Performance evaluation of 22 CMIP6 models on the representation of sea surface and bottom temperatures on the northern North American Shelves in the North Atlantic, Arctic, and North Pacific Oceans

Ahmadreza Alleosfour, Zeliang Wang, Brendan DeTracey, Blair Greenan, David Brickman, Chantelle Layton, Peter S. Galbraith, Frédéric Cyr, Nadja Steiner, James Christian

Fisheries and Oceans Canada
Bedford Institute of Oceanography
1 Challenger Drive
Dartmouth, Nova Scotia, B2Y 4A2

2026

Canadian Technical Report of
Hydrography and Ocean Sciences 411

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of Hydrography and Ocean Sciences

Technical reports contain scientific and technical information of a type that represents a contribution to existing knowledge, but which is not normally found in the primary literature. The subject matter is generally related to programs and interests of the Oceans and Science sectors of Fisheries and Oceans Canada.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts.*

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Regional and headquarters establishments of Ocean Science and Surveys ceased publication of their various report series as of December 1981. A complete listing of these publications and the last number issued under each title are published in the *Canadian Journal of Fisheries and Aquatic Sciences*, Volume 38: Index to Publications 1981. The current series began with Report Number 1 in January 1982.

Rapport technique canadien sur l'hydrographie et les sciences océaniques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles mais que l'on ne trouve pas normalement dans les revues scientifiques. Le sujet est généralement rattaché aux programmes et intérêts des secteurs des Océans et des Sciences de Pêches et Océans Canada.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques*.

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page de titre.

Les établissements de l'ancien secteur des Sciences et Levés océaniques dans les régions et à l'administration centrale ont cessé de publier leurs diverses séries de rapports en décembre 1981. Vous trouverez dans l'index des publications du volume 38 du *Journal canadien des sciences halieutiques et aquatiques*, la liste de ces publications ainsi que le dernier numéro paru dans chaque catégorie. La nouvelle série a commencé avec la publication du rapport numéro 1 en janvier 1982.

Canadian Technical Report of
Hydrography and Ocean Sciences 411

2026

PERFORMANCE EVALUATION OF 22 CMIP6 MODELS ON THE REPRESENTATION OF SEA SURFACE AND BOTTOM TEMPERATURES ON THE NORTHERN NORTH AMERICAN SHELVES IN THE NORTH ATLANTIC, ARCTIC, AND NORTH PACIFIC OCEANS

by

Ahmadreza Alleosfour[1], Zeliang Wang[1], Brendan DeTracey[1], Blair Greenan[1], David Brickman[1], Chantelle Layton[1], Peter S. Galbraith[2], Frédéric Cyr[3], Nadja Steiner[4], James Christian[4]

[1]Fisheries and Oceans Canada
Bedford Institute of Oceanography
1 Challenger Drive
Dartmouth, Nova Scotia, B2Y 4A2

[2]Fisheries and Oceans Canada
Maurice Lamontagne Institute
Mont-Joli, Québec, G5H 3Z4

[3]Fisheries and Oceans Canada
Northwest Atlantic Fisheries Center
80 East White Hills Road
St. John's, Newfoundland and Labrador, A1C 5X1

[4]Fisheries and Oceans Canada
Institute of Ocean Sciences
9860 West Saanich Road
Sidney, British Columbia, V8L 4B2

Cat. No. Fs 97-18/411E-PDF ISBN 978-0-660-98428-5 ISSN 1488-5417

Correct citation for this publication:

Alleosfour, A., Wang, Z., DeTracey, B., Greenan, B., Brickman, D., Layton, C., Galbraith, P.S., Cyr, F., Steiner, N., and Christian, J. 2026. Performance evaluation of 22 CMIP6 models on the representation of sea surface and bottom temperatures on the northern North American Shelves in the North Atlantic, Arctic, and North Pacific Oceans. Can. Tech. Rep. Hydrogr. Ocean. Sci. 411: ix + 69 p.

Table of Contents

List of Tables

List of Figures

ABSTRACT

Alleosfour, A., Wang, Z., DeTracey, B., Greenan, B., Brickman, D., Layton, C., Galbraith, P.S., Cyr, F., Steiner, N., and Christian, J. 2026. Performance evaluation of 22 CMIP6 models on the representation of sea surface and bottom temperatures on the northern North American Shelves in the North Atlantic, Arctic, and North Pacific Oceans. Can. Tech. Rep. Hydrogr. Ocean. Sci. 411: ix + 69 p.

The performance of 22 CMIP6 Earth System Models (ESM) was evaluated for both sea surface and bottom temperature for northern North American shelf waters (i.e., Canadian Shelves). For the ocean surface, the HadISST sea surface temperature product was used to evaluate the models. In order to address the scarcity of sub-surface *in-situ* observations, an ocean reanalysis product, GLORYS12, was used to validate the CMIP6 models for the bottom temperature. *In-situ* observations were available for some parts of the Canadian shelves, and they were used to complement the GLORYS12 product. The shelf waters were divided into eighteen regions for evaluation. Three statistical measures: bias, standard deviation, and trend, were introduced to examine model performance. They were aggregated together using a modified version of the Kling-Gupta efficiency score to rank these models. Our analysis suggests that in general, the evaluated CMIP6 models have inconsistent performance for the sea surface temperature and bottom temperature, while some models demonstrate better performance than others. Considering both sea surface and bottom temperatures, the top model for both the North Atlantic and Pacific is CNRM-CM6-1-HR, while for the Arctic, MPI-ESM1-2-LR and CESM2-WACCM share the top rank.

RÉSUMÉ

Alleosfour, A., Wang, Z., DeTracey, B., Greenan, B., Brickman, D., Layton, C., Galbraith, P.S., Cyr, F., Steiner, N., and Christian, J. 2026. Performance evaluation of 22 CMIP6 models on the representation of sea surface and bottom temperatures on the northern North American Shelves in the North Atlantic, Arctic, and North Pacific Oceans. Can. Tech. Rep. Hydrogr. Ocean. Sci. 411: ix + 69 p.

Le rendement de 22 modèles du système terrestre (MST) de la phase 6 du projet d'intercomparaison de modèles couplés (CIMP6) a été évalué concernant la représentation des températures de la surface et du fond de la mer dans les eaux de la partie nord des plateaux nord-américains (c.-à-d. les eaux des plateaux canadiens). En ce qui concerne la température de la surface de la mer, l'évaluation des modèles a été réalisée au moyen des données de l'ensemble du Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST). Afin de tenir compte de la rareté des observations *in situ* effectuées sous la surface, le produit de réanalyse océanique GLORYS12 a été utilisé pour la validation des valeurs de température du fond issues des modèles CMIP6. Des observations *in situ* étaient disponibles pour certaines zones des plateaux canadiens; elles ont été utilisées pour compléter le produit GLORYS12. Les eaux des plateaux canadiens ont été divisées en dix-huit régions aux fins d'évaluation. Trois mesures statistiques, soit le biais, l'écart-type et la tendance, ont été introduites aux fins d'examen du rendement des modèles. Elles ont été regroupées à l'aide d'une version modifiée de la note d'efficience de Kling-Gupta aux fins de classement des modèles. Notre analyse indique qu'en général, le rendement des modèles CMIP6 évalués est variable pour la température de la surface et la température du fond de la mer, mais certains modèles affichent un meilleur rendement que d'autres. Si les températures de la surface et du fond de la mer sont considérées, le meilleur modèle pour l'Atlantique Nord et le Pacifique est le CNRM-CM6-1-HR, tandis que pour l'Arctique, le MPI-ESM1-2-LR et le CESM2-WACCM se partagent le premier rang.

1 Introduction

The ocean covers approximately 71% of the Earth's surface, and it plays a key role in climate change. It has been undergoing significant changes due to anthropogenic greenhouse gas (GHG) emissions. What the future ocean conditions will be is a question people must ask and needs to be answered. Climate models are the only tool that can project the future states of the Earth. In the 1990s, the World Climate Research Program (WCRP) promoted a set of experiments known as the Coupled Model Intercomparison Project (CMIP), aiming at better understanding past climate changes and making projections and uncertainty estimates about the future (e.g., Meehl et al. (2000); Taylor et al. (2012)). Significant efforts have been made to improve global climate model simulations. The sixth phase of CMIP (CMIP6; Eyring et al. (2015)) is the latest modeling effort for simulating and projecting various aspects of climate change for which a new set of scenarios has been developed. The previous version of CMIP, CMIP5, used Representative Concentration Pathways (RCP) to represent greenhouse gas concentration trajectories. In contrast, the new scenarios in CMIP6 represent different socio-economic developments as well as different pathways of atmospheric greenhouse gas concentrations. They are projections of new sets of emissions and land-use scenarios based on Shared Socio-economic Pathways (SSP; Riahi et al. (2017)).

In order to investigate how well the CMIP6 models can represent ocean changes on large and local scales, numerous studies have examined the performance of CMIP6 models on global and regional scales (e.g., Koelling et al. (2023); Liu et al. (2022); Wang et al. (2022)). CMIP6 models have been reported to realistically reproduce mean and extreme climates compared to observations. However, few studies have focused on the representation of shelf waters by the CMIP6 models. To our knowledge, only Wang et al. (2023) evaluated the CMIP6 model performance on the representation of the Gulf of Maine and Scotian Shelf waters. Coastal and shelf waters support extensive and productive fisheries. Hydrographic changes in these waters are known to influence regional ecosystem dynamics and fisheries (e.g., Greenan et al. (2019); Stanley et al. (2018); Wang et al. (2020)), and can have both direct and indirect influences on fish populations (Loder & Wang, 2015).

To rank the performance of CMIP6 models, the models are evaluated against ocean temperature, which is a crucial component of the Earth's climate system. Ocean surface temperature plays a significant role in moderating the weather and climate of the surrounding area through the interaction of ocean-land-atmosphere (Deng et al. 2023; Griffies et al. 2015) and improves our understanding of the interactions between the atmosphere and the ocean (Bayr et al. 2019). There are several Sea Surface Temperature (SST) products available that may be used in the evaluation of the SST of CMIP6 models. Loder and Wang

(2015) evaluated SST products using SST data from several stations with long time coverages for the Northwest Atlantic Ocean and found that the general trends from these products are mostly consistent, while differences were evident in biases. Since the majority of CMIP6 models are coarse resolution models with resolutions close to 1°, the 1° product of HadISST is selected to evaluate the modeled SSTs. Observational bottom temperature data are available in some regions for some seasons, e.g. summer and/or fall, though generally sparse. The GLORYS12 well represents observed bottom temperatures in the North Atlantic, Arctic, and North Pacific shelves, in terms of long-term trend, bias, and standard deviation (McKee et al. 2023).

In this report, we will evaluate the performance of the CMIP6 models in terms of the representation of sea surface and bottom temperatures of the northern North American shelves, including all Canadian shelves and the whole Gulf of Maine, the southern Chukchi Sea, the Bering Sea, and the Alaskan Shelf.

