frontiers

Prescreening-based subset selection for improving predictions of Earth system models for regional environmental management of red tide

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 to Frontiers in Earth Science

Special Collection: Rapid, Reproducible, and Robust Environmental Modeling for Decision Support: Worked Examples and Open-Source Software Tools

September 2021

Keywords: regional environmental management, harmful algae blooms of red tide, climate models and Earth system models, HighResMIP of CMIP6, multi-model ensemble methods, sub-ensemble selection and subset selection, decision-relevant metrics

Abstract

The high-resolution Earth system models (ESMs) of the Coupled Model Intercomparison Project Phase 6 (CMIP6), which have unprecedented resolution and model fidelity, can provide many regional environmental management services that were previously not possible without downscaling. Using CMIP6 and reanalysis data, we present the multi-model ensemble methods of prescreening and subset selection to improve seasonal prediction of the ensemble of ESMs. In the prescreening step, the independent ensemble members are categorized based on their ability to reproduce main features of the physically interpretable relationships of interest. The ensemble size is then updated by selecting the subsets that improve the performance of the ensemble prediction. We discuss these ensemble methods using a case study of red tide prediction along the West Florida Shelf in the Gulf of Mexico, which has substantial environmental and socioeconomic impacts on the State of Florida. Red tide occurs worldwide and is the common name of harmful algal blooms that are caused by mixotrophic dinoflagellate such as *Karenia brevis*. Results show that prescreening-based subset selection can help to improve the ensemble prediction. This finding is pertinent to regional environmental management applications and other climate services. Additionally, our analysis follows the FAIR Guiding Principles for scientific data management and stewardship such that data and analysis tools are findable, accessible, interoperable, and reusable. As such, the open source Colab notebooks developed for data analysis are annotated in the manuscript. This allows for efficient and transparent testing of the results' sensitivity to different modeling assumptions. Moreover, this serves as a starting point to build upon for red tide management, using the publicly available CMIP, Coordinated Regional Downscaling Experiment (CORDEX), and reanalysis data.

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, and plant ecology and landuse, these models can provide many services at regional and seasonal scales that are important for a wide range of stakeholders. Seasonal 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 red tide case study using the high-resolution ESMs of CMIP6.

Red tide is a common name of harmful algae blooms that occur worldwide, and are caused by toxic mixotrophic dinoflagellate such as the *Karenia brevis*. Red tide contributes to water quality degradation worldwide, resulting in many undesirable effects. For example, the occurrence of red tide in the Gulf of Mexico has severe environmental and socioeconomic impacts on the State of Florida, USA. These impacts include affecting 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 of the algal blooms), 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 of red tide in the Gulf of Mexico have many drivers including regional ocean currents (i.e., Loop Current), local and deep-ocean upwelling, river flow, sediment transport, submarine groundwater discharge, nutrients from multiple sources (e.g., river, groundwater, ocean, and atmospheric deposition), African Sahara dust, tropical cyclones, and wind-direction (Brand and Compton, 2007; Heil et al., 2014; Weisberg et al., 2014; Maze et al., 2015). Loop Current, which is a warm ocean current that penetrates through the Gulf of Mexico, is an important driver for the occurrence of red tide in Florida (Weisberg et al., 2014; Maze et al., 2015; Perkins, 2019). Maze et al., (2015) show that the Loop Current is a necessary condition for a large red tide bloom to occur, and point out that the Loop Current can be “the first definitive predictor of bloom possibility” Using the Loop Current for red tide bloom prediction as a case study, this paper presents a process-based subset-selection method for improving predictive performance of ESMs.

To improve raw outputs directly given by ESMs for providing useful services to societal decision making, a combination of multiple methods is often used such as bias-correction to account for systematic errors (Szabó-Takács et al., 2019; Wang et al., 2019), ensemble recalibration to improve ensemble characteristics (Manzanas et al., 2019), downscaling to improve the spatial and temporal resolution (Gutowski Jr. et al., 2016; Gutowski et al., 2020), and ensemble methods to select and combine different models. Ensemble methods are an active research area as multi-model ensemble can

be more robust then a single-model ensemble (DelSole et al., 2014; Al Samouly et al., 2018; Wallach et al., 2018). Single model ensemble is a single ESM model with multiple realizations given perturbed parameters, initialization, physics, and forcings. Multi-model ensemble refers to an ensemble of multiple ESMs with single or multiple realizations of each ESM. Ensemble methods aim at selecting and combining multiple ESMs to form a robust and diverse ensemble of models. Ensemble methods include model weighting by assigning lower weights to less favorable models (Knutti, 2010; Weigel et al., 2010), bagging by using subsets of data or variables (Ahmed et al., 2019), subset-selection in which the best performing independent models are selected (Chandler, 2013; Herger et al., 2018; Ahmed et al., 2019; Hemri et al., 2020), and the combination of these methods (e.g., using subset selection prior to model weighting).

This study focuses on subset selection, which has not received adequate attention in climate and Earth system research (DelSole et al., 2013; Herger et al., 2018). In subset selection, a subset of models, which have better performance in a set of models, are selected as ensemble members. One model could perform better than other models due to more accurate parameterizations, higher spatial resolution, more tight calibration to relevant data sets, inclusion of more physical components, more accurate initialization, and imposition of more complete or more accurate external forcings (Haughton et al., 2015). In addition, one model could perform better than another model for a specific application as we show in this study. Accordingly, a question that often arises in multi-model combination is whether the original set of models should be screened such that "poor" models are excluded before model combination (DelSole et al., 2013). One argument is that combining all "robust" and "poor" models to form an ensemble (e.g., by assigning lower weights for poorly performing models than others) is an intuitive solution that has advantage over subset selection that uses the best performing model (Haughton et al., 2015). One justification is that, while the "poor" model can be useless by itself, it is useful when combined with other models due to error cancellation (Knutti et al., 2010; DelSole et al., 2013; Herger et al., 2018). Another justification is that no small set of models can represent the full range of possibilities for all variables, regions and seasons (Parding et al., 2020). However, it has also been argued that, the objective of subset selection is to create an ensemble of well-chosen, robust and diverse models, and thus if the subset contains a large enough number of the highest ranked and independent models, then it will have the characteristics that reflect the full ensemble (Evans et al., 2013).

Subset selection has several advantages and practical needs. First, a thorough evaluation is generally required to remove doubtful and potentially erroneous simulations (Sorland et al., 2020), and to avoid the least realistic models for a given region (McSweeney et al., 2015). Second, predictive performance can generally improve from model diversity rather than from larger ensemble (DelSole et al., 2014). A reason for this is that as more models are included in an ensemble, the amount of new information diminishes in proportion, which may lead to overly confident climate predictions (Pennell and Reichler, 2011). Accordingly, several studies (Herger et al., 2018; Ahmed et al., 2019; Hemri et al., 2020) developed evaluation frameworks in which subset selection is performed prior to model weighting. A third advantage of subset selection is to identify models based on physical relationships highlighting the importance of process-based model evaluation. For example, Knutti et al. (2017) defined the metric of September Arctic sea ice extent, showing that models that have more sea ice in 2100 than observed today and models that have almost no sea ice today are not suitable for the projection of future sea ice. There is no obvious reason to include these "poor model" that cannot simulate the main process of interest. Likewise, for our case study, we show that models that are unable to simulate Loop Current in the northern position are unsuitable for our environmental management objectives. Yun et al. (2017) indicate that incorporating such process-based information is important for highlighting key underlying mechanistic processes of the individual models of the ensemble. Fourth, subset selection allows for flexibility in terms of metrics and thresholds to tailor the multi-

model ensemble for the needs of specific applications (Bartók et al., 2019). As noted by Jagannathan et al. (2020), model selection studies are often based on evaluations of broad physical climate metrics (e.g., temperature averages or extremes) at regional scales, without additional examination of local-scale decision-relevant climatic metrics, which can provide better insights on model credibility and choice. For example, Bartók et al. (2019) and Bartók et al. (2019) employ subset selection to tailor the ensemble for energy sector needs, and local agricultural need in California, respectively. Finally, another practical need for subset selection is that, due to high computational cost, it is common that only a small subset of models can be considered for downscaling (Ahmed et al., 2019; Parding et al., 2020; Sorland et al., 2020).

Although there is a need for an efficient and versatile method that finds a subset which maintains certain key properties of the ensemble, few work has been done in climate and Earth system research (Herger et al., 2018). Without a well-defined guideline on optimum subset selection (Herger et al., 2018; Ahmed et al., 2019; Bartók et al., 2019; Parding et al., 2020), it is unclear how to best utilize the information of multiple imperfect models with the aim of optimizing the ensemble performance and reducing the presence of duplicated information (Herger et al., 2018). It may be difficult to predict exactly how many models are necessary to meet certain criteria, and subsets with good properties in one region are not guaranteed to maintain the same properties in other regions (Ross and Najjar, 2019). Typically, modelers make their own somewhat subjective subset choices, and use equal weighting for the models in the subset (Herger et al., 2018). A commonly used approach is model ranking, typically based on model performance to select the top models, which is generally the top three to five models (Jiang et al., 2015; Xuan et al., 2017; Hussain et al., 2018; Ahmed et al., 2019). For example, to derive an overall rank for each model, Ahmed et al. (2019) use comprehensive rating metric to combine information from multiple goodness-of-fit measures for multiple climate variables based on the ability to mimic the spatial or temporal characteristics of observations. Then to form the multi-model ensemble, Ahmed et al. (2019) select the four top-ranked models to evaluate the two cases of equal weighting and a bagging technique of random forest regression. A limitation of this approach is the arbitrary choice of the number of the top ranked model to include. For example, Ross and Najjar (2019) evaluate six subset-selection methods with respect to performance, and investigate the sensitivity of the results to the number of model chosen. They show that selection methods and models used should be carefully chosen. To aid this common approach of subset selection, Parding et al. (2020) present an interactive tool to compare subsets of CMIP5 and CMIP6 models based on their representation of the present climate, with user-determined weights indicating the importance of different regions, seasons, climate variables, and skill scores. This allows the users to understand the implications of their different subjective weights and ensemble member choices.

A less subjective approach for subset selection is to use a method that is designed to address specific key properties of the ensemble. In other words, a subset-selection method finds a subset which maintains certain key properties of the ensemble. Key properties include any combination of several criteria. Performance is one criterion, which reflects the model's skills in representing past and present climate and Earth system states. Examples include subset-selection methods to favor skilled models (Bartók et al., 2019), and to eliminate models with poorest representation of the present system states (Parding et al., 2020). A second criterion is the range of projected climate and Earth system changes. For example, McSweeney et al. (2015) developed a subset-selection method that captures the maximum possible range of changes in surface temperature and precipitation for three continental-scale regions. Model spread is the third criterion. This is to ensure that the ensemble contains representative models that conserve as much as possible the original spread in climate sensitivity and climate future scenarios with respect to variables of interest (Mendlik and Gobiet, 2016; Bartók et al., 2019). Another related criterion for subset selection is to produce a smaller ensemble that captures extreme events (Cannon, 2015; Farjad et al., 2019). Although some sectors are affected by mean

196 climate changes, the most acute impacts are related to extreme events (Eyring et al., 2019). Accounting
197 for extreme events and model spread criteria are generally assessed with clustering based methods. For
198 example, Mendlik and Gobiet (2016) first use principal component to find common patterns of climate
199 change within the multi-model ensemble, then use cluster analysis to detect model similarities with
200 regard to these multivariate patterns, and subsequently generate a subset of representative simulations
201 by sampling models from each cluster. Thus, the approach of Mendlik and Gobiet (2016) accounts for
202 both model spread and model independence. With respect to model independence criterion, Sanderson
203 et al. (2015) propose a stepwise model elimination procedure that maximizes intermodel distances to
204 find a diverse and robust subset of models. Similarly, Evans et al. (2013) and Herger et al. (2018) use
205 an indicator method with binary weights to find a small subset of models that reproduce certain
206 performance and independence characteristics of the full ensemble. Binary weights are either zero or
207 one for models to be either discarded or retained, respectively. An additional criterion that is
208 particularly important from many climate services is to consider decision-relevant metrics (Bartók et
209 al., 2019; Jagannathan et al., 2020). Since a primary goal of climate research is to identify how climate
210 affects society and to inform decision making, a community generally needs rigorous regional-scale
211 evaluation for different impacted sectors that include agriculture, forestry, water resources,
212 infrastructure, energy production, land and marine ecosystems, and human health (Eyring et al., 2019).
213 A subset-selection method can be for a general model evaluation irrespective of the application
214 (Sanderson et al., 2017). Alternatively, a subset-selection method can be for regional model evaluation
215 with sector-specific information (Elliott et al., 2015). This includes focusing on special criterion such
216 as extreme events (Zscheischler et al., 2018), and application-specific metrics as in this study.

217 This study complements an important aspect of subset selection by explicitly considering application
218 specific metrics for subset selection based on a prescreening step. To find more skillful and realistic
219 models for a specific process or application, we develop an indicator-based subset-selection method
220 with a prescreening step. In a prescreening step, models are scored based on physical relationships and
221 their ability to reproduce key features of interest, highlighting the importance of process-based and
222 application specific evaluation of climate models. Our method extends the indicator method based on
223 binary weights of Herger et al. (2018), by scoring each model based on evolving binary weights, which
224 are either zero or one for models to be either discarded or selected, respectively, as explained the
225 method section. Thus, irrespective of the general predictive performance of the model for the variables
226 of interest (e.g., temperature, sea surface height, wind speed, and precipitation), the model performance
227 is evaluated based on suitability to specific applications for a given problem definition with key
228 features of interest. In our case study of red tide, models that cannot reproduce key features of interest
229 are the models that cannot simulate the process of Loop Current penetration into the Gulf of Mexico,
230 for example, along with other key features as explained in the method section. Using CMIP6 and
231 reanalysis data described in the method section, we show that this prescreening-based subset-selection
232 step can help reduce the ensemble size without degrading the predictive performance. We additionally
233 illustrate the caveats of using non-representative models given the notation of error cancellation,
234 showing that that a parsimonious ensemble can be more robust.

