomme, C., Zampieri, M., Turco, M., Doblas-Reyes, F.J., 2018. Land-surface initialisation improves seasonal climate prediction skill for maize yield forecast. Sci. Rep. 8 (1) https://doi.org/10.1038/s41598-018-19586-6.

Chang, P., Zhang, S., Danabasoglu, G., Yeager, S.G., Fu, H., Wang, H., Castruccio, F.S., Chen, Y., Edwards, J., Fu, D., Jia, Y., Laurindo, L.C., Liu, X., Rosenbloom, N., Small, R.J., Xu, G., Zeng, Y., Zhang, Q., Bacmeister, J., Bailey, D.A., Duan, X., DuVivier, A.K., Li, D., Li, Y., Neale, R., Stössel, A., Wang, L.i., Zhuang, Y., Baker, A., Bates, S., Dennis, J., Diao, X., Gan, B., Gopal, A., Jia, D., Jing, Z., Ma, X., Saravanan, R., Strand, W.G., Tao, J., Yang, H., Wang, X., Wei, Z., Wu, L., 2020. An unprecedented set of high-resolution earth system simulations for understanding multiscale interactions in climate variability and change. J. Adv. Model. Earth Syst. 12 (12) https://doi.org/10.1029/2020MS002298.

Cherchi, A., Fogli, P.G., Lovato, T., et al., 2019. Global mean climate and main patterns of variability in the CMCC-CM2 coupled model. J. Adv. Model. Earth Syst. 11, 185–209. https://doi.org/10.1029/2018MS001369.

Christensen, J.H., Kjellström, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight assignment in regional climate models. Clim. Res. 44 (2-3), 179–194.

De Felice, M., Soares, M.B., Alessandri, A., Troccoli, A., 2019. Scoping the potential usefulness of seasonal climate forecasts for solar power management. Renew Energy 142, 215–223. https://doi.org/10.1016/j.renene.2019.03.134.

DelSole, T., Yang, X., Tippett, M.K., 2013. Is unequal weighting significantly better than equal weighting for multi-model forecasting? Q. J. R. Meteorolog. Soc. 139, 176–183. https://doi.org/10.1002/qj.1961.

DelSole, T., Nattala, J., Tippett, M.K., 2014. Skill improvement from increased ensemble size and model diversity. Geophys. Res. Lett. 41, 7331–7342. https://doi.org/10.1002/2014GL060133.

Dixon, A.M., Forster, P.M., Beger, M., 2021. Coral conservation requires ecological climate-change vulnerability assessments. Front. Ecol. Environ. 19 (4), 243–250.

Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. Tellus Ser A-Dyn Meteorol Oceanol 57, 234–252. https://doi.org/10.1111/j.1600-0870.2005.00104.x.

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

Elshall, A.S., Ming, Y., Kranz, S.A., Harrington, J., Yang, X., Wan, Y., Maltrud, M., 2022. Prescreening-Based Subset Selection for Improving Predictions of Earth System Models With Application to Regional Prediction of Red Tide. Frontiers in Earth Science 10 (786223). https://doi.org/10.3389/feart.2022.786223.

Elshall, A.S., Pham, H.V., Tsai, F.-C., Yan, L., Ye, M., 2015. Parallel inverse modeling and uncertainty quantification for computationally demanding groundwater-flow models using covariance matrix adaptation. J. Hydrol. Eng. 20 (8) https://doi.org/10.1061/(ASCE)HE.1943-5584.0001126.

Elshall, A.S., Arik, A.D., El-Kadi, A.I., Pierce, S., Ye, M., Burnett, K.M., Wada, C.A., Bremer, L.L., Chun, G., 2020a. Groundwater sustainability: a review of the interactions between science and policy. Environ. Res. Lett. 15 (9), 093004.

Elshall, A.S., Ye, M., Finkel, M., 2020b. Evaluating two multi-model simulation–optimization approaches for managing groundwater contaminant plumes. J. Hydrol. 590, 125427 https://doi.org/10.1016/j.jhydrol.2020.125427.