2 Data Source and Methodology

This report focuses on eighteen regions in the North Atlantic, Arctic, and Pacific oceans (Figure 1 to 6). The SST evaluation covers years 1955 to 2014, since 2014 is the last year of the CMIP6 historical simulation, and the Bottom Temperature (BT) evaluation against GLORYS12 covers years 1993 to 2014, as 1993 is the first year of the GLORYS12 product. Comparison against *in-situ* BT observations was limited to some of the Atlantic and Pacific regions, and included years prior to 1993. An extra region in the Eastern Bering Sea (EBS) (Figure 5) and three available stations, GAK1 (Gulf of Alaska), NV (Northern Vancouver Island), and SV (Southern Vancouver Island) (Figure 6) were used for BT comparison for the North Pacific. Regarding SST, all months were included in calculating the annual mean including the ice-covered seasons. A list of all regions and stations and their abbreviations are provided in **Error! Reference source not found.**. The bathymetry from ETOPO1 (Amante & Eakins, 2009) was used to define these regions. The bathymetric depth that defined a region edge was chosen based on the shelf slope and the complexity of the resulting bathymetric contour. For the Gulf of Maine and Scotian Shelf, a depth of 200 m, on the Newfoundland and Labrador shelves a depth of 400 m, and a depth of 600 m were used for all Arctic and Pacific regions. The whole basins were used for the Gulf of St. Laurence, Hudson Bay, and Baffin Bay. Note that there were areas with depth >600 m in the Gulf of St. Lawrence and Baffin Bay, however, these areas were excluded for simplicity and were small compared to the whole basins. The majority of the CMIP6 models have coarse resolutions (~ 1° × 1° in longitude and latitude), and most of the deep channels/areas are not resolved. Where the shelf region was wide, the models had similar regionally-averaged depth with

a small standard deviation, while for narrow shelves with a steep continental slope, some models deviated from the average depth among other models. The regionally-averaged depth of each model and its standard deviation for each region are provided in Appendix S 41 while cell counts are provided in Appendix S 1.

For the evaluation of SST, the HadISST data (Rayner et al. 2003), a 1° SST product with global ocean coverage was used. This product has been widely used for climate monitoring (Folland et al. 2001; Liu et al. 2022; Rickard et al. 2023; Wang et al. 2022; Yang et al. 2020; Yu et al. 2023) and model validation (Gregory et al. 2002). Due to the paucity of observed bottom temperature data, the high-resolution (1/12°) ocean product GLORYS12 was used for the evaluation of BT. GLORYS12 has demonstrated good agreement with available observational data (McKee et al. 2023) on the Northeast U.S. continental shelf in comparison to eight widely used reanalysis products (Castillo-Trujillo et al. 2023), and also along the US west coast continental shelf (Amaya et al. 2023). In the Arctic, where data sampling is inconsistent/scattered, McKee et al. (2023) compared GLORYS12 bottom temperature against four stations in the Chukchi Sea and showed that GLORYS12 has a warm bias but captured the variance in the seasonal cycle at three stations, and cold bias and weak seasonal cycle at the fourth station. In the North Atlantic, GLORYS12 comparison against the observation has shown a good fit to the observations, having a small bias (warmer in the Eastern Scotian Shelf, cooler in the central and western regions), except in the Gulf of St. Lawrence. In the Pacific, GLORYS12 has a good correlation with observations (at 0.97), having similar trend direction as the observations and almost the same magnitude, while it showed a smaller standard deviation than observation (McKee et al. 2023).

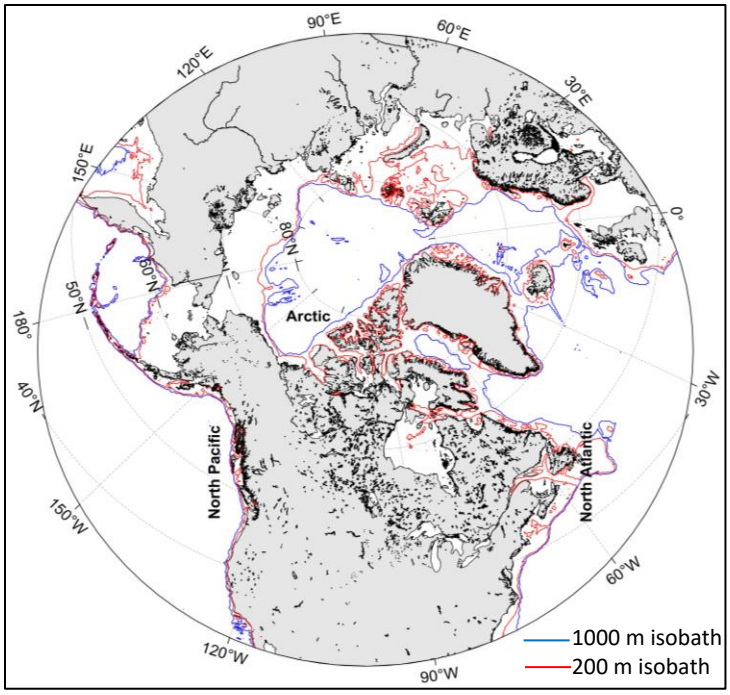

Figure 1. Map of the regions of interest in the North Atlantic, Arctic, and North Pacific. The red and blue contours are the 200 and 1000 m isobaths, respectively.

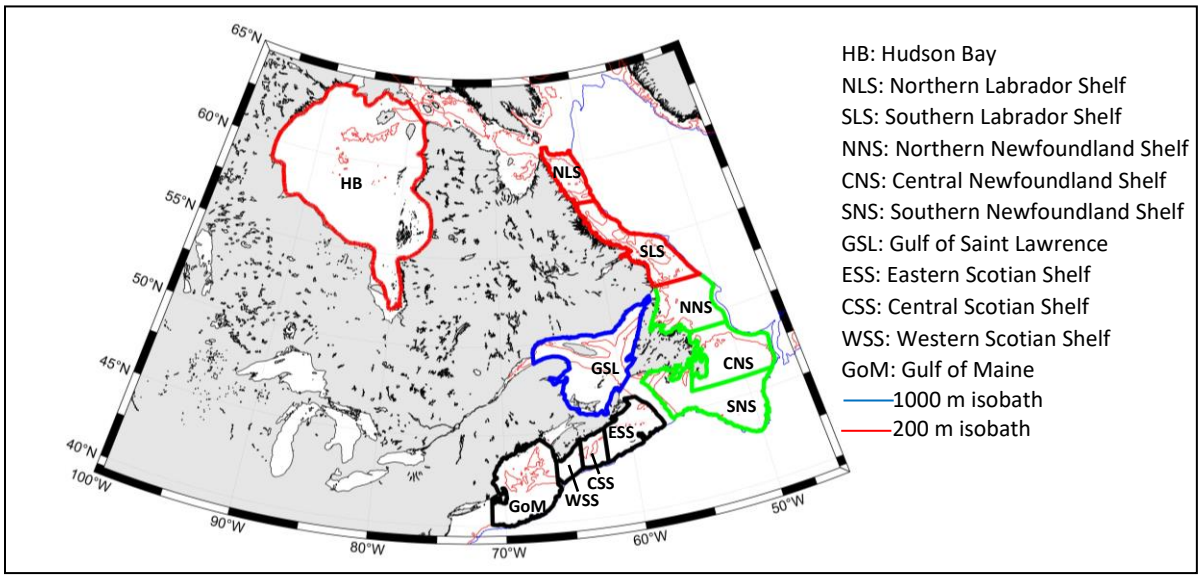

Figure 2. Regions in the North Atlantic including Gulf of Maine (GoM), Western Scotian Shelf (WSS), Central Scotian Shelf (CSS), Eastern Scotian Shelf (ESS), Gulf of Saint Lawrence (GSL), South Newfoundland Shelf (SNS), Central Newfoundland Shelf (CNS), Northern Newfoundland Shelf (NNS), Southern Labrador Shelf (SLS), Northern Labrador Shelf (NLS), and Hudson Bay (HB). The thick colored lines represent the boundary of each region. The red and blue contours are the 200 and 1000 m isobaths, respectively.

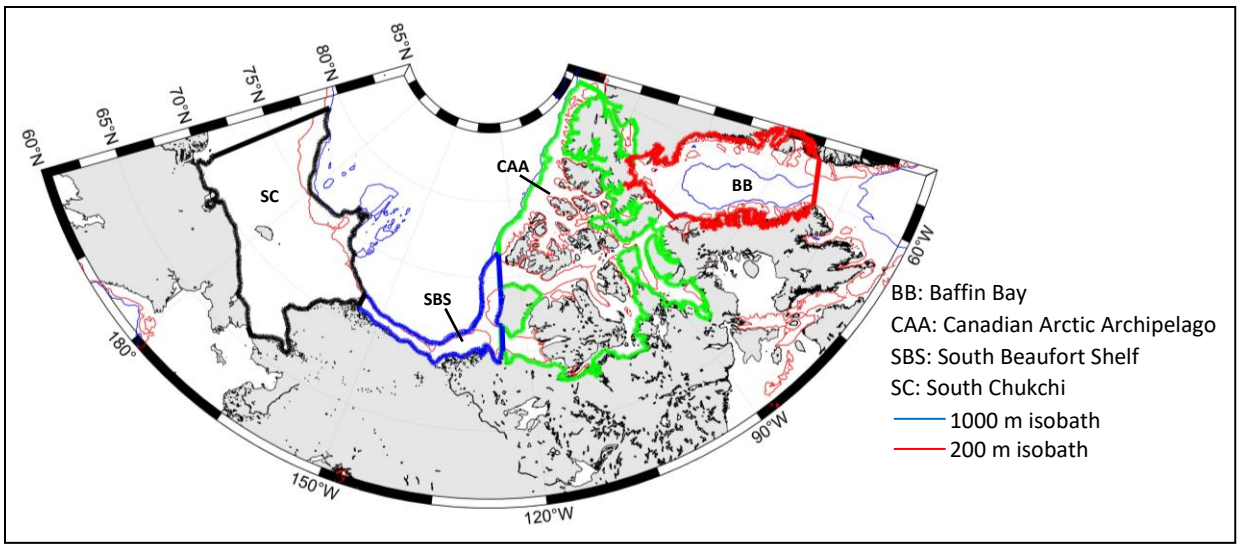

Figure 3. Regions in the Arctic include Baffin Bay (BB), the Canadian Arctic Archipelago (CAA), the Southern Beaufort Sea (SBS), and Southern Chukchi (SC). The thick colored lines represent the boundary of each region. The red and blue contours are the 200 and 1000 m isobaths, respectively.