235 In the remainder of the manuscript, we present in Section 2 the red tide case study including the CMIP6
236 data, reanalysis data, and *Karenia brevis* data. Section 2 also presents the prescreening-based subset
237 selection method. Section 3 presents the results, which is following in Section 4 by providing a
238 discussion on subset selection, challenges of seasonal prediction, and the study limitations and outlook.
239 Finally, we summarize our main findings, and draw conclusions in Section 5.

2. Methods

2.1 FAIR Guiding Principles

We follow the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016). Accordingly, the data and codes that are used and developed for this study are Findable, Accessible, Interoperable, and Reusable (FAIR). With respect to the "findable" criterion, our data and codes for data analysis are presented in Colab notebooks (Elshall, 2021). These notebooks provide the developed python codes along with rich metadata, which are data describing and giving information about the used CMIP data, reanalysis data and *Karenia brevis* data (Section 2.2). With respect to the "Accessible" criterion, the notebooks are opensource available on GitHub (Elshall, 2021), and are based on the Google Colaboratory (Colab), which is a free interactive computing environment. Colab extends the Jupyter notebook environment by allowing interactive cloud computing with data storage on Google Drive. With respect to the "interoperable" criterion, which refers to the exchange and use of information, the notebooks provide rich metadata with additional analysis details not found in the manuscript. This allows users to make use of the presented information by rerunning the codes to reproduce the results, and to understand the sensitivity of the results to different assumptions and configurations as described in the manuscript. Also, the codes can be used to visualize additional data and results that are not shown in the manuscript. With respect to the "reusable" criterion, all the used data are publicly available, and the developed codes have publicly data usage licenses. This allows the users to build additional components to the codes as discussed in the manuscript.

2.2 Data

The *Karenia brevis* cell count used in this study are from the harmful algal bloom database of the Fish and Wildlife Research Institute at the Florida Fish and the Wildlife Conservation Commission (FWRI, 2020). In the study area (Fig. 1) and given the study period from 1993-01 to 2014-12, we identify 15 time intervals of large blooms, and 29 time intervals with no bloom; each time interval is six-month long. Following Maze et al. (2015), a large 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. Similar to Maze et al. (2015) we define no bloom as the absence of large bloom. The Colab notebook "*Karenia brevis* data processing" (Elshall 2021) provides the data processing details.

We use global reanalysis data, which combine observations with shortrange weather forecast using weather forecasting models to fill the gaps in the observational records. We use the Copernicus Marine Environment Monitoring Service (CMEMS) monthly gridded observation reanalysis product that has the product identifier Global_Reanalysis_PHY_001_030 (Drévillon et al., 2018; Fernandez and Lellouche, 2018). The used CMEMS reanalysis product is a global ocean eddy-resolving reanalysis with approximatively 8 km horizontal resolution covering the altimetry from 1993 onward. Similar to CMIP6 data, we only focus on sea surface height above geoid, which is has the variable name zos according to the Climate and Forecast Metadata Conventions (CF Conventions).

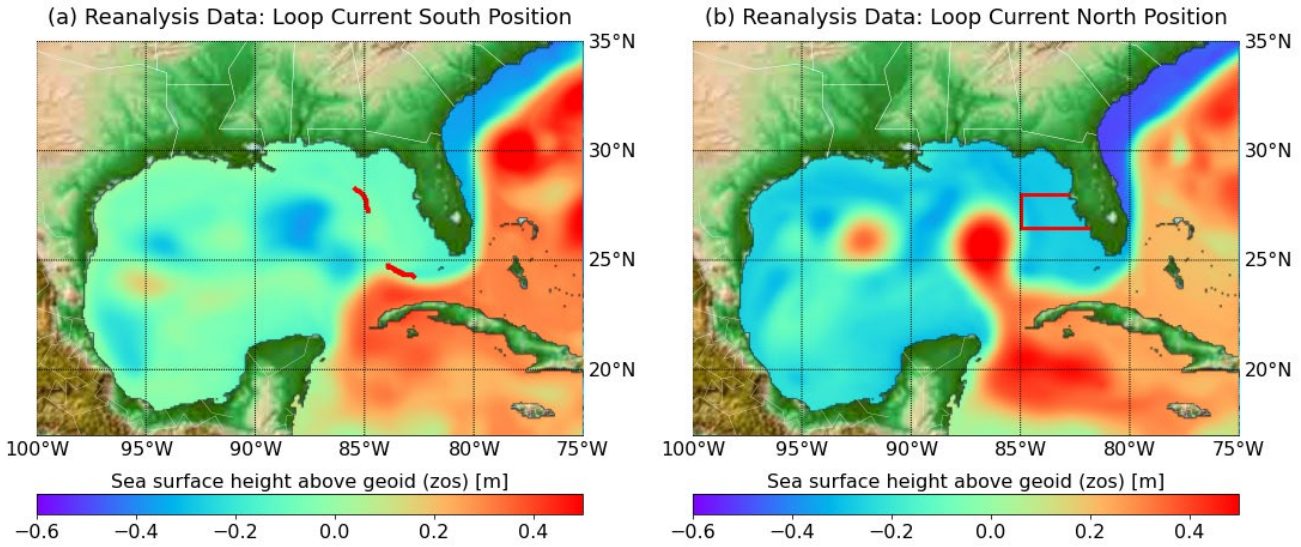

Figure 1. Observation reanalysis data of sea surface height above geoid (zos) [m] showing (a) LC-S and (b) LC-N. Two red segments along the 300m isobath in (a) are used to determine Loop Current position. The area where red tide blooms are considered by Maze et al. (2015) and this study is shown in the red box of (b)

We use 41 CMIP6 model runs from 14 different models developed by eight institutes (Roberts et al., 2018a; Cherchi et al., 2019; Golaz et al., 2019; Held et al., 2019; Roberts et al., 2019; Voldoire et al., 2019; Chang et al., 2020; Haarsma et al., 2020). The study period is from 1993-01 to 2014-12. We select CMIP6 model runs from the historical experiment (Eyring et al., 2016) and the hist-1950 experiment (Haarsma et al., 2016), which are sibling experiments that use historical forcing of recent past until 2015. The historical simulation that starts from 1850 uses all-forcing simulation of the recent past (Eyring et al., 2016). The hist-1950 experiment that starts from 1950 uses forced global atmosphere-land simulations with daily 0.25° sea surface temperature and sea-ice forcings, and aerosol optical properties (Haarsma et al., 2016). For high-resolution models, our selection criteria are to select all model runs with gridded monthly "sea surface height above geoid" , which is has the variable name zos according to the Climate and Forecast Metadata Conventions (CF Conventions), with nominal resolution less than or equal to 25 km. For each model we only consider the variable zos. Given the available CMIP6 data until September 2020 when this study started, this resulted in 33 model runs. We mainly focus on high-resolution models with eddy-rich ocean resolution, which is important for simulating Loop Current. For our analysis purpose, we include two models with standard resolution. One is EC-Earth3P with nominal ocean resolution of about 100km given in the hist-1950 experiment with three model runs, and E3SM-1-0 with variable ocean resolution of 30-60 km given in the historical experiment with five model runs.

2.3 Model independence

To account for model independence, we use institutional democracy (Leduc et al., 2016), which can be regarded as a first proxy to obtain an independent subset (Herger et al., 2018), reflecting a priori definition of dependence. For the same institution we created further subsets for different grids. This is the case for the standard- and medium-resolution models of EC-Earth-Consortium that use ORCA1 and ORCA025 grids, respectively. It is also the case for the high-resolution and medium-resolution model of MOHC-NERC that uses ORCA12 and ORC025 grids, respectively. The ORCA family is a series of global ocean configurations with tripolar grid of various resolutions. Thus, the considered 14 models that are listed aphetically by model name in Table 1, results in 11 independent model subsets.

Table1. Independent model subsets based on institutional democracy and using ocean grid as a secondary criterion when applicable.

Independent model subset (IMS)	Institution	Country	Model (Reference)	Experiment ID	Members	Ocean model resolution	Ocean Model	Ocean grid	ESM nominal resolution
IMS01	NCAR	USA	CESM1-CAM5-SE-HR (Chang et al. 2020)	hist-1950	r1i1p1f1	0.1◦ (11 km) nominal resolution	POP2	POP2-HR	25 km
IMS02	CMCC	Italy	CMCC-CM2-HR4 (Cherchi et al. 2019)	hist-1950	r1i1p1f1	0.25◦ from the Equator degrading at the poles	NEMO v3.6	ORCA025	25 km
			CMCC-CM2-VHR4 (Cherchi et al. 2019)	hist-1950	r1i1p1f1	0.25◦ from the Equator degrading at the poles	NEMO v3.6	ORCA025	25 km
IMS03	CNRM-CERFACS	France	CNRM-CM6-1-HR (Voldoire et al. 2019)	hist-1950	r(1-3)i1p1f2	0.25◦ (27-28 km) nominal resolution	NEMO v3.6	eORCA025	25 km
			CNRM-CM6-1-HR (Voldoire et al. 2019)	historical	r1i1p1f2	0.25◦ (27-28 km) nominal resolution	NEMO v3.6	eORCA025	25 km
IMS04	DOE-E3SM-Project	USA	E3SM-1-0 (Golaz et al. 2109)	historical	r(1-5)i1p1f1	60 km in mid-latitudes and 30 km at the equator and poles	MPAS-O	EC60to30	100 km
IMS05	EC-Earth-Consortium	Europe	EC-Earth3P (Haarsma et al. 2020)	hist-1950	r(1-3)i1p2f1	about 1◦ (110 km)	NEMO v3.6	ORCA1	100 km
IMS06	EC-Earth-Consortium	Europe	EC-Earth3P-HR (Haarsma et al. 2020)	hist-1950	r(1-3)i1p2f1	about 0.25◦ (27-28 km)	NEMO v3.6	ORCA025	25 km
IMS07	ECMWF	Europe	ECMWF-IFS-HR (Roberts et al. 2018)	hist-1950	r(1-6)i1p1f1	25 km nominal resolution	NEMO v3.4	ORCA025	25 km
IMS08			ECMWF-IFS-MR (Roberts et al. 2018)	hist-1950	r(1-3)i1p1f1	25 km nominal resolution	NEMO v3.4	ORCA025	25 km
IMS09	NOAA-GFDL	USA	GFDL-CM4 (Held et al 2019)	historical	r1i1p1f1	0.25◦ (27-28 km) nominal resolution	MOM6	tri-polar grid	50 km
			GFDL-ESM4 (Held et al 2019)	historical	r(2-3)i1p1f1	0.25◦ (27-28 km) nominal resolution	MOM6	tri-polar grid	50 km
IMS10	NERC	UK	HadGEM3-GC31-HH (Roberts et al. 2019)	hist-1950	r1i1p1f1	8 km nominal resolution	NEMO v3.6	ORCA12	10 km
	MOHC-NERC	UK	HadGEM3-GC31-HM (Roberts et al. 2019)	hist-1950	r1i(1-3)p1f1	25 km nominal resolution	NEMO v3.6	ORCA12	50 km
IMS11	MOHC	UK	HadGEM3-GC31-MM (Roberts et al. 2019)	hist-1950	r1i(1-3)p1f1	25 km nominal resolution	NEMO v3.6	ORCA025	100 km
			HadGEM3-GC31-MM (Roberts et al. 2019)	historical	r(1-4)i1p1f3	25 km nominal resolution	NEMO v3.6	ORCA025	25 km

For each independent model subset (IMS), multiple perturbed runs of (parameter) realizations (r), initializations (i), physics (p), and forcings (f) are considered. For example, IMS01 has only one model run r1i1p1f1, and IMS11 has seven model runs, three with perturbed initialization r1i(1-3)p1f1, and four with perturbed parameter realizations r(1-4)i1p1f3 as shown in Table 1. Note that this naming convention are relative given different modeling groups. For example, the coupled E3SM-1-0

simulations (Golaz et al., 2019) use five ensemble members that are r(1-5)i1p1f1 representing five model runs with different initialization. Each ensemble member (i.e., independent model subset, IMS) in Table 1 contains one or models, and each model has one or more model runs. These model runs of each ensemble member should not simply be included in a multi-model ensemble as they represent the same model, hence artificially increasing the weight of models with more model runs. On the other hand, using only one model run per ensemble member discards the additional information provided by these different runs (Brunner et al., 2019). Accordingly, the zos data of each ensemble member is averaged in the way described in Section 2.2.

While with the default model independence criteria of institutional democracy and ocean grid we identify 11 ensemble members listed in Table 1, the Colab notebook "Model Independence" (Elshall, 2021) provides other model independence criteria that can be investigated by the users. One model independence criterion is to assume that all models in Table 1 are independent, resulting in 16 ensemble members. Another criterion is the default criteria with the additional assumption that ensemble members with historical and hist-1950 experiments are independent, resulting in 13 ensemble members. A third criterion is the default criteria with the additional assumption that the models with high and medium resolutions are independent, resulting in 12 ensemble members. The code additionally allows for any user defined criteria. While the presented results in this manuscript all based on the default model independence criteria, the user can efficiently and transparently investigate the sensitivity of the prescreening and subset selection results under different model independence criteria.

2.4 Loop Current position and *Karenia brevis* blooms

The mechanisms of initiation, growth, maintenance, and termination of red tides have not been fully understood. Yet Loop Current, which is a warm ocean current that moves into the Gulf of Mexico, is an important factor that controls the occurrence of red tide (Weisberg et al., 2014; Maze et al., 2015; Perkins, 2019). Maze et al. (2015) shows that the difference between time intervals of large blooms and no blooms is statistically significant for the Loop Current's position. Maze et al. (2015) also show that the Loop current in a north position penetrating through the Gulf of Mexico is a necessarily condition for a large *Karenia brevis* bloom to occur. As such, when the Loop Current is in the south position shown in Fig. 1a, which is hereinafter denoted as Loop Current-South (LC-S), then there is no large bloom (Maze et al., 2015). When the Loop Current is in the north position shown in Fig. 1b, which hereinafter is denoted as Loop Current-North (LC-N), then there could be either large blooms or no blooms. This relationship between the loop current positions and *Karenia brevis* is based on retention time. With approximately 0.3 divisions per day, *Karenia brevis* is a slow growing dinoflagellate that requires an area with mixing slower than the growth rate to form a bloom (Magaña and Villareal, 2006). As such, LC-N increases the retention rate allowing bloom formation, if other conditions are ideal (Maze et al., 2015). While there are several studies that establish different relationships between Loop Current and *Karenia brevis* (Weisberg et al., 2014, 2019; Maze et al., 2015; Liu et al., 2016), the aim of this study is not to support or refute any of these relationships, but to use the study of Maze et al. (2015) for the purpose of our subset selection analysis.