Elshall, A., Ye, M., Kranz, S., et al., 2021. Machine learning for red tide prediction in the Gulf of Mexico along the West Florida Shelf. Accessed 17 Dec 2021 Earth Space Sci. Open Archive. https://doi.org/10.1002/essoar.10509597.1.

Elshall, A.S. (2020) Sea surface height above geoid: AVISO altimetry data versus ESM simulations of Loop Current.

Elshall, A.S. (2021) Python and MATLAB codes for application-specific optimal model weighting of GCMs with a red tide example (v1.0). Zenodo. 10.5281/zenodo.5499459.

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. Geosci. Model Dev. 9 (5), 1937–1958.

Eyring, V., Cox, P.M., Flato, G.M., Gleckler, P.J., Abramowitz, G., Caldwell, P., Collins, W.D., Gier, B.K., Hall, A.D., Hoffman, F.M., Hurtt, G.C., Jahn, A., Jones, C.D., Klein, S.A., Krasting, J.P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G.A., Pendergrass, A.G., Pincus, R., Ruane, A.C., Russell, J.L., Sanderson, B.M., Santer, B. D., Sherwood, S.C., Simpson, I.R., Stouffer, R.J., Williamson, M.S., 2019. Taking climate model evaluation to the next level. Nat. Clim. Change 9 (2), 102–110.

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

Fiedler, T., Pitman, A.J., Mackenzie, K., Wood, N., Jakob, C., Perkins-Kirkpatrick, S.E., 2021. Business risk and the emergence of climate analytics. Nat. Clim. Change 11 (2), 87–94.

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

Golaz, J.-C., Caldwell, P.M., Roekel, L.P.V., et al., 2019. The DOE E3SM coupled model version 1: overview and evaluation at standard resolution. J. Adv. Model. Earth Syst. 11, 2089–2129. https://doi.org/10.1029/2018MS001603.

Haarsma, R.J., Roberts, M.J., Vidale, P.L., et al., 2016. High resolution model intercomparison project (HighResMIP v1.0) for CMIP6. Geosci. Model Dev. 9, 4185–4208. https://doi.org/10.5194/gmd-9-4185-2016.

Hansen, N., Müller, S.D., Koumoutsakos, P., 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. 11, 1–18. https://doi.org/10.1162/106365603321828970.

Hansen, N., Ostermeier, A., 2001. Completely derandomized self-adaptation in evolution strategies. Evol. Comput. 9 (2), 159–195.

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

Heil, C.A., Dixon, L.K., Hall, E., et al., 2014. Blooms of Karenia brevis (Davis) G. Hansen & Ø. Moestrup on the West Florida Shelf: nutrient sources and potential management strategies based on a multi-year regional study. Harmful Algae 38, 127–140. https://doi.org/10.1016/j.hal.2014.07.016.

Held, I.M., Guo, H., Adcroft, A., et al., 2019. Structure and performance of GFDL's CM4.0 climate model. J. Adv. Model. Earth Syst. 11, 3691–3727. https://doi.org/10.1029/2019MS001829.

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

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., Sanderson, B.M., 2018. Selecting a climate model subset to optimise key ensemble properties. Earth Syst. Dyn. 9 (1), 135–151.

Herger, N., Abramowitz, G., Sherwood, S., Knutti, R., Angélil, O., Sisson, S.A., 2019. Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change. Clim. Dyn. 53 (3-4), 1581–1596.

Hoch, K.E., Petersen, M.R., Brus, S.R., Engwirda, D., Roberts, A.F., Rosa, K.L., Wolfram, P.J., 2020. MPAS-Ocean simulation quality for variable-resolution North American coastal meshes. J. Adv. Model. Earth Syst. 12 (3) https://doi.org/10.1029/2019MS001848.