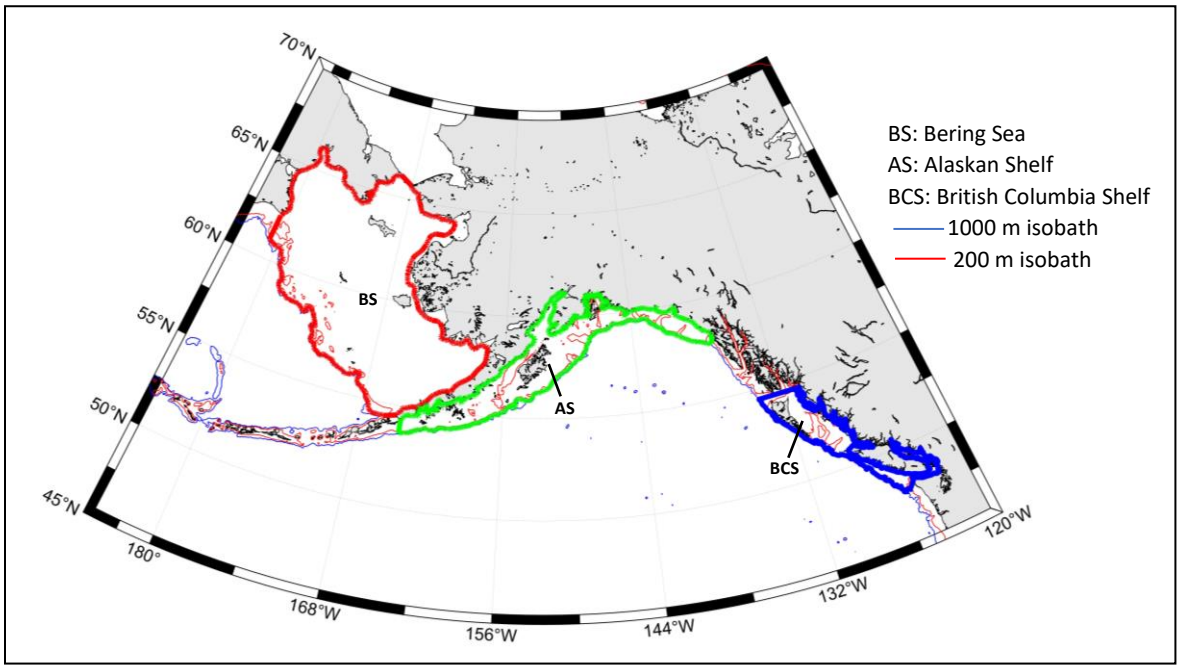

Figure 4. Regions in the North Pacific include the Bering Sea (BS), Alaska Shelf (AS), and British Columbia Shelf (BCS). The thick colored lines represent the boundary of each region. The red and blue contours are the 200 and 1000 m isobaths, respectively.

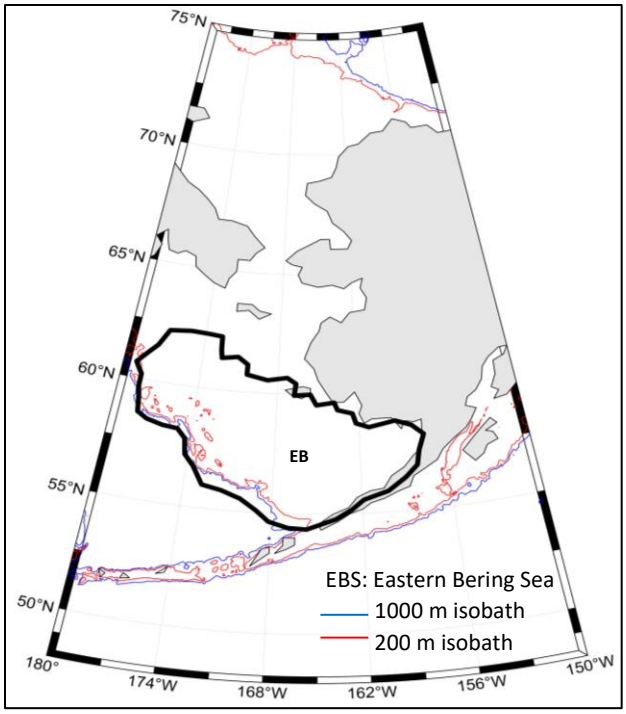

Figure 5. Eastern Bering Sea region. The thick black line represents the boundary of the region. The red and blue contours are the 200 and 1000 m isobaths, respectively.

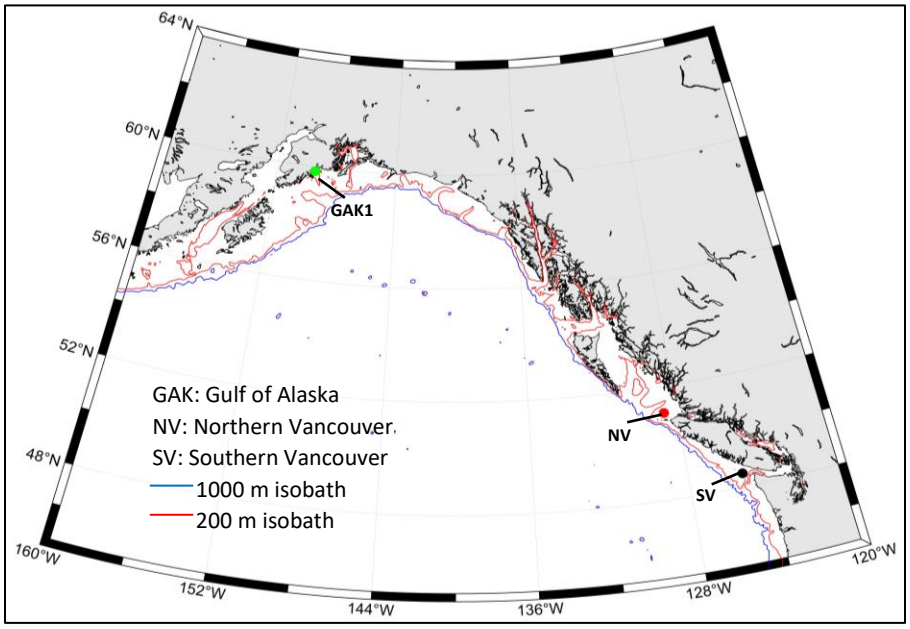

Figure 6. Gulf of Alaska (GAK1), Northern Vancouver Island (NV), and Southern Vancouver Island (SV). The red and blue contours are the 200 and 1000 m isobaths, respectively.

Table 1.Region and station names used in this report and their abbreviations.

Regions				
North Atlantic				
GoM	Gulf of Maine		CNS	Center Newfoundland Shelf
WSS	West Scotian Shelf		NNS	North Newfoundland Shelf
CSS	Centre Scotian shelf		SLS	South Labrador Shelf
ESS	East Scotian Shelf		NLS	North Labrador Shelf
GSL	Gulf of St. Lawrence		HB	Hudson Bay
SNS	South Newfoundland Shelf			
Arctic				
BB	Baffin Bay		SBS	South Beaufort Sea
CAA	Canadian Arctic Archipelago		SC	South Chukchi
North Pacific				
BS	Bering Sea		AS	Alaska Shelf
EBS	East Bering Sea		BCS	British Columbia Shelf
Stations				
GAK1	Gulf of Alaska Mooring		SV	South Vancouver
NV	North Vancouver			

2.1 Sea Surface Temperature

2.1.1 HadISST

The Met Office Hadley Centre dataset provides monthly gridded SST and ice cover at a horizontal resolution of 1° by 1° over the period 1871 to the present day. A two-stage reduced-space optimal interpolation is used for constructing the SST data. The impact of the surface melt effects on the retrievals (for the sea ice field) is compensated by using a satellite microwave-based sea ice concentration. A statistical relationship between SST and sea ice concentration is used to estimate the temperature near sea ice. To adjust the bias, HadISST uses data from the satellite-borne Advanced Very High-Resolution Radiometer (AVHRR), but this correction only started in 1981 when the AVHRR data begin. In this report, the monthly SST data from 1955 to 2014 is used. This dataset has shown the ability to resolve the large-scale variability in the ocean, for instant in the North Pacific (Chelton & Risien, 2016). More information on HadISST can be found in Rayner et al. (2003) and Chelton et al. (1998).

2.2 Bottom Temperature

2.2.1 GLORYS12V1

GLORYS12 is a global eddy-resolving physical ocean and sea ice reanalysis with 1/12° horizontal resolution, fifty vertical layers with high resolution in the surface layers, covering the period 1993 to the present (Lellouche et al. 2021). The core model component is the NEMO platform (Madec et al. 2019) and the prescribed surface forcing is from the European Centre for Medium-Range Weather Forecast (ECMWF) ERA-interim reanalysis (Dee et al. 2011). A reduced-order Kalman filter is used to assimilate the altimeter

sea level anomaly, satellite sea surface temperature, sea ice concentration, and *in situ* temperature and salinity profiles. It can capture the small-scale variability of surface dynamics due to its high resolution and is a reliable source for detecting climate variability and seasonal cycles both on global and regional scales. For more details about the GLORYS12, please refer to Lellouche et al. (2021).

2.2.2 *In Situ* Observations

Historical observations of BT data are available for the Scotian, Newfoundland, Labrador, and the North Pacific shelves (Table 2) and the dataset is the same as McKee et al. (2023), where they evaluated the GLORYS12 against this dataset. These additional data were also used in the model evaluation process.

Table 2. Available time period of observations in some of the North Atlantic and Pacific regions.

July	Spring	Fall	Summer	Yearly mean
ESS (1970-2014)	CNS (1980-2014)	CNS (1980-2014)	EBS (1982-2014)	GAK1 (1987-2014)
CSS (1970-2014)	SNS (1980-2014)	SNS (1980-2014)	-	-
WSS (1970-2014)	NV (1998-2014)	NLS (1980-2014)	-	-
-	SV (1985-2014)	SLS (1980-2014)	-	-
-	BCS (1998-2014)	NNS (1980-2014)	-	-
-	-	NV (1998-2014)	-	-
-	-	SV (1985-2014)	-	-

2.2.2.1 Scotian Shelf

The *in-situ* observations were collected by the DFO Maritimes Region Ecosystem Survey on the Scotian Shelf conducted every year since 1970 during July and August. This is the only dataset having complete spatial coverage in that area (Claytor et al. 2014; Hebert et al. 2021). The Conductivity, Temperature and Depth (CTD) profiler data is collected as part of an ecosystem trawl survey to support fish stock assessments and is based on a depth-stratified random design (ICES, 2004,2005). Measurements cover the entire water column within 5 m of the bottom depth. The data are interpolated to a 0.2° by 0.2° grid (Hebert et al. 2023). In this report, the 1970-2014 period is used, and spatial averages are calculated for each region.

2.2.2.2 Newfoundland and Labrador Shelves

The gridded bottom temperature used for the Newfoundland and Labrador shelves is from Cyr et al. (2019). These data are derived using a combination of all available profiles in the area including DFO surveys and hydrographic missions, international oceanographic campaigns and Argo float profile data

obtained through the Canadian Atlantic Shelf Temperature-Salinity (CASTS) dataset (Coyne et al. 2023). The data are bin-averaged to a regular 0.1° by 0.1° grid for each season (April-June for spring and September-December for fall) to get one seasonal profile for each grid cell, before being linearly interpolated horizontally to fill the gaps. For the bottom temperature, the closest value to the GEBCO-2014 grid bathymetry (version 20141103), to a maximum of 50-meter difference, is picked for each grid point. Also, any observation deeper than 1000 meter was removed.

2.2.2.3 North Pacific Shelf

The bottom temperature dataset is from the National Oceanic and Atmospheric Administration (NOAA) Fisheries campaign which aims to determine the abundance and distribution of various bottom fish species in the Eastern Bering Sea. It started in 1982 and has increased the number of stations over time. The survey takes place each summer for two to four months, mostly in June and July. From 1982-1989, they used expendable bathythermographs to collect data; after 1989 digital bathythermographs attached to bottom trawl nets were used. The final product is an average of bottom temperature data by station and year. For more information on data collection and pre-processing see Buckley et al. (2009) and Lauth et al. (2019), and for the dataset see Kearney (2021).