The LC and its eddies can be detected from sea surface height variability. When the difference between the average sea surface height of the north and south segments along the 300 m isobath (Fig. 1a) is positive and negative, this is a good proxy for identify LC-N and LC-S, respectively (Maze et al., 2015). The zos data processing steps to determine the Loop Current positions (i.e., LC-N and LC-S) are as follows:

(1) The zos data is preprocessed for the north and south segments (Fig.1a) for all model runs and observation analysis data. Model runs and observation reanalysis data are sampled using nearest neighborhood method along the line points (approximately spaced at 1 km interval between two neighboring points) of the north and south segments (Fig. 1a). The nearest neighborhood sampling

is performed using the python package of xarray project (http://xarray.pydata.org) that handles NetCDF (Network Common Data Form) data formats with file extension NC that is used typically for climate data (e.g., CMIP and reanalysis data). As such, we have a zos datum $h_{(j,k,l,m,n,t)}$ for a model run with index j, an ensemble member with index k, a spatial point along the segment with index l, a segment (i.e., the north or south segment in Fig. 1a) with index m, a model and reanalysis datasets temporal interval (i.e., one month) with index n, and a prediction interval with index t.

(2) The expectation of zos data is taken for all model runs $j \in [1, J]$ of each ensemble member M_k

$$h_{k,l,m,n,t} = E_j(h_{j,k,l,m,n,t} \mid M_k) \tag{1}$$

The size J of each ensemble member varies depending on the number of model runs in the ensemble member, with the minimum $J = 1$ for ensemble member IMS01 and the maximum $J = 7$ for ensemble member IMS11 (Table 1). This step is not required for the reanalysis data since there is only one realization.

(3) The zos data is averaged for all ensemble members $k \in [1, K]$

$$h_{l,m,n,t} = E_k\left(E_j(h_{j,k,l,m,n,t} \mid M_k)\right) \tag{2}$$

where k is the index of each ensemble member M_k. The size K of the multi-model ensemble varies based on subset selection (Section 2.6), which determines the inclusion and exclusion of ensemble members. For example, using all available ensemble members without any subset selection results in $K = 11$ that is all the independent model subsets in Table 1. If we evaluate k for only one ensemble member for prescreening purpose (Section 2.6), then $K = 1$.

(4) For each of the north and south segments the expected zos is calculated for each segment

$$h_{m,n,t} = E_l\left[E_k\left(E_j(h_{j,k,l,m,n,t} \mid M_k)\right)\right] \tag{3}$$

(5) The zos data of the north segment is subtracted from the south segment

$$h_{n,t} = \Delta_m\left[E_l\left[E_k\left(E_j(h_{j,k,l,m,n,t} \mid M_k)\right)\right]\right] \tag{4}$$

resulting in zos difference data $h_{n,t}$ with $n \in [1, N]$ and $t \in [1, T]$. As such, N represents the interval length such that $N = 3$ for a season interval, and $N = 6$ for a semiannual interval, and T represents the number of intervals. For example, given $N = 6$ as considered in this study and the 22-year study period, then $T = 44$.

(6) The maximum $h_{n,t}$ in the 6-month interval is selected to obtain the zos anomaly per time interval

$$h_t = \max_{h_n}\left(\Delta_m\left[E_l\left[E_k\left(E_j(h_{j,k,l,m,n,t} \mid M_k)\right)\right]\right]\right) \tag{5}$$

For each zos anomaly datum h_t, positive and negative values are used as an indicator of LC-N dominated interval and LC-S dominated interval, respectively. Selecting the maximum value $\max_{h_n}(.)$ is more robust than using the average value, which may dilute the signals since the Loop Current position is a cycling event, recalling that loop current has a random and chaotic cycle with the average period of 8–18 months per cycle (Sturges and Evans, 1983; Maze et al., 2015).

The objective of this analysis is not to model the LC cycle, but rather to use the relationship between Loop Current position and *Karenia brevis* bloom of Maze et al. (2015) to obtain a heuristic coarse-temporal-resolution relation between Loop Current position and *Karenina brevis*. Thus, the h_t values given by Eq. 5 can be expressed as an indicator function for LC-N:

$$H_{LC-N}(h_t) = \begin{cases} 1, & h_t \geq 0 \\ 0, & h_t < 0 \end{cases} \tag{6}$$

and LC-S:

$$H_{LC-S}(h_t) = \begin{cases} 1, & h_t < 0 \\ 0, & h_t \geq 0 \end{cases} \tag{7}$$

such that $H_{LC-N}(h_t) = 1$ and $H_{LC-S}(h_t) = 1$ indicate a LC-N interval and LC-S interval, respectively. Eqs. 6 and 7 are convenient to use since we are not interested in the value of zos anomaly between the north and south segments per se, but rather in sign difference. Finally, Eqs. 5-7 are valid for both model simulation and observation reanalysis data, which hereinafter are donated as h_t and $h_{t,obs}$, respectively.

2.5 Model performance metrics

A model performance is based on its ability to reproduce the observed phenomena. We define three qualitative metrics to prescreen for physical relationships, and four quantitative metrics of the model performance. Based on this prescreening we can do subset selection. For prescreening, a process-based metric is needed, for example, to understand if the model can simulate certain mechanistic aspects of the problem of interest. For example, Christensen et al. (2010) use metrics that capture aspects of model performance in reproducing large-scale circulation patterns and meso-scale signals. A qualitative metric reflects if the model is suitable or unsuitable for reproducing key features of the problem. For example, Knutti et al. (2017) define a metric for the presence and absence of September Arctic sea ice, such that models with more sea ice in 2100 than observed today and models that have almost no sea ice today are not suitable for the projection of future sea ice. In our case study, models that cannot reproduce key features of interest would be the models that cannot (i) simulate the penetration of LC into the Gulf of Mexico, (ii) represent the alternation of LC in the North and South positions given the empirical method (Eqs. 5-7), and (iii) reproduce the higher frequency of Loop Current in the northern and southern positions as described below. For example, with respect to (i), the Loop Current penetrates the Gulf of Mexico extending its northward reach with eddy shedding as shown by the high-resolution model EC-Earth3P-HR (Fig. 2a-c). As such, intrusion of cooler water increases the stratification of the core of the Loop Current, and the Loop Current becomes unstable forming anticyclonic eddy that breaks from the parent Loop Current westward without reconnecting (Caldwell et al., 2019), as shown by the high-resolution model EC-Earth3P-HR (Fig. 2d-e). On the other hand, the standard-resolution model EC-Earth3P (Fig. 2f-j) cannot reproduce the observed physical phenomena, and thus unsuitable for this application. For a further illustration, Elshall (2020) shows an animation of a Loop Current cycle of year 2010 given monthly zos data for all the 41 model runs in Table 1 side-by-side with the reanalysis data. Additionally, the Colab notebook "zos data visualization" (Elshall, 2021) allows the visualization of the reanalysis data in Fig. 1 for any month in the study period 1993-2015, and CMIP data in Fig. 2 for the other ensemble members listed in Table 1 for any month in the study period. Models that are unable to simulate LC-N are unsuitable for this environmental management purpose. Justifications about selecting these three qualitative metrics and details about them are given below.

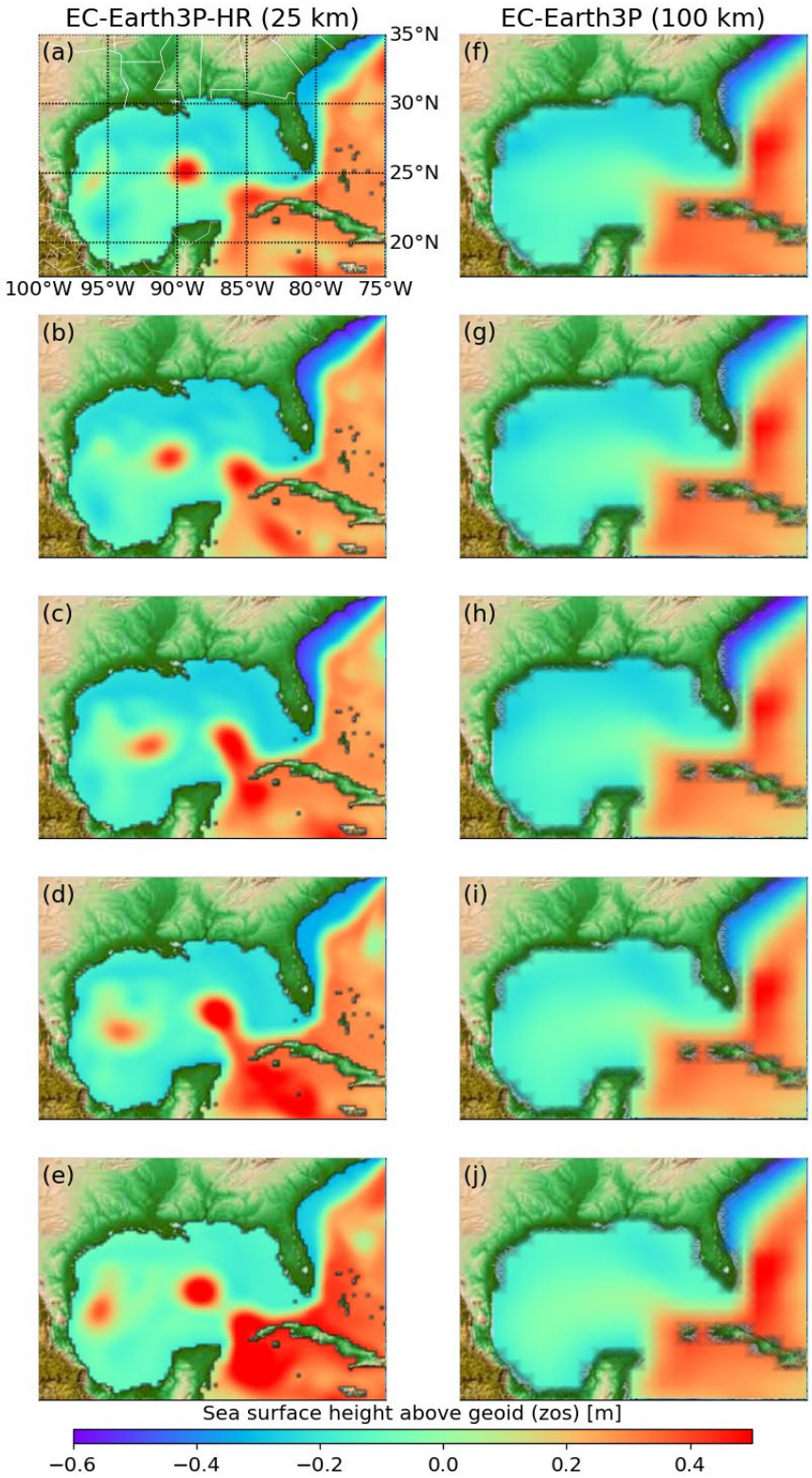

Figure 2. Snapshots of sea surface height above geoid (zos) [m] from 1993-02 to 1993-06 simulated using (a-e) a high-resolution ESM, and (f-j) standard-resolution ESM with nominal resolution of 10km and 100km, respectively.

The binary qualitative metrics (y_1 - y_3) used for prescreening are as follows:

Physical phenomena simulation (y_1): Accurate simulation of Loop Current positions is generally a challenging task, yet the objective of this first metric is to determine if the model can simulate LC-N irrespective of the accuracy. Thus, the model receives a score one $y_1 = 1$ if it can simulate LC-N (e.g., Figs. 2a-e), and zero $y_1 = 0$ otherwise (e.g., Figs. 2f-j), i.e.,

$$y_1 = \begin{cases} 1, & \sum_{t=1}^{T} H_{LC-N}(h_t) > 0 \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) = 0 \end{cases} \tag{8}$$

such that $\sum_{t=1}^{T} H_{LC-N}(h_t)$ is the count on LC-N intervals given the total number of intervals $T = 44$ as explained before.

Oscillating event representation (y_2): This metric is specific to the method of Maze et al. (2015) for determining LC-N and LC-S. If the sea surface height is consistently higher at the north segment than at the south segment, then the model is unable to represent alternation of LC-N and LC-S according to the proxy method of Maze et al. (2015). In this case, the model receives a score zero $y_2 = 0$, and one $y_2 = 1$ otherwise, i.e.,

$$y_2 = \begin{cases} 1, & 0 < \sum_{t=1}^{T} H_{LC-N}(h_t) < T \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) = T \end{cases} \tag{9}$$

Oscillating event realism (y_3): If the frequency of LC-N is greater than that of LC-S for a model, the model receives the score of one $y_3 = 1$ and zero $y_3 = 0$ otherwise, i.e.,

$$y_3 = \begin{cases} 1, & \sum_{t=1}^{T} H_{LC-N}(h_t) \geq \sum_{t=1}^{T} H_{LC-S}(h_t) \\ 0, & \sum_{t=1}^{T} H_{LC-N}(h_t) < \sum_{t=1}^{T} H_{LC-S}(h_t) \end{cases} \tag{10}$$

It is more realistic that the frequency of LC-N is greater than that of LC-S. In the study of Maze et al. (2015), the ratio of the LC-S intervals $\sum_{t=1}^{T} H_{LC-N}(h_t)$ to the total number of intervals $T = 60$ is 0.267 , given their altimetry data product with study period of 15 years and 3-month interval (i.e., $N = 3$). In this study the ratio of LC-S to total number of intervals is 0.273, given our reanalysis product with $T = 44$ and $N = 6$ as previously explained.