Jacox, M.G., Alexander, M.A., Siedlecki, S., Chen, K.e., Kwon, Y.-O., Brodie, S., Ortiz, I., Tommasi, D., Widlansky, M.J., Barrie, D., Capotondi, A., Cheng, W., Di Lorenzo, E., Edwards, C., Fiechter, J., Fratantoni, P., Hazen, E.L., Hermann, A.J., Kumar, A., Miller, A.J., Pirhalla, D., Pozo Buil, M., Ray, S., Sheridan, S.C., Subramanian, A., Thompson, P., Thorne, L., Annamalai, H., Aydin, K., Bograd, S.J., Griffis, R.B., Kearney, K., Kim, H., Mariotti, A., Merrifield, M., Rykaczewski, R., 2020. Seasonal-to-interannual prediction of North American coastal marine ecosystems: forecast methods, mechanisms of predictability, and priority developments. Prog. Oceanogr. 183, 102307.

Knutti, R., Sedláček, J., Sanderson, B.M., Lorenz, R., Fischer, E.M., Eyring, V., 2017. A climate model projection weighting scheme accounting for performance and interdependence. Geophys. Res. Lett. https://doi.org/10.1002/2016GL072012.

Kourakos, G., Mantoglou, A., 2008. Remediation of heterogeneous aquifers based on multiobjective optimization and adaptive determination of critical realizations. Water Resour. Res. 44 https://doi.org/10.1029/2008WR007108.

Leduc, M., Laprise, R., de Elía, R., Šeparović, L., 2016. Is institutional democracy a good proxy for model independence? J. Clim. 29, 8301–8316. https://doi.org/10.1175/JCLI-D-15-0761.1.

Li, S., Huang, G., Li, X., Liu, J., Fan, G., 2021. An assessment of the antarctic sea ice mass budget simulation in CMIP6 historical experiment. Front. Earth Sci. 9, 649743 https://doi.org/10.3389/feart.2021.649743.

Liu, Y., Weisberg, R.H., Lenes, J.M., Zheng, L., Hubbard, K., Walsh, J.J., 2016. Offshore forcing on the "pressure point" of the West Florida Shelf: anomalous upwelling and its influence on harmful algal blooms. J. Geophys. Res. Oceans 121 (8), 5501–5515.

Lledo, L., Torralba, V., Soret, A., Ramon, J., Doblas-Reyes, F.J., 2019. Seasonal forecasts of wind power generation. Renew Energy 143, 91–100.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E.M., Knutti, R., 2018. Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. J. Geophys. Res.: Atmos. 123 (9), 4509–4526.

Lowe, R., Stewart-Ibarra, A.M., Petrova, D., García-Díez, M., Borbor-Cordova, M.J., Mejía, R., Regato, M., Rodó, X., 2017. Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador. Lancet Planet. Health 1 (4), e142–e151.

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

Manzanas, R. (2020) Assessment of model drifts in seasonal forecasting: sensitivity to ensemble size and implications for bias correction. J. Adv. Model. Earth Syst. 12: e2019MS001751. 10.1029/2019MS001751.

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

Merrifield, A.L., Brunner, L., Lorenz, R., Medhaug, I., Knutti, R., 2020. An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. Earth Syst. Dyn. 11 (3), 807–834.

Mishra, N., Prodhomme, C., Guemas, V., 2019. Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. Clim. Dyn. 52, 4207–4225. https://doi.org/10.1007/s00382-018-4404-z.

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

Payne, M.R., Hobday, A.J., MacKenzie, B.R., Tommasi, D., 2019. Editorial: seasonal-to-decadal prediction of marine ecosystems: opportunities, approaches, and applications. Front. Mar. Sci. 6 https://doi.org/10.3389/fmars.2019.00100.

Perkins, S., 2019. Inner workings: ramping up the fight against Florida's red tides. PNAS 116, 6510–6512. https://doi.org/10.1073/pnas.1902219116.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1.

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

Roberts, M.J., Baker, A., Blockley, E.W., Calvert, D., Coward, A., Hewitt, H.T., Jackson, L. C., Kuhlbrodt, T., Mathiot, P., Roberts, C.D., Schiemann, R., Seddon, J., Vannière, B., Vidale, P.L., 2019. Description of the resolution hierarchy of the global coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments. Geosci. Model Dev. 12 (12), 4999–5028.