2.2.2.4 North Pacific Stations

On the Pacific shelf, three stations (Figure 6) are used for the bottom temperature comparison against the models. The first one is the Gulf of Alaska GAK1 station at the mouth of Resurrection Bay (49.845°N, 149.466°W), which is the longest shelf time series in the North Pacific (1970 to present); sampling from 1970 to 1990 was mostly by ships-of-opportunity, and from 1990 onward monthly CTD profiles to within 10 m of the bottom were collected by the University of Alaska. The second and third stations (Figure 6) are Northern Vancouver Island (NV) at 51.06°N, 128.7°W and Southern Vancouver Island (SV) at 48.48°N, 125.25°W. The dataset comes from the Canadian Integrated Ocean Observing System (CIOOS) (Liu et al. 2022). The data were measured by CTDs deployed by DFO during scientific surveys from 1965 to the present and are divided into spring and fall.

2.3 Methods

This study focuses on the assessment of a 60-year historical period (1955-2014) of 22 CMIP6 models in capturing the long-term changes in the sea surface temperature and bottom temperature. Henceforth 1955-2014 was used in the SST evaluation, and 1993-2014 for the BT evaluation. As mentioned previously, some long-time observed BT data were available and they were used in the evaluation as well: Scotian Shelf, 1970-2014; Newfoundland Shelf, 1980-2014; Pacific Shelf including Southern Vancouver Island (SV), 1986-2014 for fall and spring; Northern Vancouver Island (NV), 1998-2014 for fall and spring; GAK1

station, annual average for 1987-2014; and finally, Eastern Bering Sea (EBS), 1982-2014 summer. For each region, the yearly and regionally-averaged values were calculated for the comparison, once missing values are removed.

Three statistical measures were used to quantify the performance of each model in representing the observed surface and bottom temperature: model bias (Bs), standard deviation (STD), and linear trend (Tr). A consideration of the bottom temperature was the depth of each model grid at these regions. Due to the coarse resolution of the models, they are generally incapable of resolving deep channels. Each model's depth distribution generally contains a small number of grid cells at the tails of distribution while generally located close to the regional mean depth. The number of each model grid's counts for each region is provided in Table S 1.

A scoring method modified from Kling-Gupta Efficiency (KGE) is used to score all models (Gupta et al. 2009). Each model is ranked according to its regional average score over all regions.

The bias is the mean of the difference between the model results (Model) and the observations (OBS) and is defined as:

Eq. 1 $$Bs = \frac{1}{N}\sum_{year=N1}^{year=N2}(Model_{year} - OBS_{year})$$

, where N is the number of years.

The standard deviation (σ_M) describes the dispersion of the data around the mean. It is clustered around the mean if the standard deviation (STD) is low and is defined for each model as follows:

Eq. 2 $$\sigma_M = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Model_i - \mu_{Model})^2}$$

, where μ is the mean of the model (M), and N is the total number of years.

The linear trend only considers the changes over the whole period which then is divided by the number of decades to provide the trend per decade. For each CMIP6 model and each observation dataset (e.g., HadISST for SST and GLORYS12 for BT), we employed simple linear regression (SLR) to statistically test whether the slope differs from zero (p = 0.05), as well as to calculate the linear trend (slope) over the analysis period to quantify the rate of change in sea surface and bottom temperatures. Qualitative comparisons among SLR results for CMIP6 models relative to observation datasets were conducted to assess model result trend bias from potentially not capturing certain physical processes or external

factors. Slope magnitudes for each single CMIP6 model were employed in subsequent skill performance calculations.

To measure the predictive accuracy of a model, the original Kiling-Gupta Efficiency (Gupta et al. 2009) is used which consists of the decomposition of the errors into its constituent components , linear correlation, variability, and bias and provide a more robust metric by removing the dependence of bias on the standard deviation of observed data which is defined as follows:

Eq. 3 $$KGE = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_M}{\sigma_O} - 1)^2 + (\frac{\mu_M}{\mu_O} - 1)^2}$$

, where r is the correlation, σ and μ are the standard deviations and mean of the model (M) and observation(O), respectively. The KGE ranges from -∞ to 1 and values less than zero indicate that the model performs no better than the mean of observations. As this report is more focused on the three measures of bias, standard deviation, and trend, a new version of Eq. 3 is adopted. As the range of model biases are wider than trend and standard deviation, the scoring results were more affected by the bias than the other two statistics. The same feature exists between the standard deviation and trend, where the CMIP6 models shows a higher range of standard deviation than trend. To fix this issue, these statistics are each normalized to their ranges among the models to balance the weight of each term in the final result. In the case of STD, the difference between the 22-model STD and the observation for each region is normalized to have a range between 0 and 1. A similar approach is used for the trend, except that the priority is given to the models that show the same sign as the observation trend, even if they have a higher distance from the observation. The model trends with the same sign as observations get the range between 0 to 0.4 (lower range to be used in Eq. 4 and showing higher scores) and opposite signs range from 0.5 to 1 (higher range to be used in Eq. 4 and showing lower scores). For instance, if the observation trend is -0.3 °C/decade and Model A and B trends are -0.9 and 0.1, Model A gets a lower range (i.e., better score) than Model B, even though Model B has a lower difference (0.4) than Model A (0.6) against the observation trend. This new form of Eq. 3 is called Modified Kling-Gupta Efficiency (MKGE) in this report and is defined as follows:

Eq. 4 $$MKGE = 1 - \sqrt{(Bias_{normalized})^2 + (\Delta\_STD_{normalized})^2 + (\Delta\_Trend_{Normalized})^2}$$

This MKGE has a range of -0.73 to 1 with the best scores equal to 1. With the inclusion of the trend, the scoring captures trend, STD, and bias in this study. It should be noted that since Eq. 4 is normalized to the

range of each variable at each location, each score is a localized score for each region and is not comparable among regions.

The 22 CMIP6 models get their scores from Eq. 4 for each region and then ranks from 1 to 22 in that region. Then, for each ocean, the average rank of all regions in that ocean is calculated and is used as a base for the final raking in that ocean. As an example, a lower average rank indicates a better overall performance of a model in that ocean. Also, the individual ranks in each region could be used to find the best performance model in an specific region in an ocean. At the end, the top four models are suggested based on the final ranks for each ocean. The primary rationale behind selecting the top 4 models was their close average regional rankings. Rather than singling out one model as superior, we opted for the top four to account for minor fluctuations in the average regional rank. The statistics and scoring for all models are presented in the Appendices tables S.2 to S.38.

3 Results

In this section, the ranking breakdown and the statistics for the top four models are provided, and then the plots for the top four models are presented.

3.1 Sea Surface Temperature

The results for each major region are presented. Based on the performance of the 22 CMIP6 models, the top four models for simulating SST are recommended.

3.1.1 Atlantic Shelf Water

Table 3 shows the model's individual ranking in each region in the North Atlantic and the average ranking ("Region Average" column in Table 3) which is used to assign the final rank in the last column of Table 3. The model CNRM-CM6-1-HR gets a region average of 4.5 and is ranked as the top model for this region. The MRI-ESM2-0 and NorESM2-LM are ranked the second and third places, at region average of 5.2 and 6.4, respectively. Finally, the model MIROC6 is ranked fourth with an region average of 7.6. The lowest region averages belong to EC-Earth3, and ACCESS-ESM1-5 at 19.3 and 16.7, respectively, ranking 22nd and 21st in the North Atlantic regions. It is worth noting that generally, a model with good rank for one region might not have a good rank in another region. For example, MRI-ESM2-0 has a lower rank in GoM (ranked 19; Table 3) but consistently has a good performance in all other North Atlantic shelf regions.

Table 3. North Atlantic Shelf model ranking based on the Modified Kling-Gupta Efficiency (MKGE) for SST. The lower the rank, the better the model performance. The **CNRM-CM6-1-HR** , **MRI-ESM2-0**, **NorESM2-LM**, and **MIROC6** get the overall rank of 1, 2, 3, and 4, respectively.

Models/Region	North Atlantic Shelf												
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	Region Average	Overall Rank
ACCESS-CM2	22	21	21	14	12	12	17	17	17	12	7	15.6	19
ACCESS-ESM1-5	21	20	18	17	20	22	16	19	16	10	5	16.7	21
AWI-CM-1-1-MR	2	12	13	15	5	15	10	11	7	7	15	10.2	10
CAMS-CSM1-0	3	2	2	1	1	13	13	15	15	15	14	8.5	8
CanESM5	5	9	7	5	11	6	19	18	19	19	17	12.3	12
CESM2	14	15	19	19	22	19	6	9	3	1	3	11.8	11
CESM2-WACCM	13	16	20	21	21	18	9	10	8	6	9	13.7	14
CMCC-CM2-SR5	7	10	10	11	13	5	11	3	2	3	16	8.3	6
CNRM-CM6-1	18	17	12	9	14	8	15	16	20	20	19	15.3	18
CNRM-CM6-1-HR	**1**	**3**	**3**	**7**	**7**	**4**	**4**	**2**	**4**	**8**	**6**	**4.5**	**1**
CNRM-ESM2-1	15	8	1	2	18	11	18	21	21	22	22	14.5	16
EC-Earth3	20	19	16	10	19	20	22	22	22	21	21	19.3	22
GISS-E2-1-G	11	13	17	20	15	21	21	14	14	9	20	15.9	20
IPSL-CM6A-LR	8	11	6	6	6	3	1	5	6	16	18	7.8	5
MIROC6	**16**	**1**	**4**	**3**	**16**	**1**	**3**	**4**	**9**	**14**	**13**	**7.6**	**4**
MIROC-ES2L	12	14	14	18	17	17	12	12	12	17	12	14.3	15
MPI-ESM1-2-HR	10	4	11	12	9	9	5	6	5	13	10	8.5	8
MPI-ESM1-2-LR	4	6	9	13	4	10	7	7	10	11	11	8.4	7
MRI-ESM2-0	**19**	**5**	**8**	**8**	**8**	**2**	**2**	**1**	**1**	**2**	**1**	**5.2**	**2**
NorESM2-LM	**6**	**7**	**5**	**4**	**2**	**7**	**8**	**8**	**11**	**4**	**8**	**6.4**	**3**
TaiESM1	17	18	22	22	3	16	14	13	13	5	2	13.2	13
UKESM1-0-LL	9	22	15	16	10	14	20	20	18	18	4	15.1	17

Similarly, the fourth ranked model, MIROC6, has a good rank in all Atlantic regions except in four (GoM, GSL, NLS, and HB) where the model performance decreases substantially (ranks 16, 16, 14, and 13, respectively; Table 3). If the interest is in using these models as the boundary forcing to conduct ocean model downscaling for a particular region, the model with a better rank in this region could be a good candidate. For instance, CNRM-CM6-1-HR consistently shows better ranks on the Scotian, Newfoundland, and Labrador shelves (Table 3). Also, MRI-ESM2-0 has good ranks for all the North Atlantic regions other than GoM. The breakdown of the statistics for the top four models is provided for all regions in Table 4. Regarding the bias, the top four models show inconsistency in different regions, except on the SNS and CNS where all of the top four models show a cold bias in comparison to the HadISST. The first-ranked model, CNRM-CM6-1-HR (Table 3), has a cold bias in GoM (-0.7 °C) and on the Scotian shelf up to the CNS (-0.7 °C), excluding GSL, and a warm bias for the NNS (0.3 °C) to HB (0.5 °C) (Table 4).