We define four quantitative metrics (y_4 - y_7) to evaluate the predictive performance, and the scoring rules (y_8) to evaluate complexity. These performance criteria are as follows.

Oscillating event frequency (y_4): This is the ratio of the number of a LC position (LC-S or LC-N) to the total number of intervals . Hereinafter, we refer to the oscillating event frequency as the number of LC-S to the total number of intervals T,

$$y_4 = \frac{\sum_{t=1}^{T} H_{LC-S}(h_t)}{T} \tag{11}$$

which can be compared to reanalysis data that is 0.273 as presented in the results section. Additionally, we define the oscillating event frequency error as

$$y_{4,err} = \frac{\left| \sum_{t=1}^{T} H_{LC-S}(h_t) - \sum_{t=1}^{T} H_{LC-S}(h_{t,obs}) \right|}{T} \tag{12}$$

which is the absolute difference of LC-S counts of ensemble prediction h_t and reanalysis data $h_{t,obs}$.

Temporal match error (y_5): This is a temporal match of model predictions and reanalysis data with respect to LC position for LC-N

$$y_{5,LC-N} = \frac{\sum_{t=1}^{T} H_{LC-N}(h_{t,obs}) - \sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0)}{\sum_{t=1}^{T} H_{LC-N}(h_{t,obs})} \tag{13}$$

for LC-S

$$y_{5,LC-S} = \frac{\sum_{t=1}^{T} H_{LC-S}(h_{t,obs}) - \sum_{t=1}^{T} (h_{t,obs} < 0 \wedge h_t < 0)}{\sum_{t=1}^{T} H_{LC-S}(h_{t,obs})} \tag{14}$$

and both positions

$$y_5 = \frac{T - \sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0) - \sum_{t=1}^{T} (h_{t,obs} < 0 \wedge h_t < 0)}{T} \tag{15}$$

such that $\sum_{t=1}^{T} H_{LC-N}(h_{t,obs})$ and $\sum_{t=1}^{T} H_{LC-S}(h_{t,obs})$ are the counts of the LC-N and LC-S intervals, respectively, given the observation reanalysis data $h_{t,obs}$; the terms $\sum_{t=1}^{T} (h_{t,obs} \geq 0 \wedge h_t \geq 0)$ and $\sum_{t=1}^{T} (h_{t,obs} < 0 \wedge h_t < 0)$ are the temporal match counts of model simulation and reanalysis data for LC-N and LC-S, respectively. The logical conjunction $\wedge$ gives a value of one when the statement $(h_{t,obs} \geq 0 \wedge h_t \geq 0)$ is true if $h_{t,obs} \geq 0$ and $h_t \geq 0$ are both true, otherwise gives a value of zero if false. Temporal match is the most challenging task. While ESMs are well established on climate timescale, the temporal match at seasonal timescale can be challenging (Hewitt et al., 2017). Generally speaking, the hist-1950 and historical experiments are free-running, and accordingly are neither designed nor expected to have temporal coincide with real-world conditions, which is especially true for the historical experiment. However, one aim of this study is to investigate if any temporal match is possible given the used heuristic relation for determining Loop Current position with a coarse temporal resolution of 6-month interval.

Karenia brevis error (y_6): A false negative prediction of *Karenia brevis* bloom occurs when large bloom coincides with LC-S. For the study period, we define the *Karenia brevis* error as the ratio of the number of LC-S with large bloom to the number of large-bloom N_{bloom}

$$y_6 = \frac{\sum_{t=1}^{T} (h_t < 0 \wedge H(z_t) = 1)}{N_{bloom}} \tag{16}$$

where $H(z_t)$ is an indicator function with one and zero for large bloom and no bloom, respectively.

Root-mean-square error (y_7): It is the root-mean-square error *(*RMSE) between model simulation and reanalysis data

$$y_7 = \sqrt{\frac{\sum_{t=1}^{T}(h_t - h_{t,obs})^2}{T}} \tag{17}$$

The defined metrics (y_1 - y_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 skill of these ESMs globally or regionally for general purposes. Judging the predictive skills of these models with respect to global or regional simulations of sea surface height above geoid (variable: zos) or any other variable, is beyond the scope of this work.

2.6 Prescreening and subset selection

In this study we present a subset-selection method that focuses on regional analysis with respect to a specific application. The method is based on physically interpretable relationships, and addresses application specific requirements. Such evaluation of specific regional applications is another important criterion, which is the focus of this manuscript. We develop a subset-selection method that extends the binary method of Herger et al. (2018) based on a prescreening step as shown in Fig. 3. Model independence is accounted for as described in Section 2.2, and a score is obtained for each ensemble member using three binary qualitative metrics y_1 - y_3 (Section 2.3). Binary refers to a score of either zero or one if the ensemble member is unable or able to produce the metric target. The three binary metrics (Eqs. 8-10) are evolving such that if the ensemble member fails the first metric, then it will consequently fail in the other two, and will accordingly receive a score of zero. For example, given score (y_1, y_2, y_3), the model receives a score from zero to three for score (0,0,0), (1,0,0), (1,1,0), and (1,1,1), respectively. In other words, if a model score is one for y_3 (Eq. 10) it will by default score ones for y_1 (Eq.11) and y_2 (Eq.9).

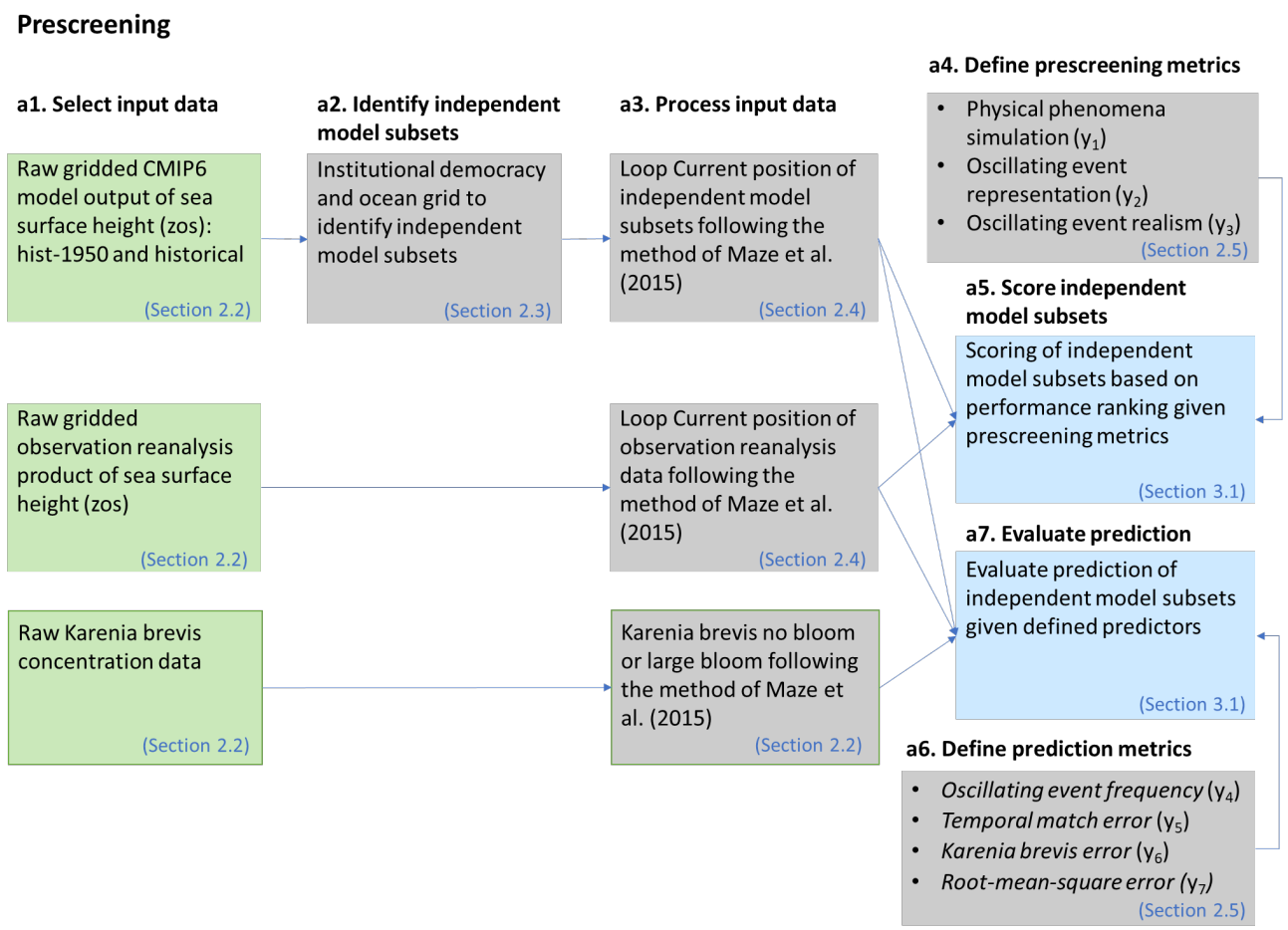

Figure 3. The prescreening method as implemented in the Colab notebook "Prescreening" (Elshall, 2021) that provides more details and the python code.

The subset selection step is shown in Fig. 4. In this step we compose five multi-model ensembles using simple-average multi-model ensemble (SME). Each SME is composed of ensemble members with different scores. For example, SME3210 contains all ensemble members with scores from zero to three, i.e., all the 11 ensemble members listed in Table1. Ensemble SME321X, SME32XX, and SME3XXX exclude ensemble members based on the three binary qualitative metrics (y_1 - y_3), respectively. These are evolving metrics such that if an ensemble member scores zero in y_1, it will score zero in y_2 and y_2, and have an overall score of zero. If a model has a score $y_3 = 1$, it will by default score one in y_1 and y_2, and have an overall score of three. As such, SME3XXX contains the best ensemble members, which are the ones with a score of three. Ensemble SME32XX contains ensemble members with scores of three and two, and so on. On the other hand, ensemble SMEXXX0 contains only the least performing ensemble members with a score of zero. More discussion on the model scores is given in the next section below. We evaluate the predictive performance of these five multi-model ensembles using the quantitative metrics (y_4 - y_7). The evaluation of these five multi-model ensembles serves multiple purposes as described in the results section.

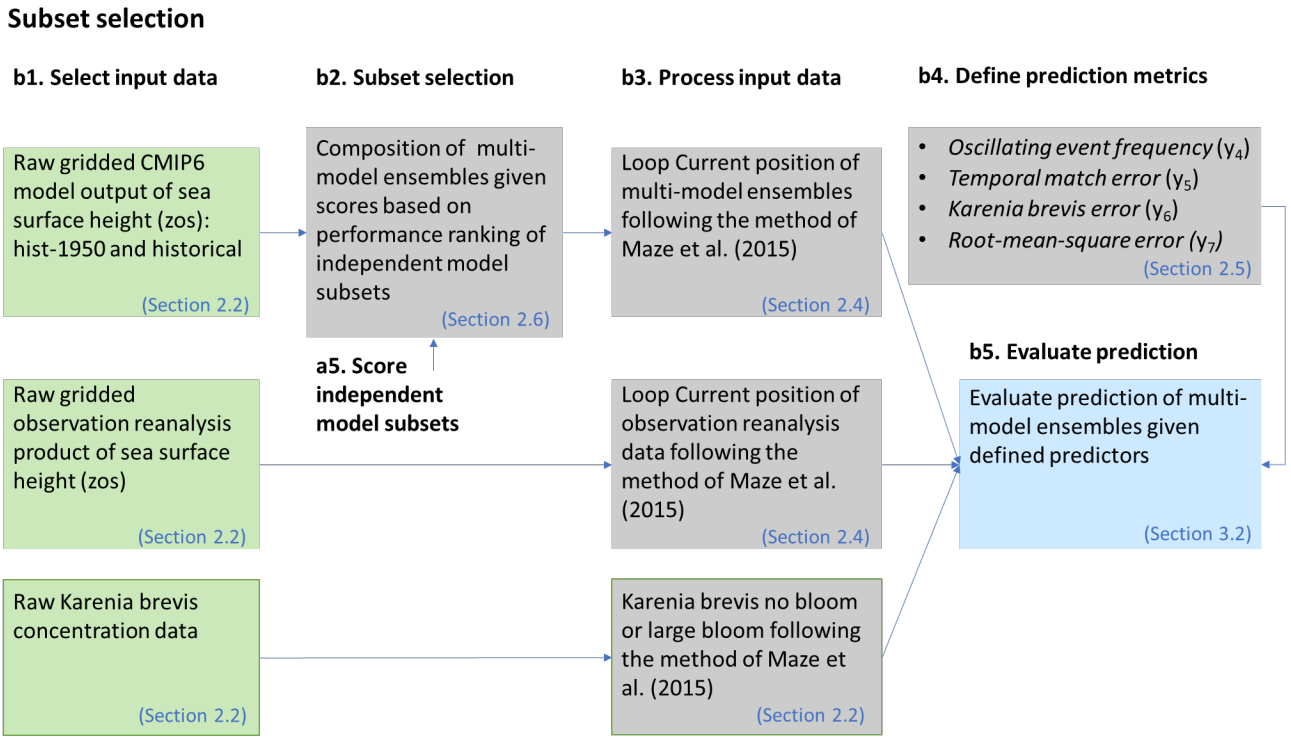

Figure 4. The subset-selection method as implemented in the Colab notebook "Subset selection" (Elshall, 2021) that provides more details and the python code. Details about activity a5 are shown in Fig. 3.