Roberts, C.D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., Keeley, S.P.E., 2018. Climate model configurations of the ECMWF integrated forecasting system (ECMWF-IFS cycle 43r1) for HighResMIP. Geosci. Model Dev. 11 (9), 3681–3712.

Ross, A.C., Najjar, R.G., 2019. Evaluation of methods for selecting climate models to simulate future hydrological change. Clim. Change 157, 407–428. https://doi.org/10.1007/s10584-019-02512-8.

Sanderson, B.M., Wehner, M., Knutti, R., 2017. Skill and independence weighting for multi-model assessments. Geosci. Model Dev. 10, 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017.

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R Soc. A-Math. Phys. Eng. Sci. 365, 2053–2075. https://doi.org/10.1098/rsta.2007.2076.

Tebaldi, C., Smith, R.L., Nychka, D., Mearns, L.O., 2005. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. J. Clim. 18, 1524–1540. https://doi.org/10.1175/JCLI3363.1.

Vajda, A., Hyvärinen, O. (2020) Development of seasonal climate outlooks for agriculture in Finland. In: Advances in Science and Research. Copernicus GmbH, pp 269–277.

van den Hurk, B., Hewitt, C., Jacob, D., Bessembinder, J., Doblas-Reyes, F., Döscher, R., 2018. The match between climate services demands and Earth System Models supplies. Clim. Serv. 12, 59–63.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., Waldman, R., 2019. Evaluation of CMIP6 DECK experiments with CNRM-CM6-1. J. Adv. Model. Earth Syst. 11 (7), 2177–2213.

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? Hydrol. Earth Syst. Sci. 23, 4033–4050. https://doi.org/10.5194/hess-23-4033-2019.

Ward, N.D., Megonigal, J.P., Bond-Lamberty, B., Bailey, V.L., Butman, D., Canuel, E.A., Diefenderfer, H., Ganju, N.K., Goñi, M.A., Graham, E.B., Hopkinson, C.S., Khangaonkar, T., Langley, J.A., McDowell, N.G., Myers-Pigg, A.N., Neumann, R.B., Osburn, C.L., Price, R.M., Rowland, J., Sengupta, A., Simard, M., Thornton, P.E., Tzortziou, M., Vargas, R., Weisenhorn, P.B., Windham-Myers, L., 2020. Representing the function and sensitivity of coastal interfaces in Earth system models. Nat. Commun. 11 (1) https://doi.org/10.1038/s41467-020-16236-2.

Weigel, A.P., Knutti, R., Liniger, M.A., Appenzeller, C., 2010. Risks of model weighting in multimodel climate projections. J. Clim. 23, 4175–4191. https://doi.org/10.1175/2010JCLI3594.1.

Weisberg, R.H., Zheng, L., Liu, Y., Lembke, C., Lenes, J.M., Walsh, J.J., 2014. Why no red tide was observed on the West Florida Continental Shelf in 2010. Harmful Algae 38, 119–126.

Weisberg, R.H., Liu, Y., Lembke, C., et al., 2019. The coastal ocean circulation influence on the 2018 West Florida Shelf K. brevis Red Tide Bloom. J. Geophys. Res. Oceans 124, 2501–2512. https://doi.org/10.1029/2018JC014887.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J.T., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., Zebiak, S.E., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol. Appl. 24 (3), 315–325.

Xu, D., Ivanov, V.Y., 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. Int. J. Climatol. 38, 3432–3438. https://doi.org/10.1002/joc.5486.

Zhao, T., Zhang, W., Zhang, Y., Liu, Z., Chen, X., 2020. Significant spatial patterns from the GCM seasonal forecasts of global precipitation. Hydrol. Earth Syst. Sci. 24 (1), 1–16.

Zohdi, E., Abbaspour, M., 2019. Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction. Int. J. Environ. Sci. Technol. 16 (3), 1789–1806.