Table 4. SST statistics for eleven regions on the Atlantic Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), trend (Tr; unit: °C/decade with *p*-value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1955-2014 period. The HadISST evaluation dataset is indicated in blue.

Models/Regions	GoM				WSS				CSS			
	Bs	STD	Tr(p)	MKGE	Bs	STD	Tr(p)	MKGE	Bs	STD	Tr(p)	MKGE
HadISST	-	0.7	0.23(<0.005)	-	-	0.8	0.35(<0.005)	-	-	0.8	0.34(<0.005)	-
CNRM-CM6-1-HR	-0.7	0.7	0.24(<0.005)	0.68	-1.0	0.7	0.24(<0.005)	0.62	-1.0	0.7	0.24(<0.005)	0.64
MRI-ESM2-0	1.0	0.8	0.13(0.029)	-0.08	1.0	0.8	0.16(0.006)	0.49	0.3	0.8	0.17(0.002)	0.46
NorESM2-LM	3.5	0.7	0.22(<0.005)	0.44	3.5	0.8	0.20(<0.005)	0.47	1.5	0.7	0.22(<0.005)	0.58
MIROC6	2.5	0.8	0.29(<0.005)	-0.02	2.6	0.8	0.27(<0.005)	0.69	1.9	0.7	0.25(<0.005)	0.64
	ESS				GSL				SNS			
HadISST	-	0.7	0.28(<0.005)	-	-	0.5	0.20(<0.005)	-	-	0.7	0.20(<0.005)	-
CNRM-CM6-1-HR	-1.1	0.6	0.22(<0.005)	0.50	0.5	0.5	0.15(<0.005)	0.50	-1.0	0.6	0.22(<0.005)	0.61
MRI-ESM2-0	-0.7	0.7	0.20(<0.005)	0.48	-0.2	0.5	0.15(<0.005)	0.46	-0.7	0.7	0.20(<0.005)	0.77
NorESM2-LM	0.0	0.7	0.20(<0.005)	0.57	-0.6	0.6	0.19(<0.005)	0.65	-0.6	0.6	0.14(0.001)	0.53
MIROC6	1.1	0.7	0.23(<0.005)	0.58	2.5	0.7	0.21(<0.005)	-0.13	-0.1	0.7	0.22(<0.005)	0.86
	CNS				NNS				SLS			
HadISST	-	0.6	0.19(<0.005)	-	-	0.5	0.16(<0.005)	-	-	0.4	0.15(<0.005)	-
CNRM-CM6-1-HR	-0.7	0.5	0.19(<0.005)	0.69	0.3	0.4	0.15(<0.005)	0.85	0.5	0.3	0.10(<0.005)	0.74
MRI-ESM2-0	-0.5	0.5	0.19(<0.005)	0.81	-0.3	0.5	0.17(<0.005)	0.85	-0.3	0.4	0.13(<0.005)	0.88
NorESM2-LM	-0.9	0.6	0.11(0.012)	0.54	-0.7	0.6	0.10(0.012)	0.61	-0.8	0.5	0.08(0.025)	0.53
MIROC6	-0.3	0.7	0.14(<0.005)	0.74	0.2	0.6	0.12(<0.005)	0.75	0.1	0.6	0.12(<0.005)	0.63
	NLS				HB							
HadISST	-	0.3	0.08(<0.005)	-	-	0.3	0.12(<0.005)	-				
CNRM-CM6-1-HR	0.8	0.3	0.08(<0.005)	0.57	0.5	0.4	0.09(<0.005)	0.67				
MRI-ESM2-0	0.4	0.3	0.09(<0.005)	0.78	0.1	0.4	0.10(<0.005)	0.76				
NorESM2-LM	-0.4	0.3	0.05(0.033)	0.73	-0.6	0.4	0.14(<0.005)	0.65				
MIROC6	0.3	0.5	0.13(<0.005)	0.43	0.5	0.5	0.14(<0.005)	0.34				

On the other hand, the third-ranked model, NorESM2-LM (Table 3), overestimates the SST in the GoM (bias of 3.5 °C) and Scotian Shelf, except for ESS with zero bias, while underestimating it elsewhere in North Atlantic shelf waters. The magnitude of bias is different in each region for each model with the top two models showing a bias lower than 1.1 °C, while the other two models (the third and fourth) have biases that reach as high as 2.5 to 3.5 °C in GoM and WSS.

All of the top four models are consistent in capturing the STD in all the North Atlantic regions; however, the fourth-ranked model, MIROC6, tends to have higher STD from the Scotian Shelf to Labrador Shelf and

Hudson Bay. The four top ranked models agree with observed trends, both in direction and magnitude; however, in most cases they underestimate these trends compared to HadISST. Also, the top four models exhibit p-value below 0.05, indicating statistically significant trends, which is consistent with the HadISST trend p-value. The statistics for all models and their MKGE scores for SST for each region are provided in detail in the Appendix in Tables S 2 to S 19.

Table 4 shows the regional MKGE scores for the top four models (better performance as the scores get closer to one). The top four models have positive scores higher than 0.5 in all regions, except MRI-ESM2-0 (with a score of -0.08) in the GoM and MIROC6 in the GoM and GSL with scores of -0.02 and -0.13, respectively. Qualitative comparisons are provided with time series plots for the top four ranked models against the HadISST in Figure 7-9. It should be noted that these four models are the top ones for the entire North Atlantic shelf waters. In some regions, they might show better ranks while underperforming in other regions or seasons. For instance, in the GoM and WSS, NorESM2-LM and MIROC6 (also at GSL) show higher biases than the other two models but generally follow the HadISST (Table 4; Figure 7). Going northward from the Scotian Shelf to the Newfoundland and Labrador Shelves and Hudson Bay, the bias in the four models decreases and the models show almost the same reduction in STD as the HadISST (Table 4; Figures 8-9).

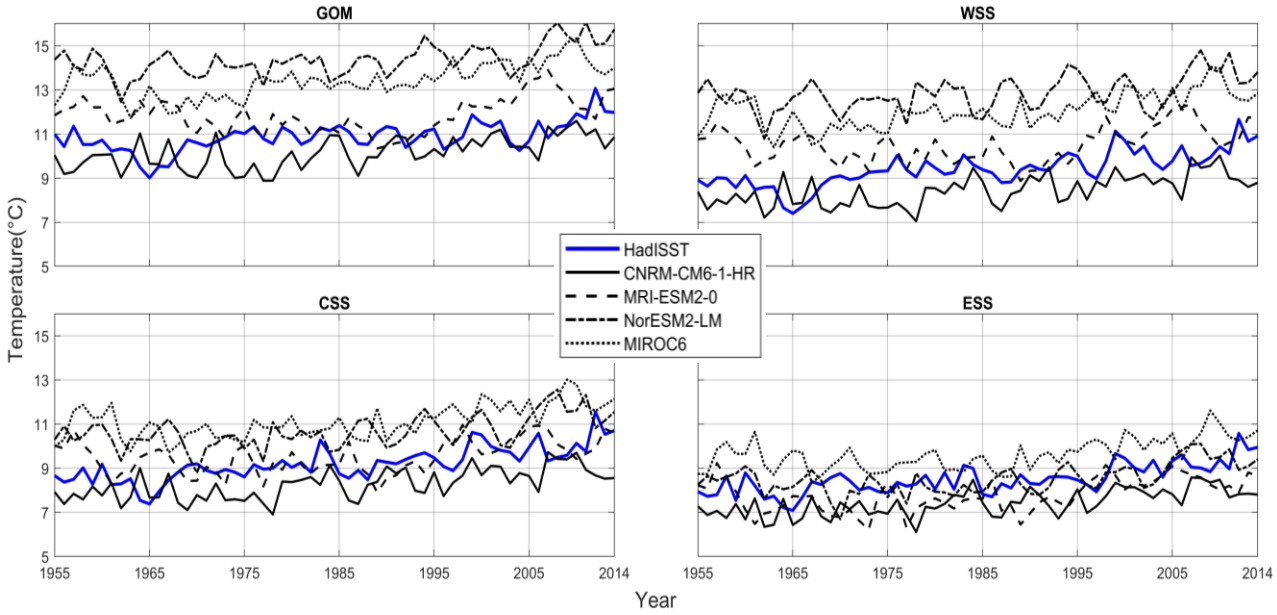

Figure 7.Regionally-averaged annual time series of sea surface temperature from top four CMIP6 models (black) for four Atlantic regions, GoM, WSS, CSS, and ESS against HadISST (blue) for the period 1955-2014.

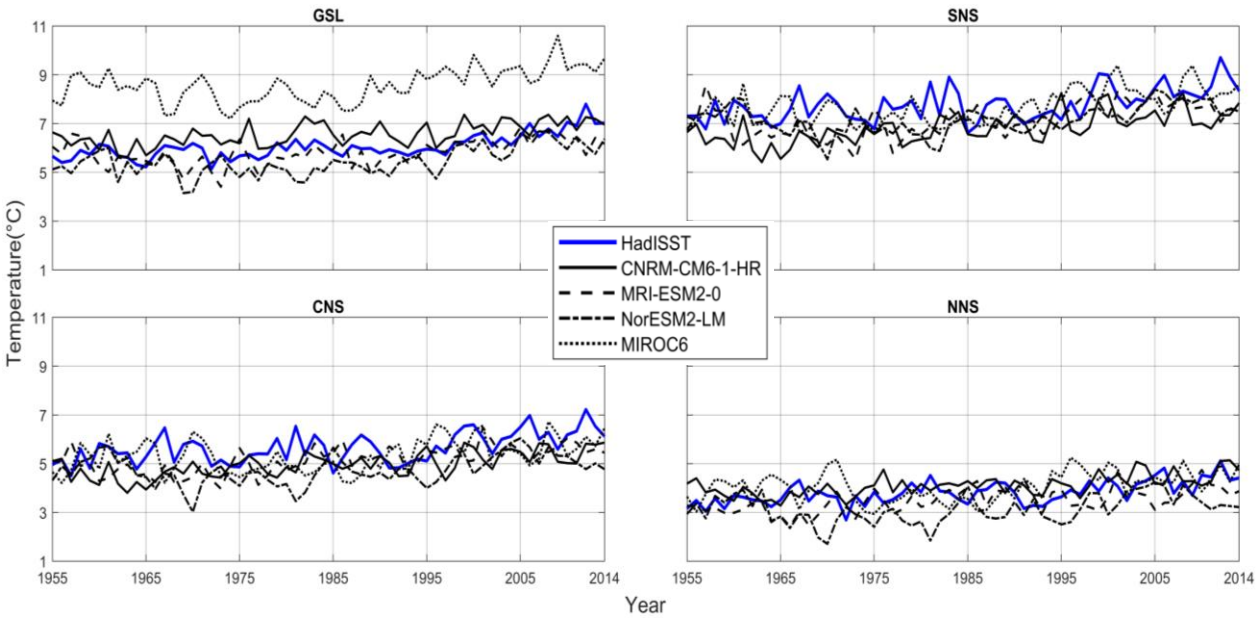

Figure 8. Regionally-averaged annual time series of sea surface temperature from top four CMIP6 models (black) for four Atlantic regions, GSL, SNS, CNS, and NNS against HadISST (blue) for the period 1955-2014.