3. Results

3.1 Prescreening

We plot the oscillation of the Loop Current position for each ensemble member (Fig. 5), following the zos data processing steps described in Section 2.4. This is to conduct qualitative comparison between the reanalysis data (Fig. 5a) and the prediction of each ensemble member (Fig.5b-l). Accordingly, we score the ensemble member given its performance with respect to three binary evolving metrics (y_1-y_3). The score is zero if the ensemble member fails to pass all the three metrics. This is the case for E3SM-1-0 of DOE-E3SM-Project (Fig. 5e) and the EC-Earth3P of EC-Earth-Consortium (Fig. 5f). As these two ensemble members do not pass the first metric of physical phenomena simulation (y_1) that is the simulation of the LC-N, then accordingly they score zero in the next two metrics of oscillating event representation (y_2) and oscillating event realism (y_3). This is not unexpected as these two ensemble members are standard-resolution ESMs, which do not have improved process description as the high-resolution ESMs do. The standard-resolution grids EC60to30 of E3SM-1-0 and ORCA1 of EC-Earth3P do not explicitly resolve the mesoscale eddies and boundary currents, but rather require global parametrization of mesoscale eddies. For example, EC60to30 is an eddy closure (EC) grid with global parameterization that is not designed to resolve regional spatial phenomena. On the other hand, with a high horizontal resolution, the eddy-permitting grids such as eORCA12, ORCA12, eORCA025, and ORCA025 (Table 1) can resolve mesoscale eddies, and do not require ocean eddy flux parameterization. For comparison of high- and standard-resolution grid see also Fig.2. On the other hand, the model runs of CESM1-CAM5-SE-HR of NCAR (Fig. 5b) and CNRM-CM6-1-HR of CNRM-CERFACS (Fig. 5d) can simulate LC-N, but without a sign difference of zos at the two segments (Fig. 1a), and accordingly fail in the second metric of oscillating event representation (y_2). These two ensemble members receive a score of one. This score does not indicate that the sea surface

height simulation of these models is poor in general, but rather that these models are unsuitable for this target given the problem definition. The ensemble members of CMCC-CM2-(V)HR4 of CMCC (Fig. 5c), EC-Earth3P-HR of EC-Earth-Consortium (Fig. 5g), and GFDL-CM4/ESM4 of NOAA-GFDL (Fig. 5j), pass the second metric, but fail on the oscillating event realism (y_3). These ensemble members show a higher LC-S frequency than LC-N, which is not consistent with the reanalysis data (Fig. 5a). Accordingly, these three ensemble members receive a score of two. Finally, the ensemble members that pass the three evolving binary metrics and receive a score of three are ECMWF-IFS-HR of ECMWF (Fig. 5h), ECMWF-IFS-MR of ECMWF (Fig. 5i), HadGEM3-GC31-HH/HM of MOHC-NERC (Fig. 5k), and HadGEM3-GC31-MM of MOHC-NERC (Fig. 5l). Visual inspection shows that these four ensemble members are qualitatively similar to the reanalysis data (Fig. 5a) with respect to Loop Current position oscillation.

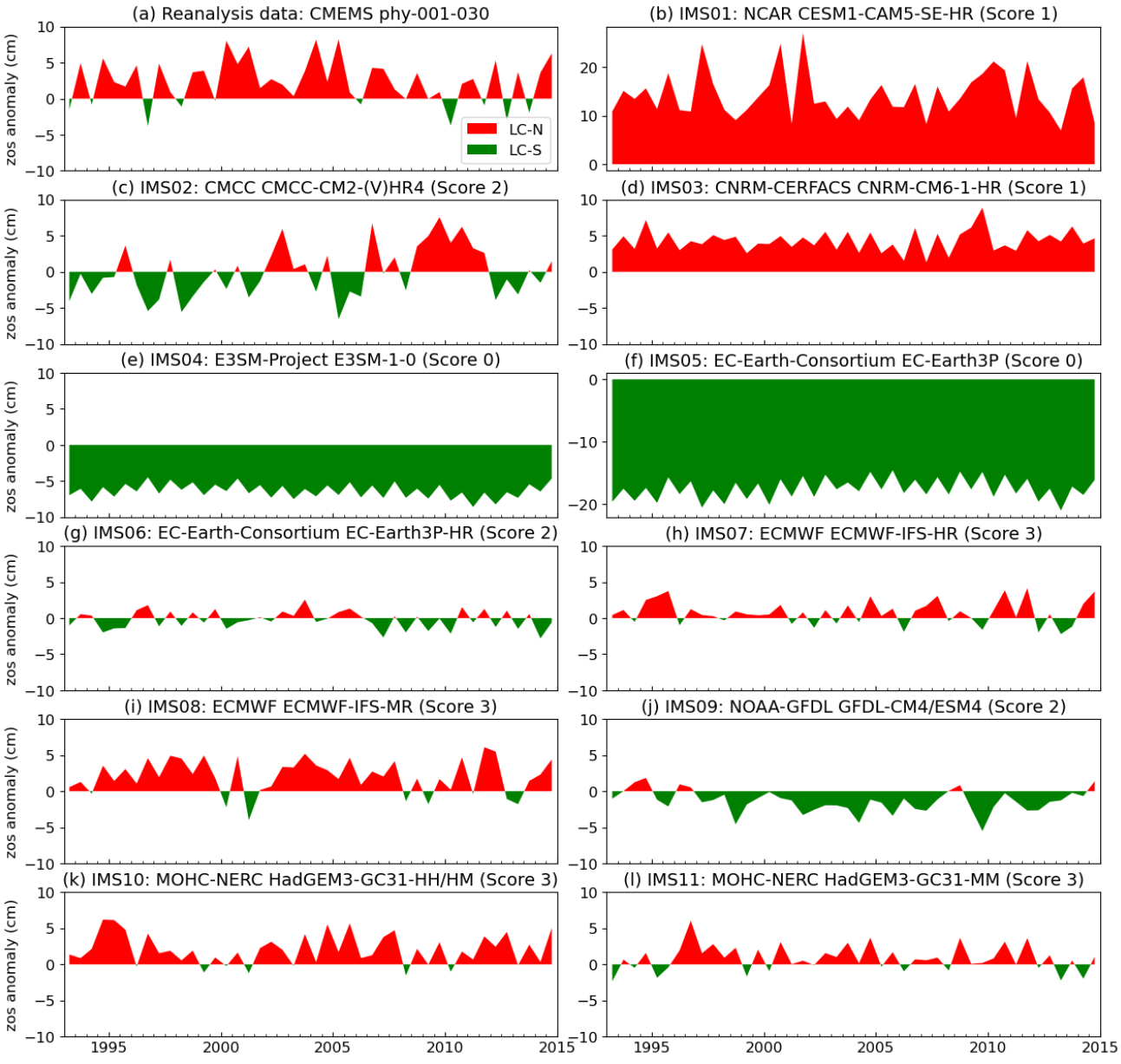

Figure 5. The surface height above geoid (zos) anomaly (Eq.5) of (a) reanalysis data, and (b-l) enesmble members (i.e, indpenednt model subsets). The title of the reanalysis data shows the data provider name, and product ID. The title of ensemble member shows ensemble member number, modeling group name, model name(s), and ensmble member score.

Using metrics $y_4 - y_7$, we evaluate the predictive performance of these 11 ensemble members with respect to reanalysis data as shown in Table 2. According to Maze et al. (2015) there are no red tide blooms for LC-S, and there are either large blooms or no blooms for LC-N. The results of our reanalysis data shown in Table 2 are consistent with Maze et al. (2015) such that none of the 12 intervals of LC-S has large blooms for the study period. Out of the 32 intervals of LC-N, 15 intervals have large blooms. This indicates that LC-N is a necessarily condition for the large bloom to occur and be sustained. Given the reanalysis data, the LC-S frequency is 0.273 for our 22-year study period, which is comparable to Maze et al. (2015), which is 0.267 for their 15-year study period. The ensemble members IMS07, IMS10, IMS11, and IMS08 have the best agreement with the reanalysis data showing LC-S frequencies (y_4) of 0.295, 0.318, 0.205, and 0.182, respectively. These correspond to the oscillating event

frequency errors ($y_{4,err}$) of 0.022, 0.045, -0.068, and -0.091, respectively. Ensemble members that can simulate the oscillation of LC-N and LC-S and have the best temporal match are IMS08, IMS07, IMS10, and IMS11 with temporal match error (y_5) of 27%, 34%, 34%, and 41%, respectively. Given the high-resolution model runs, IMS08, IMS07, IMS10, and IMS11 have the lowest *Karenia brevis* error (y_6) of 0.1, 0.3, 0.3, and 0.3, respectively. IMS09, IMS08, IMS10, IMS03 have the lowest RMSE (y_7) of 3.77, 3.87, 3.88, 4.02, respectively. While no ensemble member is consistently ranked as the top ensemble member given the four metrics, IMS08 is ranked twice as the top ensemble member given the two metrics y_5 and y_6. Thus, this analysis shows that there is no single ensemble member that consistently perform better with respect to all metrics, and that different ensemble members show both over and underestimation of zos anomaly. These two remarks indicate the importance of using a multi-model ensemble.

Table 2. Raw data of Loop Current at North (LC-N) and South (LC-S) positions, and their relation to the occurrence of large blooms for reanalysis data, and each ensemble member (i.e., independent model subset, IMS). The ensemble size is the number of model runs per ensemble member, and the reanalysis data has only one realization. Note given Score(y_1, y_2, y_3) the model receives a score from 0 to 3 for Score(0,0,0), Score(1,0,0), Score(1,1,0), and Score(1,1,1), respectively.

IMS	Ensemble Size	Count		Count LC-N		Count LC-S		Temporal Match			RMSE	Score
		LC-N	LC-S	No-Bloom	Large-Bloom	No-Bloom	Large-Bloom	LC-N	LC-S	Total		
Reanalysis data	1	32	12	17	15	12	0	32	12	44	0	3
IMS01	1	44	0	29	15	0	0	32	0	32	13.16	1
IMS02	2	20	24	14	6	15	9	15	7	22	5.48	2
IMS03	4	44	0	29	15	0	0	32	0	32	4.02	1
IMS04	5	0	44	0	0	29	15	0	12	12	9.27	0
IMS05	3	0	44	0	0	29	15	0	12	12	20.16	0
IMS06	3	20	24	13	7	16	8	13	5	18	4.34	2
IMS07	6	31	13	21	10	8	5	24	5	29	3.77	3
IMS08	3	36	8	22	14	7	1	28	4	32	3.87	3
IMS09	3	8	36	6	2	23	13	5	9	14	5.06	2
IMS10	4	35	9	24	11	5	4	26	3	29	3.88	3
IMS11	7	30	14	20	10	9	5	22	4	26	4.08	3

3.2 Subset selection

There is generally no specific guideline on the composition of multi-model ensemble of ESMs. While composing information from multiple imperfect ensemble members can be an arbitrarily task, the prescreening step can help find subsets that maintain key features of the problem of interest. We first discuss the two ensembles of SME3210 and SME321X. The ensemble SME3210, which includes both high- and standard-resolution model runs, is generally a flawed ensemble composition, since we know from prior existing knowledge of other studies (Caldwell et al., 2019; Hoch et al., 2020) that standard-resolution ESMs are generally incapable of simulating Loop Current. On the other hand, SME321X is the most straightforward ensemble composition that acknowledges prior information, and includes all high-resolution runs that are capable of simulating Loop Current. We consider SME321X as our reference ensemble. Fig. 6 shows the predictive performance of the four multi-model ensembles. Large red tide blooms do not occur for LC-S given reanalysis data (Fig. 6a). Comparing reanalysis data (Fig. 6a) and the multi-model ensembles (Fig.6b-e) shows that ensembles based on prior information (i.e., SME321X, SME32XX, and SME321X) corresponds better to reanalysis data than without accounting for prior information (i.e., SME3210).

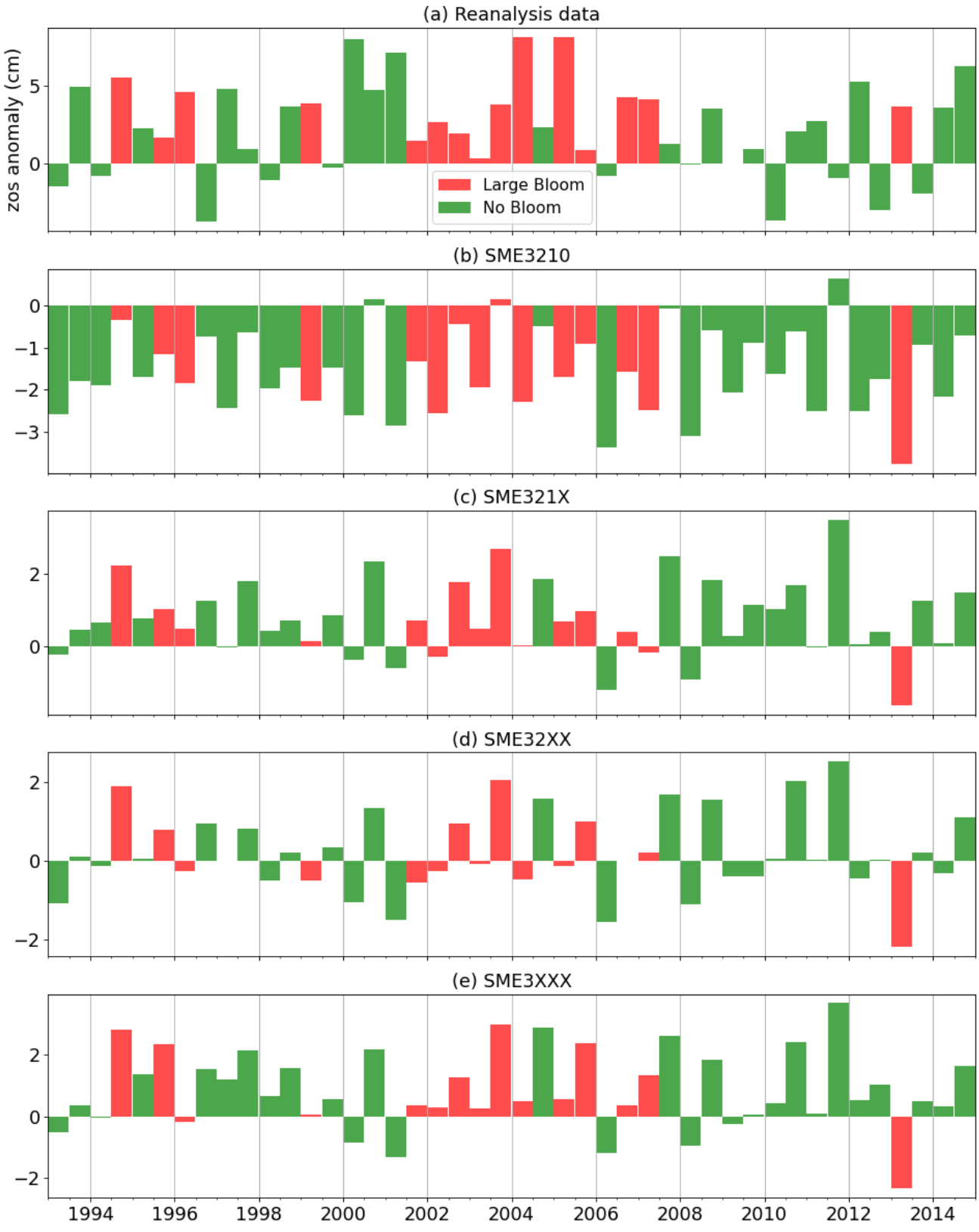

Figure 6. Temporal match of large bloom/no bloom with Loop Current positions given the surface height above geoid (zos) anomaly (Eq.5) of (a) reanalysis data, and (b-e) simulations of four multi-model ensembles. Positive and negative bars indicate Loop Current North (LC-N) and Loop Current South (LC-S), respectively.