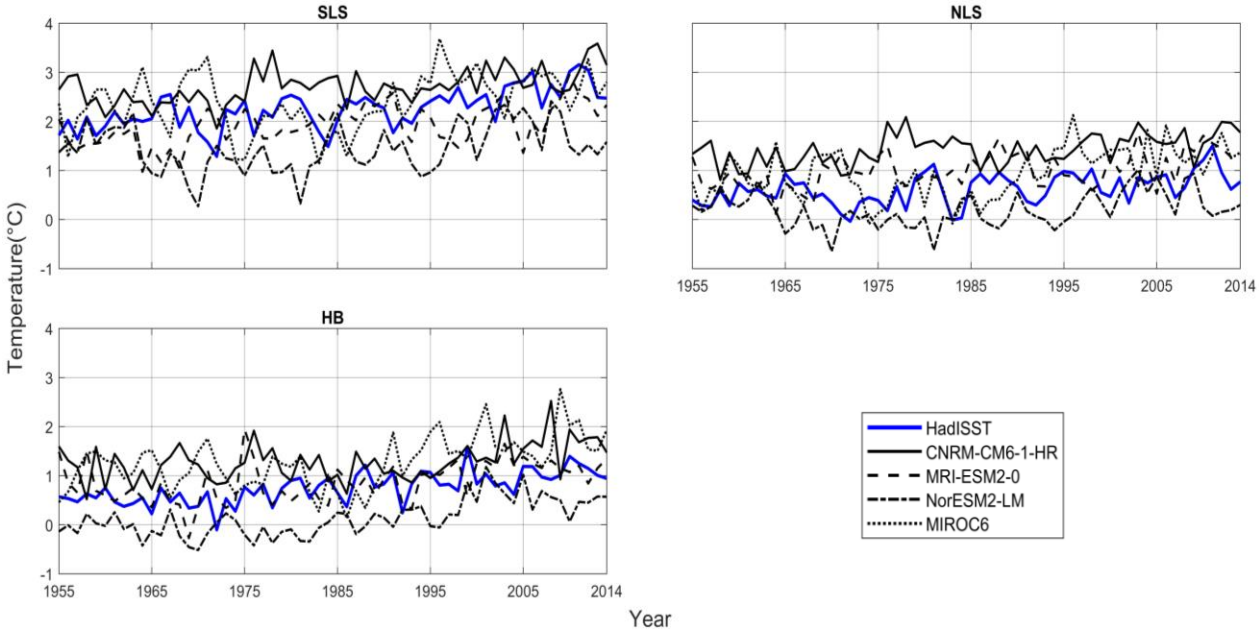

Figure 9. Regionally-averaged annual time series of sea surface temperature from top four CMIP6 models (black) for three Atlantic regions, SLS, NLS, and HB (in black) against HadISST (blue) for the period 1955-2014.

3.1.2 Arctic Shelf Water

The same SST metrics and plots are provided for the four Arctic regions in the following tables, and the time series plots of the top four models against the HadISST are depicted at the end. The performance of the 22 CMIP6 models is based on the four regions and the average ranks are more variable (standard deviation of 4.8) than the Atlantic region (STD of 4.1), indicating a wider range of performance for the Arctic than for the Atlantic. Based on the ranking table for the Arctic (Table 5), CESM2 has the lowest region average at 3.5 and is ranked as the top model for the Arctic. This model shows a better rank for CAA, SBS, and SC and a lower rank for BB (ranks seventh); on average, it ranks the first. The second-best model is CESM2-WACCM with a region average of 4.8, but like to CESM2 it has a low rank in BB. The third and fourth highest ranked models, MPI-ESM1-2-LR and MRI-ESM2-0, have region average ranks at 6.8 and 7, respectively. These two models show better performance (ranks first and third) in BB than the CESM2 and CESM2-WACCM.

Table 6 provides the statistics breakdown for the top four models. The top four models get a high MKGE score (>0.5) for all four regions. Regarding bias, the four models have bias lower than 0.7 °C in all four regions and mostly have cold biases in these regions; CESM2-WACCM is the only one with cold biases for all four Arctic regions. All four models have STD consistent with the HadISST data, however, MPI-ESM1-2-LR and MRI-ESM2-0 models show slightly higher STD in BB and SBS. All four models closely capture the trend of HadISST in all four regions, however, MPI-ESM1-2-LR always underestimates the trend. Among the top four models, MRI-ESM2-0 and MPI-ESM1-2-LR have p-value above 0.05 at BB (at 0.191) and SBS (at 0.207), while the HadISST has a statistically significant trend in the two regions.

The time series plots of the top four models against the HadISST are depicted in Figure 10. The models show a similar range of variability across regions as HadISST, showing a higher spread in BB, SBS, and SC and lower SST variability in the CAA.

Table 5. Arctic Shelf model ranking based on the Modified Kling-Gupta Efficiency (MKGE) for SST. The lower the rank, the better the model performance. The **CESM2**, **CESM2-WACCM**, **MPI-ESM1-2-LR**, and **MRI-ESM2-0** get the overall rank of 1, 2, 3, and 4, respectively.

Models/Regions	Arctic Shelf					
	BB	CAA	SBS	SC	Region Average	Overall Rank
ACCESS-CM2	13	5	4	13	8.8	10
ACCESS-ESM1-5	5	11	13	1	7.5	5
AWI-CM-1-1-MR	6	20	12	5	10.8	11
CAMS-CSM1-0	11	18	8	16	13.3	14
CanESM5	18	9	7	9	10.8	11
CESM2	**7**	**2**	**1**	**4**	**3.5**	**1**
CESM2-WACCM	**9**	**6**	**2**	**2**	**4.8**	**2**
CMCC-CM2-SR5	12	22	21	22	19.3	20
CNRM-CM6-1	21	21	22	19	20.8	22
CNRM-CM6-1-HR	17	13	17	14	15.3	18
CNRM-ESM2-1	20	19	20	20	19.8	21
EC-Earth3	22	3	16	18	14.8	17
GISS-E2-1-G	19	17	9	21	16.5	19
IPSL-CM6A-LR	8	12	19	17	14.0	16
MIROC6	14	16	15	10	13.8	15
MIROC-ES2L	4	15	14	15	12.0	13
MPI-ESM1-2-HR	2	8	18	6	8.5	6
MPI-ESM1-2-LR	**1**	**4**	**11**	**11**	**6.8**	**3**
MRI-ESM2-0	**3**	**7**	**10**	**8**	**7.0**	**4**
NorESM2-LM	10	14	3	7	8.5	6
TaiESM1	16	10	5	3	8.5	6
UKESM1-0-LL	15	1	6	12	8.5	6

Table 6. SST statistics for four regions on the Arctic Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), trend (Tr; unit: °C/decade , p value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1955-2014 period. The HadISST evaluation dataset is indicated in blue.

Models/Regions	BB					CAA			
	Bs	STD	Tr(p)	MKGE		Bs	STD	Tr(p)	MKGE
HadISST	-	0.2	0.07(<0.005)	-		-	0.1	0.05(<0.005)	-
CESM2	-0.6	0.2	0.07(<0.005)	0.56		-0.1	0.1	0.04(<0.005)	0.78
CESM2-WACCM	-0.6	0.2	0.07(<0.005)	0.53		-0.1	0.1	0.03(<0.005)	0.66
MPI-ESM1-2-LR	0.2	0.3	0.04(0.054)	0.78		-0.0	0.1	0.03(<0.005)	0.70
MRI-ESM2-0	-0.2	0.3	0.03(0.191)	0.71		0.3	0.1	0.04(<0.005)	0.64
	SBS					SC			
HadISST	-	0.4	0.07(0.01)	-		-	0.3	0.12(<0.005)	-
CESM2	-0.2	0.4	0.10(<0.005)	0.78		0.1	0.4	0.15(<0.005)	0.61
CESM2-WACCM	-0.3	0.4	0.10(<0.005)	0.74		-0.2	0.3	0.09(<0.005)	0.71
MPI-ESM1-2-LR	-0.6	0.6	0.06(0.207)	0.31		-0.7	0.2	0.06(<0.005)	0.14
MRI-ESM2-0	0.0	0.6	0.14(<0.005)	0.33		0.3	0.4	0.13(<0.005)	0.40

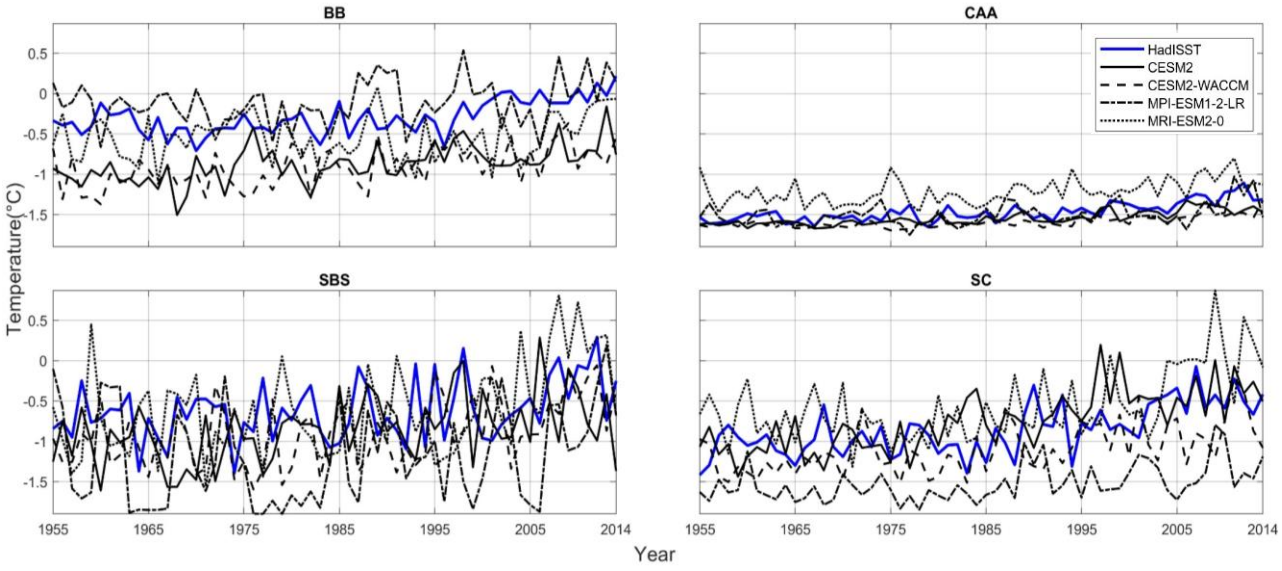

Figure 10. Regionally-averaged annual time series of sea surface temperature from top four CMIP6 models (black) for four Arctic regions, BB, CAA, SBS, and SC against HadISST (blue) for the period 1955-2014.

3.1.3 Pacific Shelf Water

The ranking table (Table 7), breakdown of the top four models statistics (Table 8), and time series plots (Figure 11) are provided for the Pacific SST. The top model in the Pacific shelf is ACCESS-CM2 with a region average of 1.3 followed by TaiESM1 and CESM2-WACCM with region average ranks of 3.0 and 3.7, respectively. In the fourth place, CNRM-CM6-1-HR gets a 4.7 region average. The lowest rank models for SST in these regions are CMCC-CM2-SR5 and ACCESS-ESM1-5 at 22 and 21, respectively. While both ACCESS-CM2 and ACCESS-EMS1-5 share many components (with different versions), such as atmospheric, oceanic, sea-ice, and land surface; the latter is an Earth System Model (ESM) which has biogeochemical processes that can interact with physical climate.

The SST statistics for the top four models (Table 8) show a cold bias for the top two models (ACCESS-CM2 and TaiESM1) in three regions, while CNRM-CM6-1-HR shows a warm bias of 0.7 °C for BS, 1.2 °C for AS, and 2.3 °C for BCS. In a situation where a model shows both cold and warm biases in a region, the mean bias could average to zero. On the BCS, the top model, ACCESS-CM2, shows this near zero bias (Table 8). The four models show good consistency in capturing the STD, but they usually show a higher STD than HadISST (Table 8). The long-term trend in HadISST shows a reduction from BS to the BCS. Among the top four models, only TaiESM1 captures this pattern in the trend. However, it shows a slight negative trend on the BCS. Otherwise, the four models show generally positive trends in SST in these regions but mostly

overestimate its magnitude. The HadISST only show a statistically significant trend in BS, and only TaiESM1 and CESM2-WACCM show the p-value below 0.05 in this region. In the other two regions, AS and BCS, while the HadISST have p-value higher than 0.05, the top four models have mixed results, some with higher p-value than 0.05 and some below. The time series plots of the top four models against the HadISST are depicted in Figure 11. The range of variation in SST in the four models is consistent with HadISST, and the warm bias in the CNRM-CM6-1-HR is clear in the three regions.

Table 7. Pacific Shelf model ranking based on the Modified Kling-Gupta Efficiency (MKGE) for SST. The lower the rank, the better the model performance. The **ACCESS-CM2**, **TaiESM1**, **CESM2-WACCM**, and **CNRM-CM6-1-HR** get the overall rank of 1, 2, 3, and 4, respectively.

Models/Region	Pacific Shelf				
	BS	**AS**	**BCS**	**Region Average**	**Overall Rank**
ACCESS-CM2	**2**	**1**	**1**	**1.3**	**1**
ACCESS-ESM1-5	20	20	21	20.3	21
AWI-CM-1-1-MR	7	8	5	6.7	7
CAMS-CSM1-0	9	15	16	13.3	13
CanESM5	15	17	12	14.7	14
CESM2	4	6	8	6.0	5
CESM2-WACCM	**3**	**4**	**4**	**3.7**	**3**
CMCC-CM2-SR5	22	22	22	22.0	22
CNRM-CM6-1	16	10	18	14.7	14
CNRM-CM6-1-HR	**1**	**3**	**10**	**4.7**	**4**
CNRM-ESM2-1	10	7	11	9.3	9
EC-Earth3	21	11	15	15.7	16
GISS-E2-1-G	19	18	19	18.7	20
IPSL-CM6A-LR	11	5	3	6.3	6
MIROC6	8	13	14	11.7	11
MIROC-ES2L	17	19	13	16.3	18
MPI-ESM1-2-HR	12	9	6	9.0	8
MPI-ESM1-2-LR	13	21	20	18.0	19
MRI-ESM2-0	6	16	7	9.7	10
NorESM2-LM	14	14	9	12.3	12
TaiESM1	**5**	**2**	**2**	**3.0**	**2**
UKESM1-0-LL	18	12	17	15.7	16

Table 8. SST statistics for three regions on the Pacific Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), and trend (Tr; unit: °C/decade, p value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1955-2014 period. The HadISST evaluation dataset is indicated in blue.

Models/Regions	BS					AS					BCS			
	Bs	STD	Tr(p)	MKGE		Bs	STD	Tr(p)	MKGE		Bs	STD	Tr(p)	MKGE
HadISST	-	0.4	0.06(0.04)	-		-	0.5	0.03(0.47)	-		-	0.4	0.02(0.56)	-
ACCESS-CM2	-1.3	0.5	0.01(0.691)	0.58		-1.0	0.6	0.05(0.335)	0.77		0.0	0.5	0.09(0.008)	0.86
TaiESM1	-0.1	0.7	0.18(<0.005)	0.41		-1.4	0.6	0.02(0.647)	0.71		-0.1	0.6	-0.04(0.336)	0.74
CESM2-WACCM	-0.5	0.7	0.12(0.014)	0.55		-0.8	0.7	0.05(0.375)	0.67		0.2	0.6	0.09(0.034)	0.66
CNRM-CM6-1-HR	0.7	0.6	0.05(0.293)	0.66		1.2	0.5	0.12(<0.005)	0.67		2.3	0.5	0.14(<0.005)	0.38

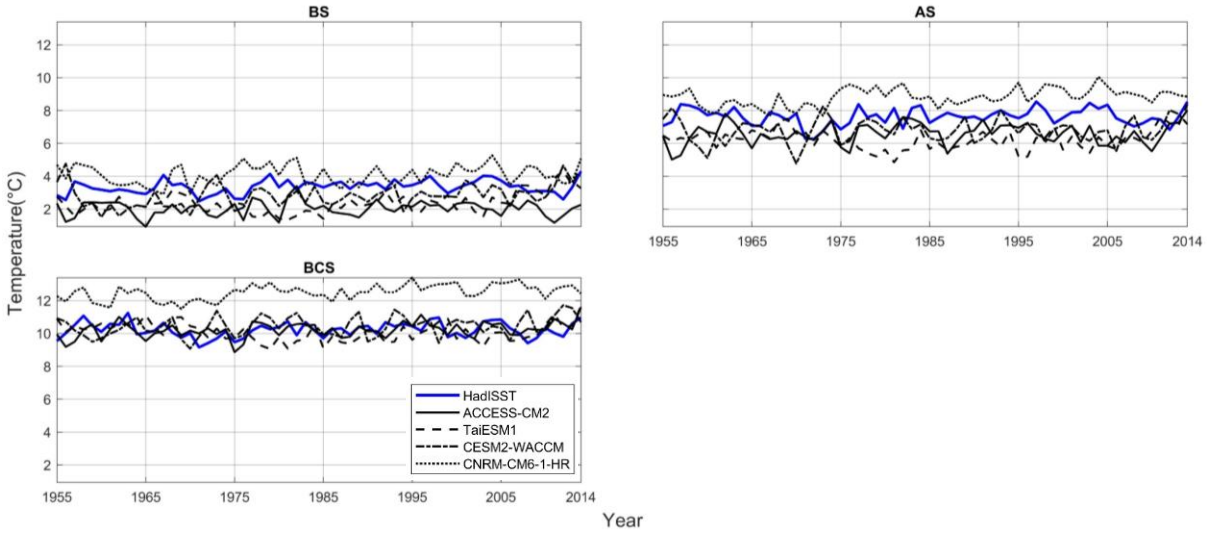

Figure 11. Regionally-averaged annual time series of sea surface temperature from top four CMIP6 models (black) for three Pacific regions, BS, AS, and BCS against HadISST (blue) for the period 1955-2014.

3.2 Bottom Temperature

In this section, models are compared with BT from GLORYIS12 and available *in situ* observational data for each major region. Based on the performance of the 22 CMIP6 models, the top four models are recommended and the breakdown statistics and plots for these four models (against GLORYS12 and available observations) are provided. The same metrics and procedure as previously used in the SST evaluation are used here as well.

3.2.1 Atlantic Shelf Water

Table 9 summarizes the model ranking based on bottom temperature on the North Atlantic shelf. CNRM-ESM2-1 ranks first with a region average of 2.8 for the eleven regions in the North Atlantic. CNRM-CM6-

1-HR follows closely in second place with a region average of 3.2. These two models show good ranks in these regions, however, the CNRM-CM6-1-HR score degrades considerably on the SLS and NLS and ranks 8th and 10th in these two regions. The third and fourth places belong to MRI-ESM2-0 and MPI-ESM1-2-LR with region averages of 7.4 and 8.7, respectively. While MRI-ESM2-0 shows good ranks in the North Atlantic regions, its rank decreases to 14th in the GSL. Also, MPI-ESM1-2-LR has a lower rank in the GoM, WSS, and CSS than for other regions in the North Atlantic.

Table 9. North Atlantic Shelf ranking based on the Modified Kling-Gupta Efficiency (MKGE) for bottom temperature. The lower the rank, the better the model performance. The **CNRM-ESM2-1**, **CNRM-CM6-1-HR**, **MRI-ESM2-0**, and **MPI-ESM1-2-LR** get the overall rank of 1, 2, 3, and 4, respectively.

Models/Region	North Atlantic Shelf												
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	Region Average	Overall Rank
ACCESS-CM2	18	4	8	11	15	20	20	7	13	14	10	12.7	14
ACCESS-ESM1-5	22	22	22	18	16	22	13	10	9	7	13	15.8	20
AWI-CM-1-1-MR	11	6	9	2	4	1	11	16	12	16	19	9.7	6
CAMS-CSM1-0	14	16	11	12	8	14	9	21	16	6	16	13.0	15
CanESM5	15	20	18	14	13	11	3	3	1	1	17	10.5	9
CESM2	4	9	7	8	21	18	16	18	20	21	4	13.3	16
CESM2-WACCM	3	10	10	10	19	13	18	17	17	15	5	12.5	13
CMCC-CM2-SR5	13	12	12	7	9	12	8	6	11	9	9	9.8	7
CNRM-CM6-1	19	14	13	6	3	5	7	11	5	11	11	9.5	5
CNRM-CM6-1-HR	**2**	**1**	**1**	**1**	**1**	**3**	**2**	**4**	**8**	**10**	**2**	**3.2**	**2**
CNRM-ESM2-1	**1**	**2**	**2**	**4**	**2**	**4**	**1**	**1**	**4**	**4**	**6**	**2.8**	**1**
EC-Earth3	5	19	20	21	20	21	15	2	2	5	22	13.8	17
GISS-E2-1-G	10	5	3	9	12	15	21	13	10	12	3	10.3	8
IPSL-CM6A-LR	12	11	16	22	11	10	4	5	6	3	18	10.7	10
MIROC6	17	17	17	15	22	16	22	22	21	18	7	17.6	21
MIROC-ES2L	9	15	15	17	17	17	17	20	19	19	1	15.1	19
MPI-ESM1-2-HR	7	3	5	19	7	9	14	14	14	20	15	11.5	11
MPI-ESM1-2-LR	**21**	**13**	**14**	**5**	**6**	**6**	**5**	**9**	**3**	**2**	**12**	**8.7**	**4**
MRI-ESM2-0	**6**	**8**	**6**	**3**	**14**	**7**	**6**	**8**	**7**	**8**	**8**	**7.4**	**3**
NorESM2-LM	20	18	19	16	10	19	19	19	22	22	21	18.6	22
TaiESM1	8	7	4	13	18	8	12	15	18	17	14	12.2	12
UKESM1-0-LL	16	21	21	20	5	2	10	12	15	13	20	14.1	18

It is worth mentioning that the lowest ranks for BT in North Atlantic shelf waters belong to NorESM2-LM and MIROC6 with region averages of 18.6 and 17.6, ranking 22nd and 21st overall.

Table 10 provides the breakdown of the statistics for the top four models for the bottom temperature in the North Atlantic.

Table 10. Bottom temperature statistics for eleven regions on the Atlantic Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), and trend (Tr; unit: °C/decade , p value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1993-2014 period. The GLORYS12 evaluation dataset is indicated in blue.