Visual examination in Fig.6 is insufficient to understand the impact of prescreening information (i.e., SME32XX and SME3XXX) in comparison to the reference ensemble SME321X without prescreening information, and qualitative metrics are needed. Table 3 quantitatively shows that including standard-resolution model runs (i.e., SME3210) results in prediction degradation with respect to the four qualitative metrics (y_4-y_7). As can be calculated from raw data in Table 3, SME321X shows relatively good agreement with the reanalysis data with a LC-S frequency (y_4) of 0.227, temporal match error (y_5) of 36%, *Karenia brevis* bloom error (y_6) of 20%, and RMSE (y_7) of 3.71.

Table 3. Raw data of Loop Current at North (LC-N) and South (LC-S) positions, and their relation to the occurrence of large blooms simple-average multi-model ensemble (SME). The ensemble size refers to the number of model runs per multi-model ensemble.

SME	Ensemble Size	Count		Count LC-N		Count LC-S		Temporal Match			RMSE
		LC-N	LC-S	No-Bloom	Large-Bloom	No-Bloom	Large-Bloom	LC-N	LC-S	Total	
Reanalysis data	1	32	12	17	15	12	0	32	12	44	0
SME3210	41	3	41	2	1	27	14	2	11	13	5.13
SME321X	33	34	10	22	12	7	3	25	3	28	3.71
SME32XX	28	23	21	17	6	12	9	17	6	23	3.92
SME3XXX	20	35	9	22	13	7	2	28	5	33	3.68
SMEXXX0	8	0	44	0	0	29	15	0	12	12	13.52

Another approach for ensemble composition is to use information from the prescreening step. These are ensembles SME32XX and SME3XXX that exclude the models that cannot represent the oscillation of LC-N and LC-S (y_2). Ensemble SME3XXX only includes model runs with realistic presentation of LC-N and LC-S (y_3). SME32XX shows degraded predictions with respect to the reference ensemble SME321X for all the four quantitative metrics (y_4-y_7). This is not unexpected since members of SME321X show both under and overestimation. For simple model average of model runs with over and underestimation the errors are expected to cancel out (Herger et al., 2018). However, this is not the case for SME3XXX that leverages on most information gained from the prescreening step (i.e., by only including the best members that meet the targets of interest). SME3XXX shows mixed predictive performance with respect to the reference ensemble showing better performance with respect to temporal match error (y_5) of 25% (versus 36% for the reference ensemble), *Karenia brevis* error (y_6) of 13% (versus 20% for the reference ensemble), and RMSE (y_7) of 3.68 (versus 3.71 for the reference ensemble), but inferior performance with respect to LC-S frequency (y_4) of 0.205 (versus 0.273 and 0.227 for the reanalysis data and reference ensemble, respectively). The relatively good performance of SME3XXX is expected, because this ensemble ensures that members with good performance are only included. Table 3 additionally shows the case of SMEXXX0, which only considers standard-resolution runs. SMEXXX0 shows a poor predictive performance with respect to all metrics. We present the SMEXXX0 ensemble to illustrate the breakthrough of the HighResMIP of CMIP6. With respect to sea surface height simulation and regional phenomena, our results clearly show the significant improvement of the high-resolution runs of CMIP6 in comparison to the standard-resolution models that are typical to CMIP5.

3.3 Ensemble composition

Our results show that using prior information is important for ensemble composition, and prescreening-based subset selection can be helpful. Fig.7 summarizes the effect of different ensemble composition criteria. Prior information appears as an important criterion that should be considered as SME3210 has the worst predictive performance with respect to the other ensembles given y_4-y_7. Prescreening-based subset selection seems to relatively improve the predictive performance given y_5-y_7, and slightly degraded performance with respect to y_4. However, pre-screening-based subsect selection has a second conceptual advantage. Given prior information, the first approach of using all the available ensemble

members (i.e., SME321X) is a straightforward choice that can result in error cancellation. The second approach of using information from prescreening results in a reduced size ensemble (i.e., SME3XXX), which maintains the most important ensemble characteristics with respect to the problem of interest. While in the first approach we attempt to maintain a more conservative ensemble, with the second approach we create an ensemble with robust ensemble members. Our results suggest that pre-screening based subset section used to substitute or prior to model weighting, which is a subject of a future research.

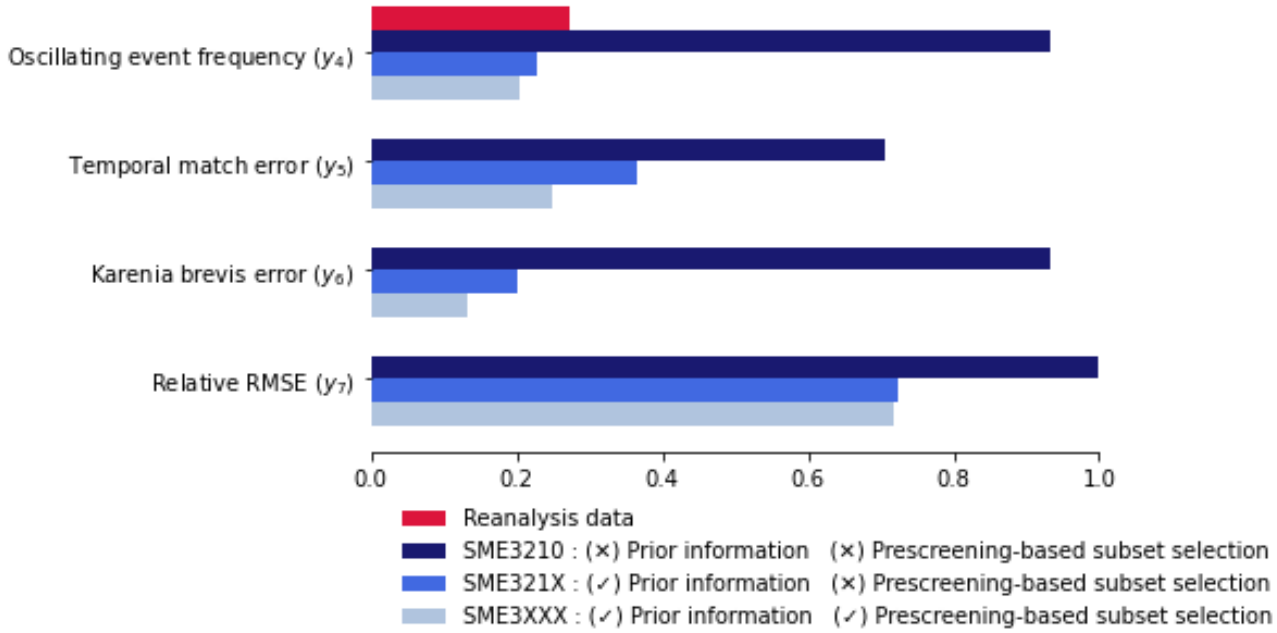

Figure7. Predictive performance (y_4-y_7) given different ensemble composition criteria.

4. Discussion

4.1 Subset selection

To find a robust ensemble that improves the predictive performance of ESMs, this article shows the importance of subset selection based on prior information, prescreening, and process-based evaluation. By evaluating the prescreening-based subset-selection method we deduce two key points as follows. First, we present additional advantages to subset selection that are not well recognized in the literature, which is the importance of subset selection based on process-based evaluation similar to Yun et al. (2017). Eliminating models from an ensemble can be justified if they are known to lack key mechanisms that are indispensable for meaningful climate projections (Weigel et al., 2010). Our study shows that models that cannot simulate the processes of interest based on a prescreening step can be excluded from the ensemble without degrading the ensemble prediction. Second, the selection of subset-selection method depends on the criteria that are relevant for the application in question such that fair comparison can only be made once the various ensemble selection approaches have been tailored to a specific problem (Herger et al., 2018). For example, the process-based evolving binary weights developed in this study is particularly important to eliminate non-representative models. Unlike other subset-selection methods in literature that can be technically challenging to implement, we present a subset-selection method that can be frequently used, as it is intuitive and straightforward to apply. This approach is an addition to subset-selection literature, and is not meant to supersede any of the existing approaches in the literature.

699

700 **4.2 Seasonal prediction limitations**

701 Seasonal prediction of ESMs is generally a challenging task (Hewitt et al., 2017; van den Hurk et al.,
702 2018). While boundaries between timescales can be loose, seasonal prediction is an intermediate
703 timescale between weather forecasting and climate prediction. As such the timescale of seasonal
704 prediction ranges from more than two weeks to slightly longer than one year, which can be
705 differentiated from the sub-seasonal timescale in weather forecasting, and the decadal timescale in
706 climate prediction that goes beyond the first year and up to 30 years (Doblas-Reyes et al., 2013).
707 Although the prediction performance of ESMs is well established for decadal climate prediction
708 (Hewitt et al., 2017), seasonal prediction for some ESMs can suffer from systematic bias that spans
709 over decades resulting in poor temporal agreement between model simulation and observation data
710 (Tokarska et al., 2020). In addition, many fine-scale processes for regional and seasonal prediction
711 (e.g., anomalies of upwelling intensity, temperature, nutrient fluxes, etc.) are sub-grid scale for global
712 models (Jacox et al., 2020). Thus, these processes cannot be resolved by the standard-resolution ESMs,
713 which generally have nominal resolution larger than 25 km. In addition, similar to weather forecasting,
714 seasonal prediction has the challenges of initializing the simulations with a realistic state of the
715 atmosphere as well as other components (i.e., ocean, land and sea ice, and land surface) of the climate
716 system (Doblas-Reyes et al., 2013).

717 Improving seasonal prediction of ESMs to provide useful services for societal decision making is an
718 active research area. Seasonal prediction of ESMs has generally been possible through statistical and
719 dynamical downscaling methods, and other similar techniques such as pattern scaling and use of
720 analogue (van den Hurk et al., 2018). For example, regional ESMs is a dynamical downscaling method
721 that has the same components of global ESMs with boundary conditions from global ESMs or
722 reanalysis data. However, regional ESMs will be subject to biases in global ESMs (Jacox et al., 2020),
723 redistricted compatibility between different model components (Giorgi and Gao, 2018), and large
724 uncertainty in the boundary conditions (Adachi and Tomita, 2020). Another approach to improve
725 seasonal prediction is to refine the resolution of the ESMs, and enhance the model fidelity to resolve
726 fine-scale features and to represent missing processes and feedbacks (Prodhomme et al., 2016; Hewitt
727 et al., 2017). An example of this approach is HigResMIP (Haarsma et al., 2016), which was introduced
728 for the first time in CMIP6 (Eyring et al., 2016). HigResMIP focuses on regional phenomena through
729 using high resolution ESMs to simulate fine-scale processes with more dynamics and less
730 parameterization (Roberts et al., 2018b). While improving the resolution and accordingly model
731 fidelity are important factors for advancing seasonal prediction capability (Roberts et al., 2018b; Deser
732 et al., 2020; Jacox et al., 2020), the prediction system (e.g., subset selection, weighting, bias correction,
733 etc.) can additionally contribute to improving seasonal prediction, as shown in this study using the case
734 study of red tide prediction in Florida based on the high-resolution models of CMIP6.

735 Yet techniques to improve temporal correspondence between predictions and observations at regional
736 scale is needed for climate services in many sectors such as energy, water resources, agriculture, and
737 health (Manzanas et al., 2019). Alternatives to more complex statistical downscaling techniques to
738 improve temporal correspondence include bias correction (Rozante et al., 2014; Oh and Suh, 2017;
739 Wang et al., 2019), ensemble recalibration (Sansom et al., 2016; Manzanas et al., 2019), and
740 postprocessing techniques such as copula-based postprocessing (Li et al., 2020). Bias adjustment can
741 range from simple adjustments in the mean and variance to more complex quantile mapping
742 alternatives, which can adjust higher order moments or probability distribution (Manzanas et al., 2019).
743 Ensemble recalibration methods transform the raw model outputs building on the temporal
744 correspondence between the ensemble mean predictions and the corresponding observations (Sansom
745 et al., 2016). For example, to improve temporal correspondence of seasonal prediction, Manzanas

(2020) use bias correction and recalibration methods to remove mean prediction bias, and intraseasonal biases from drift (i.e., lead-time dependent bias). In this study we used raw outputs without using a postprocessing method to improve temporal correspondence of seasonal prediction. Our results show that the temporal correspondence is not poor, which could be just coincident. Alternatively, this could be attribute to the chosen Loop Current position heuristic with a coarse-temporal-resolution. Accordingly, given a long 6-month period, this is not a month-by-month or season-by-season temporal match, but rather a pseudo-temporal correspondence that captures the general pattern of a dynamic process. Accordingly, using this heuristic relationship, a form of temporal relationship might be possible as long as there is no large drift. If such a temporal correspondence cannot be established for ESMs for Loop Current or other factors that drives the red tide, this would limit the use of the ESMs in terms of providing an early warning system. However, this will not affect the main purpose of the intended model, which is to understand the frequency and trend of red tide under different climate scenarios and estimating the socioeconomic impacts accordingly.

4.3 Limitations and outlook

In this study we present the advantages of subset selection using Loop Current prediction as an example. We show these advantages for the simplest case of using a deterministic analysis, and by considering only historical data. For red tide management purpose, which is to understand the frequency of red tide and the corresponding socioeconomic impacts under different climate scenarios, further steps are needed. First, using CMIP6 model projection data is important to understand the frequency and future trends of red tide under different Shared Socioeconomic Pathways (SSPs) of CMIP6 in which socio-economic scenarios are used to derive emissions scenarios without mitigation (i.e., baseline scenario) and with mitigation (i.e., climate polices). Additionally, CMIP6 data can be readily replaced by high resolution data of Coordinated Regional Downscaling Experiment (CORDEX) as soon as they become available. CORDEX which is driven by the CMIP outputs, provides dynamically downscaled climate change experiments for selected regions (Gutowski Jr. et al., 2016; Gutowski et al., 2020). Second, we need to extend our method to a probabilistic framework that considers both historical and future simulations. As historical assessment criteria are not necessarily informative in terms of the quality of model projections of future climate change, identifying the performance metrics that are most relevant to climate projections is one of the biggest challenges in ESM evaluation (Eyring et al., 2019). As the choice of model is a tradeoff between good performance in the past and projected climate change, selecting only the best performing models may limit the spread of projected climate change (Parding et al., 2020). Exploring such trade-off is warranted in a future study in which a probabilistic framework (e.g., Brunner et al., 2019) is needed to account for model performance, model independence, and the representation of future climate projections. Third, it is imperative to consider not only Loop Current, but also other factors that control red tide such as alongshore and offshore wind speed, African Sahara dust, and atmospheric CO2 concentration need to be considered. These factors can either be directly simulated (e.g., sea surface temperature) or requires a form of postprocessing (e.g., Loop Current as presented in this study). To account for these different factors simultaneously to predict red tide, machine learning is needed similar to the study of Tonelli et al. (2021) that uses CMIP6 data and machine learning to study marine microbial communities under different climate scenarios. Finally, even if the match between model simulation and observation product is not perfect, it should not be a problem because a probabilistic framework can represent these errors when inputting data to the machine learning algorithm. In summary, there are still many further steps needed to develop a probabilistic machine learning framework for regional environmental management of red tide using ESMs of CMIP6 and CORDEX when available. This study is merely a showcase for the potential of using ESMs for red tide management.

5. Conclusions

To improve ensemble performance and to avoid prediction artifacts from including non-representative models, which are models that cannot simulate the process(es) of interest, we introduce a prescreening based subset-selection method. 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. We present a generic subset-selection method to exclude non-representative models based on process-based evolving binary weights. This is the prescreening step that screens each model with respect to its ability to reproduce certain key features. This research emphasizes the importance of ensemble prescreening, which is a topic that is rarely discussed. The presented subset-selection method is flexible as it scores each model given multiple binary criteria. This allows the user to systematically evaluate the sensitivity of the results to different choices of ensemble members. Such flexibility is generally needed to allow the user to understand the implication of ensemble subset selection under different cases (e.g., historic versus historic and future simulations, etc.). Our subset selection method is not meant to replace any of the existing approaches in the literature, but to provide a straightforward and easy-to-implement approach that can be used for many climate services in different sectors as needed.

Data Availability Statement

Data and codes that support the findings of this study are openly available. Elshall (2021) provides and documents the *Karenia brevis* data, CMIP6 model data, CMEMS reanalysis data, and the python codes for data analysis and visualization, which are used in this study.

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. AE, MY, and SK: method development and execution. AE: manuscript development and writing. MY, SK, JH, XY, YW, and MM: manuscript editing and improvements. All authors read and approved the submitted version.

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

Adachi, S. A., and Tomita, H. (2020). Methodology of the Constraint Condition in Dynamical Downscaling for Regional Climate Evaluation: A Review. *Journal of Geophysical Research: Atmospheres* 125, e2019JD032166. doi:https://doi.org/10.1029/2019JD032166.

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., and Chung, E.-S. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences* 23, 4803–4824. doi:https://doi.org/10.5194/hess-23-4803-2019.

Al Samouly, A., Luong, C. N., Li, Z., Smith, S., Baetz, B., and Ghaith, M. (2018). Performance of multi-model ensembles for the simulation of temperature variability over Ontario, Canada. *Environ. Earth Sci.* 77, 524. doi:10.1007/s12665-018-7701-2.

Bartók, B., Tobin, I., Vautard, R., Vrac, M., Jin, X., Levavasseur, G., et al. (2019). A climate projection dataset tailored for the European energy sector. *Climate Services* 16, 100138. doi:10.1016/j.cliser.2019.100138.

Bett, P. E., Thornton, H. E., Lockwood, J. F., Scaife, A. A., Golding, N., Hewitt, C., et al. (2017). Skill and Reliability of Seasonal Forecasts for the Chinese Energy Sector. *J. Appl. Meteorol. Climatol.* 56, 3099–3114. doi:10.1175/JAMC-D-17-0070.1.

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

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.* 14, 124010. doi:10.1088/1748-9326/ab492f.

Caldwell, P. M., Mametjanov, A., Tang, Q., Roekel, L. P. V., Golaz, J.-C., Lin, W., 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. doi:10.1029/2019MS001870.

Cannon, A. J. (2015). Selecting GCM Scenarios that Span the Range of Changes in a Multimodel Ensemble: Application to CMIP5 Climate Extremes Indices. *Journal of Climate* 28, 1260–1267. doi:10.1175/JCLI-D-14-00636.1.

Ceglar, A., Toreti, A., Prodhomme, C., Zampieri, M., Turco, M., and Doblas-Reyes, F. J. (2018). Land-surface initialisation improves seasonal climate prediction skill for maize yield forecast. *Scientific Reports* 8, 1322. doi:10.1038/s41598-018-19586-6.

Chandler, R. E. (2013). Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371, 20120388. doi:10.1098/rsta.2012.0388.

Chang, P., Zhang, S., Danabasoglu, G., Yeager, S. G., Fu, H., Wang, H., 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. doi:https://doi.org/10.1029/2020MS002298.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., 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. doi:https://doi.org/10.1029/2018MS001369.

874 Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M. (2010). Weight
875 assignment in regional climate models. *Climate Research* 44, 179–194. doi:10.3354/cr00916.

876 De Felice, M., Soares, M. B., Alessandri, A., and Troccoli, A. (2019). Scoping the potential
877 usefulness of seasonal climate forecasts for solar power management. *Renew. Energy* 142,
878 215–223. doi:10.1016/j.renene.2019.03.134.

879 DelSole, T., Nattala, J., and Tippett, M. K. (2014). Skill improvement from increased ensemble size
880 and model diversity. *Geophysical Research Letters* 41, 7331–7342.
881 doi:https://doi.org/10.1002/2014GL060133.

882 DelSole, T., Yang, X., and Tippett, M. K. (2013). Is unequal weighting significantly better than equal
883 weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological*
884 *Society* 139, 176–183. doi:https://doi.org/10.1002/qj.1961.

885 Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights
886 from Earth system model initial-condition large ensembles and future prospects. *Nature*
887 *Climate Change* 10, 277–286. doi:10.1038/s41558-020-0731-2.

888 Dixon, A. M., Forster, P. M., and Beger, M. (2021). Coral conservation requires ecological climate-
889 change vulnerability assessments. *Frontiers in Ecology and the Environment* n/a.
890 doi:https://doi.org/10.1002/fee.2312.

891 Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L. (2013).
892 Seasonal climate predictability and forecasting: status and prospects. *WIREs Climate Change*
893 4, 245–268. doi:https://doi.org/10.1002/wcc.217.

894 Drévillon, M., Régnier, C., Lellouche, J.-M., Garric, G., and Bricaud, C. (2018). QUALITY
895 INFORMATION DOCUMENT For Global Ocean Reanalysis Products GLOBAL-
896 REANALYSIS-PHY-001-030. 48.

897 Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., et al. (2015).
898 The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1
899 (v1.0). *Geoscientific Model Development* 8, 261–277. doi:10.5194/gmd-8-261-2015.

900 Elshall, A. S. (2020). Sea surface height above geoid: AVISO altimetry data versus ESM simulations
901 of Loop Current. Available at: https://youtu.be/9Guohel814w [Accessed May 19, 2021].

902 Evans, J. P., Ji, F., Abramowitz, G., and Ekström, M. (2013). Optimally choosing small ensemble
903 members to produce robust climate simulations. *Environ. Res. Lett.* 8, 044050.
904 doi:10.1088/1748-9326/8/4/044050.

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

909 Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019).
910 Taking climate model evaluation to the next level. *Nature Clim Change* 9, 102–110.
911 doi:10.1038/s41558-018-0355-y.

912 Farjad, B., Gupta, A., Sartipizadeh, H., and Cannon, A. J. (2019). A novel approach for selecting
913 extreme climate change scenarios for climate change impact studies. *Science of The Total*
914 *Environment* 678, 476–485. doi:10.1016/j.scitotenv.2019.04.218.

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

917 Fiedler, T., Pitman, A. J., Mackenzie, K., Wood, N., Jakob, C., and Perkins-Kirkpatrick, S. E. (2021).
918 Business risk and the emergence of climate analytics. *Nature Climate Change* 11, 87–94.
919 doi:10.1038/s41558-020-00984-6.

920 FWRI (2020). HAB Monitoring Database. *Florida Fish And Wildlife Conservation Commission*.
921 Available at: http://myfwc.com/research/redtide/monitoring/database/ [Accessed December
922 23, 2020].

923 Giorgi, F., and Gao, X.-J. (2018). Regional earth system modeling: review and future directions.
924 *Atmospheric and Oceanic Science Letters* 11, 189–197.
925 doi:10.1080/16742834.2018.1452520.

926 Golaz, J.-C., Caldwell, P. M., Roekel, L. P. V., Petersen, M. R., Tang, Q., Wolfe, J. D., et al. (2019).
927 The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution.
928 *Journal of Advances in Modeling Earth Systems* 11, 2089–2129.
929 doi:10.1029/2018MS001603.

930 Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., et al. (2016). WCRP
931 COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6.
932 *Geoscientific Model Development* 9, 4087–4095. doi:10.5194/gmd-9-4087-2016.

933 Gutowski, W. J., Ullrich, P. A., Hall, A., Leung, L. R., O'Brien, T. A., Patricola, C. M., et al. (2020).
934 The Ongoing Need for High-Resolution Regional Climate Models: Process Understanding
935 and Stakeholder Information. *Bulletin of the American Meteorological Society* 101, E664–
936 E683. doi:10.1175/BAMS-D-19-0113.1.

937 Haarsma, R., Acosta, M., Bakhshi, R., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., et al. (2020).
938 HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – description, model
939 computational performance and basic validation. *Geoscientific Model Development* 13, 3507–
940 3527. doi:10.5194/gmd-13-3507-2020.

941 Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., et al. (2016). High
942 Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geoscientific*
943 *Model Development* 9, 4185–4208. doi:https://doi.org/10.5194/gmd-9-4185-2016.

944 Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J. (2015). Weighting climate model
945 ensembles for mean and variance estimates. *Clim Dyn* 45, 3169–3181. doi:10.1007/s00382-
946 015-2531-3.

947 Heil, C. A., Dixon, L. K., Hall, E., Garrett, M., Lenes, J. M., O'Neil, J. M., et al. (2014). Blooms of
948 Karenia brevis (Davis) G. Hansen & Ø. Moestrup on the West Florida Shelf: Nutrient sources
949 and potential management strategies based on a multi-year regional study. *Harmful Algae* 38,
950 127–140. doi:10.1016/j.hal.2014.07.016.

951 Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure
952 and Performance of GFDL's CM4.0 Climate Model. *Journal of Advances in Modeling Earth*
953 *Systems* 11, 3691–3727. doi:https://doi.org/10.1029/2019MS001829.

954 Hemri, S., Bhend, J., Liniger, M. A., Manzanas, R., Siegert, S., Stephenson, D. B., et al. (2020). How
955 to create an operational multi-model of seasonal forecasts? *Clim Dyn* 55, 1141–1157.
956 doi:10.1007/s00382-020-05314-2.

957 Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M. (2018).
958 Selecting a climate model subset to optimise key ensemble properties. *Earth System*
959 *Dynamics* 9, 135–151. doi:https://doi.org/10.5194/esd-9-135-2018.

960 Hewitt, H. T., Bell, M. J., Chassignet, E. P., Czaja, A., Ferreira, D., Griffies, S. M., et al. (2017). Will
961 high-resolution global ocean models benefit coupled predictions on short-range to climate
962 timescales? *Ocean Modelling* 120, 120–136. doi:10.1016/j.ocemod.2017.11.002.

963 Hoch, K. E., Petersen, M. R., Brus, S. R., Engwirda, D., Roberts, A. F., Rosa, K. L., et al. (2020).
964 MPAS-Ocean Simulation Quality for Variable-Resolution North American Coastal Meshes.
965 *Journal of Advances in Modeling Earth Systems* 12, e2019MS001848.
966 doi:10.1029/2019MS001848.

967 Hussain, M., Yusof, K. W., Mustafa, M. R. U., Mahmood, R., and Jia, S. (2018). Evaluation of
968 CMIP5 models for projection of future precipitation change in Bornean tropical rainforests.
969 *Theor Appl Climatol* 134, 423–440. doi:10.1007/s00704-017-2284-5.

970 Jacox, M. G., Alexander, M. A., Siedlecki, S., Chen, K., Kwon, Y.-O., Brodie, S., et al. (2020).
971 Seasonal-to-interannual prediction of North American coastal marine ecosystems: Forecast
972 methods, mechanisms of predictability, and priority developments. *Progress in*
973 *Oceanography* 183, 102307. doi:10.1016/j.pocean.2020.102307.

974 Jagannathan, K., Jones, A. D., and Kerr, A. C. (2020). Implications of climate model selection for
975 projections of decision-relevant metrics: A case study of chill hours in California. *Clim. Serv.*
976 18, 100154. doi:10.1016/j.cliser.2020.100154.

977 Jiang, Z., Li, W., Xu, J., and Li, L. (2015). Extreme Precipitation Indices over China in CMIP5
978 Models. Part I: Model Evaluation. *Journal of Climate* 28, 8603–8619. doi:10.1175/JCLI-D-
979 15-0099.1.

980 Knutti, R. (2010). The end of model democracy? *Climatic Change* 102, 395–404.
981 doi:10.1007/s10584-010-9800-2.

982 Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in Combining
983 Projections from Multiple Climate Models. *Journal of Climate* 23, 2739–2758.
984 doi:10.1175/2009JCLI3361.1.

985 Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V. (2017). A
986 climate model projection weighting scheme accounting for performance and interdependence.
987 *Geophysical Research Letters* 44, 1909–1918. doi:https://doi.org/10.1002/2016GL072012.

988 Leduc, M., Laprise, R., Elía, R. de, and Šeparović, L. (2016). Is Institutional Democracy a Good
989 Proxy for Model Independence? *Journal of Climate* 29, 8301–8316. doi:10.1175/JCLI-D-15-
990 0761.1.

991 Li, M., Jin, H., and Brown, J. N. (2020). Making the Output of Seasonal Climate Models More
992 Palatable to Agriculture: A Copula-Based Postprocessing Method. *Journal of Applied*
993 *Meteorology and Climatology* 59, 497–515. doi:10.1175/JAMC-D-19-0093.1.

994 Liu, Y., Weisberg, R. H., Lenes, J. M., Zheng, L., Hubbard, K., and Walsh, J. J. (2016). Offshore
995 forcing on the "pressure point" of the West Florida Shelf: Anomalous upwelling and its
996 influence on harmful algal blooms. *Journal of Geophysical Research: Oceans* 121, 5501–
997 5515. doi:10.1002/2016JC011938.

998 Lledo, L., Torralba, V., Soret, A., Ramon, J., and Doblas-Reyes, F. J. (2019). Seasonal forecasts of
999 wind power generation. *Renew. Energy* 143, 91–100. doi:10.1016/j.renene.2019.04.135.

1000 Lowe, R., Stewart-Ibarra, A. M., Petrova, D., García-Díez, M., Borbor-Cordova, M. J., Mejía, R., et
1001 al. (2017). Climate services for health: predicting the evolution of the 2016 dengue season in
1002 Machala, Ecuador. *The Lancet Planetary Health* 1, e142–e151. doi:10.1016/S2542-
1003 5196(17)30064-5.

1004 Magaña, H. A., and Villareal, T. A. (2006). The effect of environmental factors on the growth rate of
1005 Karenia brevis (Davis) G. Hansen and Moestrup. *Harmful Algae* 5, 192–198.
1006 doi:10.1016/j.hal.2005.07.003.

1007 Manzanas, R. (2020). Assessment of Model Drifts in Seasonal Forecasting: Sensitivity to Ensemble
1008 Size and Implications for Bias Correction. *Journal of Advances in Modeling Earth Systems*
1009 12, e2019MS001751. doi:https://doi.org/10.1029/2019MS001751.

1010 Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., et al. (2019).
1011 Bias adjustment and ensemble recalibration methods for seasonal forecasting: a
1012 comprehensive intercomparison using the C3S dataset. *Clim Dyn* 53, 1287–1305.
1013 doi:10.1007/s00382-019-04640-4.

1014 Maze, G., Olascoaga, M. J., and Brand, L. (2015). Historical analysis of environmental conditions
1015 during Florida Red Tide. *Harmful Algae* 50, 1–7. doi:10.1016/j.hal.2015.10.003.

1016 McSweeney, C. F., Jones, R. G., Lee, R. W., and Rowell, D. P. (2015). Selecting CMIP5 GCMs for
1017 downscaling over multiple regions. *Clim Dyn* 44, 3237–3260. doi:10.1007/s00382-014-2418-
1018 8.

1019 Mendlik, T., and Gobiet, A. (2016). Selecting climate simulations for impact studies based on
1020 multivariate patterns of climate change. *Climatic Change* 135, 381–393. doi:10.1007/s10584-
1021 015-1582-0.

1022 Mishra, N., Prodhomme, C., and Guemas, V. (2019). Multi-model skill assessment of seasonal
1023 temperature and precipitation forecasts over Europe. *Clim Dyn* 52, 4207–4225.
1024 doi:10.1007/s00382-018-4404-z.

1025 Oh, S.-G., and Suh, M.-S. (2017). Comparison of projection skills of deterministic ensemble methods
1026 using pseudo-simulation data generated from multivariate Gaussian distribution. *Theor Appl*
1027 *Climatol* 129, 243–262. doi:10.1007/s00704-016-1782-1.

1028 Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., et
1029 al. (2020). GCMeval - An interactive tool for evaluation and selection of climate model
1030 ensembles. *Clim. Serv.* 18, 100167. doi:10.1016/j.cliser.2020.100167.

1031 Payne, M. R., Hobday, A. J., MacKenzie, B. R., and Tommasi, D. (2019). Editorial: Seasonal-to-
1032 Decadal Prediction of Marine Ecosystems: Opportunities, Approaches, and Applications.
1033 *Front. Mar. Sci.* 6. doi:10.3389/fmars.2019.00100.

1034 Pennell, C., and Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*
1035 24, 2358–2367. doi:10.1175/2010JCLI3814.1.

1036 Perkins, S. (2019). Inner Workings: Ramping up the fight against Florida's red tides. *PNAS* 116,
1037 6510–6512. doi:10.1073/pnas.1902219116.

1038 Prodhomme, C., Batté, L., Massonnet, F., Davini, P., Bellprat, O., Guemas, V., et al. (2016). Benefits
1039 of Increasing the Model Resolution for the Seasonal Forecast Quality in EC-Earth. *Journal of*
1040 *Climate* 29, 9141–9162. doi:10.1175/JCLI-D-16-0117.1.

1041 Roberts, C. D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., and Keeley, S. P. E. (2018a).
1042 Climate model configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS
1043 cycle 43r1) for HighResMIP. *Geoscientific Model Development* 11, 3681–3712.
1044 doi:10.5194/gmd-11-3681-2018.

1045 Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., et al. (2019).
1046 Description of the resolution hierarchy of the global coupled HadGEM3-GC3.1 model as
1047 used in CMIP6 HighResMIP experiments. *Geoscientific Model Development* 12, 4999–5028.
1048 doi:https://doi.org/10.5194/gmd-12-4999-2019.

1049 Roberts, M. J., Vidale, P. L., Senior, C., Hewitt, H. T., Bates, C., Berthou, S., et al. (2018b). The
1050 Benefits of Global High Resolution for Climate Simulation: Process Understanding and the
1051 Enabling of Stakeholder Decisions at the Regional Scale. *Bulletin of the American*
1052 *Meteorological Society* 99, 2341–2359. doi:10.1175/BAMS-D-15-00320.1.

1053 Ross, A. C., and Najjar, R. G. (2019). Evaluation of methods for selecting climate models to simulate
1054 future hydrological change. *Climatic Change* 157, 407–428. doi:10.1007/s10584-019-02512-
1055 8.

1056 Rozante, J. R., Moreira, D. S., Godoy, R. C. M., and Fernandes, A. A. (2014). Multi-model
1057 ensemble: technique and validation. *Geoscientific Model Development* 7, 2333–2343.
1058 doi:https://doi.org/10.5194/gmd-7-2333-2014.

1059 Sanderson, B. M., Knutti, R., and Caldwell, P. (2015). A Representative Democracy to Reduce
1060 Interdependency in a Multimodel Ensemble. *Journal of Climate* 28, 5171–5194.
1061 doi:10.1175/JCLI-D-14-00362.1.

1062 Sanderson, B. M., Wehner, M., and Knutti, R. (2017). Skill and independence weighting for multi-
1063 model assessments. *Geoscientific Model Development* 10, 2379–2395. doi:10.5194/gmd-10-
1064 2379-2017.

1065 Sansom, P. G., Ferro, C. A. T., Stephenson, D. B., Goddard, L., and Mason, S. J. (2016). Best
1066 Practices for Postprocessing Ensemble Climate Forecasts. Part I: Selecting Appropriate
1067 Recalibration Methods. *Journal of Climate* 29, 7247–7264. doi:10.1175/JCLI-D-15-0868.1.

1068 Sorland, S. L., Fischer, A. M., Kotlarski, S., Kunsch, H. R., Liniger, M. A., Rajczak, J., et al. (2020).
1069 CH2018-National climate scenarios for Switzerland: How to construct consistent multi-model
1070 projections from ensembles of opportunity. *Clim. Serv.* 20, 100196.
1071 doi:10.1016/j.cliser.2020.100196.

1072 Sturges, W., and Evans, J. C. (1983). On the variability of the Loop Current in the Gulf of Mexico.
1073 *Journal of Marine Research* 41, 639–653. doi:10.1357/002224083788520487.

1074 Szabó-Takács, B., Farda, A., Skalák, P., and Meitner, J. (2019). Influence of Bias Correction
1075 Methods on Simulated Köppen−Geiger Climate Zones in Europe. *Climate* 7, 18.
1076 doi:10.3390/cli7020018.

1077 Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., et al. (2020). Past
1078 warming trend constrains future warming in CMIP6 models. *Sci Adv* 6.
1079 doi:10.1126/sciadv.aaz9549.

1080 Tonelli, M., Signori, C. N., Bendia, A., Neiva, J., Ferrero, B., Pellizari, V., et al. (2021). Climate
1081 Projections for the Southern Ocean Reveal Impacts in the Marine Microbial Communities
1082 Following Increases in Sea Surface Temperature. *Front. Mar. Sci.* 8, 636226.
1083 doi:10.3389/fmars.2021.636226.

1084 Vajda, A., and Hyvärinen, O. (2020). Development of seasonal climate outlooks for agriculture in
1085 Finland. in *Advances in Science and Research* (Copernicus GmbH), 269–277.
1086 doi:https://doi.org/10.5194/asr-17-269-2020.

1087 van den Hurk, B., Hewitt, C., Jacob, D., Bessembinder, J., Doblas-Reyes, F., and Döscher, R. (2018).
1088 The match between climate services demands and Earth System Models supplies. *Climate*
1089 *Services* 12, 59–63. doi:10.1016/j.cliser.2018.11.002.

1090 Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019).
1091 Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *Journal of Advances in*
1092 *Modeling Earth Systems* 11, 2177–2213. doi:https://doi.org/10.1029/2019MS001683.

1093 Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P. J., et al. (2018). Multimodel
1094 ensembles improve predictions of crop–environment–management interactions. *Global*
1095 *Change Biology* 24, 5072–5083. doi:https://doi.org/10.1111/gcb.14411.

1096 Wang, H.-M., Chen, J., Xu, C.-Y., Chen, H., Guo, S., Xie, P., et al. (2019). Does the weighting of
1097 climate simulations result in a better quantification of hydrological impacts? *Hydrology and*
1098 *Earth System Sciences* 23, 4033–4050. doi:10.5194/hess-23-4033-2019.

1099 Ward, N. D., Megonigal, J. P., Bond-Lamberty, B., Bailey, V. L., Butman, D., Canuel, E. A., et al.
1100 (2020). Representing the function and sensitivity of coastal interfaces in Earth system models.
1101 *Nat. Commun.* 11, 2458. doi:10.1038/s41467-020-16236-2.

1102 Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C. (2010). Risks of Model Weighting in
1103 Multimodel Climate Projections. *Journal of Climate* 23, 4175–4191.
1104 doi:10.1175/2010JCLI3594.1.

1105 Weisberg, R. H., Liu, Y., Lembke, C., Hu, C., Hubbard, K., and Garrett, M. (2019). The Coastal
1106 Ocean Circulation Influence on the 2018 West Florida Shelf K. brevis Red Tide Bloom.
1107 *Journal of Geophysical Research: Oceans* 124, 2501–2512. doi:10.1029/2018JC014887.

1108 Weisberg, R. H., Zheng, L., Liu, Y., Lembke, C., Lenes, J. M., and Walsh, J. J. (2014). Why no red
1109 tide was observed on the West Florida Continental Shelf in 2010. *Harmful Algae* 38, 119–
1110 126. doi:10.1016/j.hal.2014.04.010.

1111 White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., et al. (2017).
1112 Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorol. Appl.* 24, 315–
1113 325. doi:10.1002/met.1654.

1114 Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al.
1115 (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci*
1116 *Data* 3, 160018. doi:10.1038/sdata.2016.18.

1117 Xuan, W., Ma, C., Kang, L., Gu, H., Pan, S., and Xu, Y.-P. (2017). Evaluating historical simulations
1118 of CMIP5 GCMs for key climatic variables in Zhejiang Province, China. *Theor Appl Climatol*
1119 128, 207–222. doi:10.1007/s00704-015-1704-7.

1120 Yun, K., Hsiao, J., Jung, M.-P., Choi, I.-T., Glenn, D. M., Shim, K.-M., et al. (2017). Can a multi-
1121 model ensemble improve phenology predictions for climate change studies? *Ecological*
1122 *Modelling* 362, 54–64. doi:10.1016/j.ecolmodel.2017.08.003.

1123 Zhao, T., Zhang, W., Zhang, Y., Liu, Z., and Chen, X. (2020). Significant spatial patterns from the
1124 GCM seasonal forecasts of global precipitation. *Hydrology and Earth System Sciences* 24, 1–
1125 16. doi:https://doi.org/10.5194/hess-24-1-2020.

1126 Zohdi, E., and Abbaspour, M. (2019). *Harmful algal blooms (red tide): a review of causes, impacts*
1127 *and approaches to monitoring and prediction*. Center for Environmental and Energy
1128 Research and Studies doi:10.1007/s13762-018-2108-x.

1129 Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., et
1130 al. (2018). Future climate risk from compound events. *Nature Clim Change* 8, 469–477.
1131 doi:10.1038/s41558-018-0156-3.

1132