Models/Regions	GoM				WSS				CSS			
	Bs	STD	Tr(p)	MKGE	Bs	STD	Tr(p)	MKGE	Bs	STD	Tr(p)	MKGE
GLORYS12	-	0.6	0.39(0.060)	-	-	1.0	0.58(0.079)	-	-	0.9	0.53(0.081)	-
CNRM-ESM2-1	0.9	0.6	0.41(0.043)	0.86	2.7	1.4	0.33(0.465)	0.50	0.3	1.3	0.17(0.677)	0.58
CNRM-CM6-1-HR	-0.6	0.7	0.08(0.722)	0.71	0.2	1.2	0.28(0.483)	0.75	-0.6	1.2	0.13(0.738)	0.62
MRI-ESM2-0	-1.1	0.4	0.28(0.034)	0.47	0.8	0.4	0.41(<0.005)	0.24	0.3	0.5	0.50(<0.005)	0.45
MPI-ESM1-2-LR	2.4	0.7	-0.72(<0.005)	-0.08	1.5	0.1	-0.89(<0.005)	0.05	1.6	0.6	-0.43(0.014)	0.18
	ESS			GSL					SNS			
GLORYS12	-	0.7	0.81(<0.005)	-	-	0.3	0.38(<0.005)	-	-	0.5	0.51(<0.005)	-
CNRM-ESM2-1	1.0	1.0	0.27(0.435)	0.46	0.8	0.3	0.24(0.010)	0.79	1.6	0.6	0.24(0.197)	0.74
CNRM-CM6-1-HR	0.4	0.7	0.32(0.191)	0.75	0.0	0.3	0.33(<0.005)	0.95	1.0	0.5	0.25(0.128)	0.79
MRI-ESM2-0	3.2	0.6	0.48(0.504)	0.51	3.2	0.6	0.55(<0.005)	0.29	2.5	0.5	0.10(0.515)	0.57
MPI-ESM1-2-LR	0.0	0.5	-0.12(<0.005)	0.38	0.9	0.1	0.04(0.451)	0.56	1.4	0.6	0.13(0.434)	0.77
	CNS			NNS					SLS			
GLORYS12	-	0.4	0.30(0.025)	-	-	0.4	0.28(0.020)	-	-	0.5	0.35(0.013)	-
CNRM-ESM2-1	0.5	0.3	0.24(0.027)	0.87	0.0	0.3	0.35(<0.005)	0.92	0.4	0.3	0.34(<0.005)	0.74
CNRM-CM6-1-HR	0.5	0.3	0.27(0.010)	0.86	-0.5	0.3	0.20(0.013)	0.79	-0.4	0.3	0.22(0.006)	0.47
MRI-ESM2-0	1.3	0.3	0.17(0.043)	0.64	1.0	0.3	0.26(0.007)	0.69	1.3	0.3	0.29(<0.005)	0.54
MPI-ESM1-2-LR	0.4	0.4	0.10(<0.005)	0.62	1.0	0.4	0.42(<0.005)	0.69	0.8	0.4	0.45(<0.005)	0.77
	NLS			HB								
GLORYS12	-	0.5	0.36(0.011)	-	-	0.2	-0.11(0.024)	-				
CNRM-ESM2-1	0.8	0.4	0.40(<0.005)	0.71	0.2	0.2	0.20(<0.005)	0.23				
CNRM-CM6-1-HR	-0.2	0.2	0.20(0.009)	0.39	-0.1	0.1	0.13(<0.005)	0.33				
MRI-ESM2-0	2.0	0.4	0.37(<0.005)	0.55	0.5	0.1	0.04(0.219)	0.22				
MPI-ESM1-2-LR	0.9	0.4	0.49(<0.005)	0.78	0.9	0.1	0.10(0.006)	0.13				

All four models mostly tend to overestimate the bottom temperature in these regions, but there are some instances of cold biases. CNRM-ESM2-1, MRI-ESM2-0 (except GoM at -1.1 °C), and MPI-ESM1-2-LR always show warm biases for these regions. CNRM-CM6-1-HR shows the highest number of cold biases (6 regions out of 11) and MRI-ESM2-0 shows one instance. These four models capture the observed STD in the eleven evaluated regions and generally underestimate the BT trend, while the third-ranked model, MRI-ESM2-0, mostly underestimates the STD. Apart from MPI-ESM1-2-LR in GoM, WSS, CSS, and ESS ,in general, the

models share the same trend direction (positive) as GLORYS12 in all the regions, except in HB where GLORYS12 shows a negative trend (-0.11 °C/decade). Other than three regions, GoM, WSS, and CSS, the GLORYS12 shows statistically significant trend in bottom temperature. In general, the top four models show mixed results for the p-values in the North Atlantic region, some models have p-values below 0.05, indicating significant trends, while others have p-values above 0.05, indicating non-significant trends.

The statistics for all models and their regional MKGE scores are provided in detail in appendix Tables S 20 to S 28 for BT. The time series plots of the top four models against GLORYS12 for BT are provided in Figures 12-14. Two models, CNRM-CM6-1-HR in the GSL, and CNRM-ESM2-1 on the NNS demonstrate near zero bias (0.02 °C and 0.00 °C, respectively) (Table 10) in these two regions, while the other models have some systematic cold/warm biases.

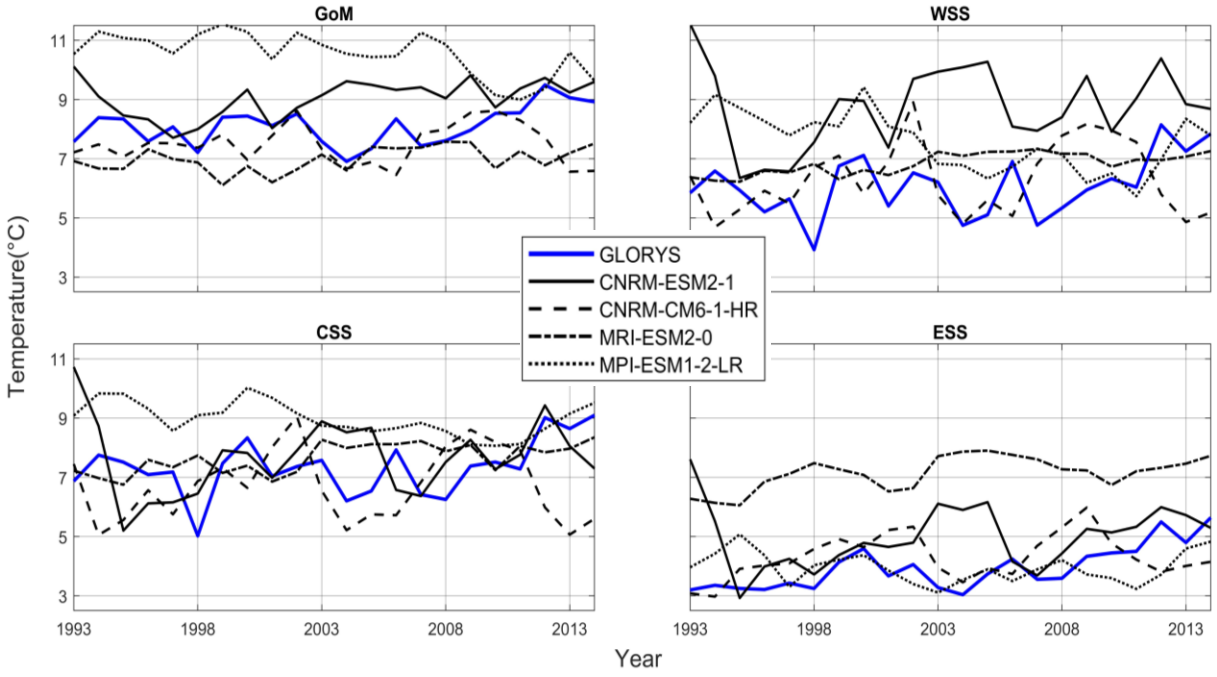

Figure 12. Regionally-averaged annual time series of bottom temperature from top four CMIP6 models (black) for four Atlantic regions, GoM, WSS, CSS, and ESS against GLORYS12 (blue) for the period 1993-2014.

The models show almost the same range of bottom temperature variability as GLORYS12, with the highest spread on the Scotian Shelf and the lowest for the Labrador Shelf and HB. In general, the top four models tend to show warmer biases in these regions, while CNRM-CM6-1-HR shows some cold biases. The models show more variable bias on the Scotian shelf, with the bias reducing with latitude, with one exception on NLS.

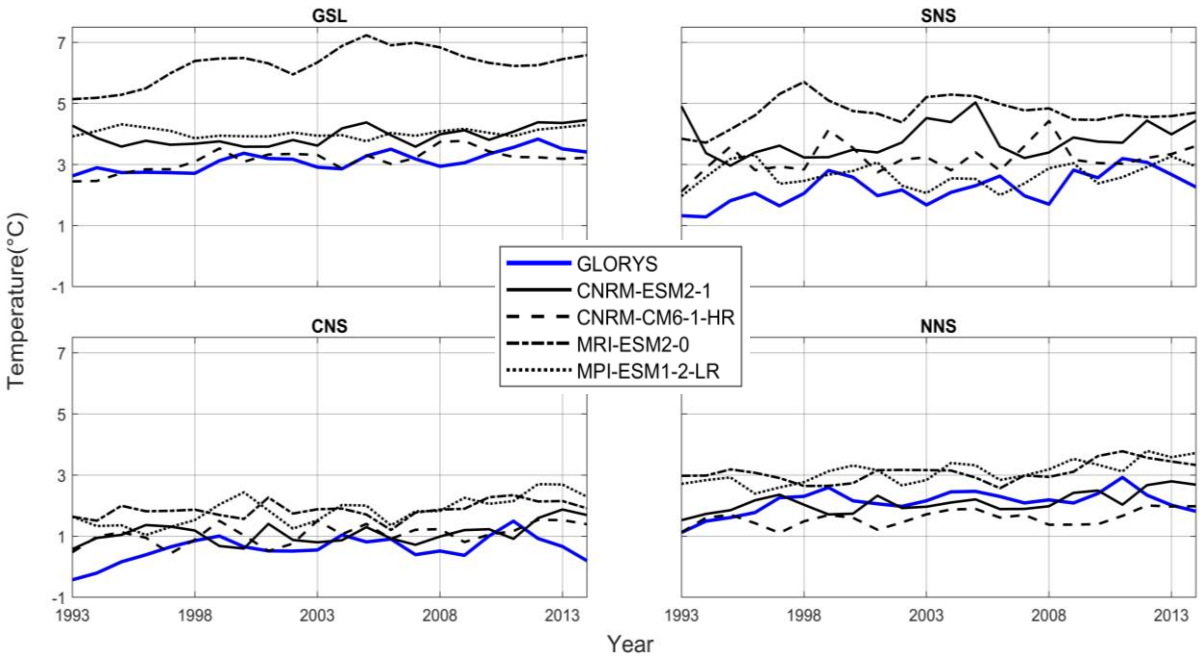

Figure 13. Regionally-averaged annual time series of bottom temperature from top four CMIP6 models (black) for four Atlantic regions, GSL, SNS, CNS, and NSS against GLORYS12 (blue) for the period 1993-2014.

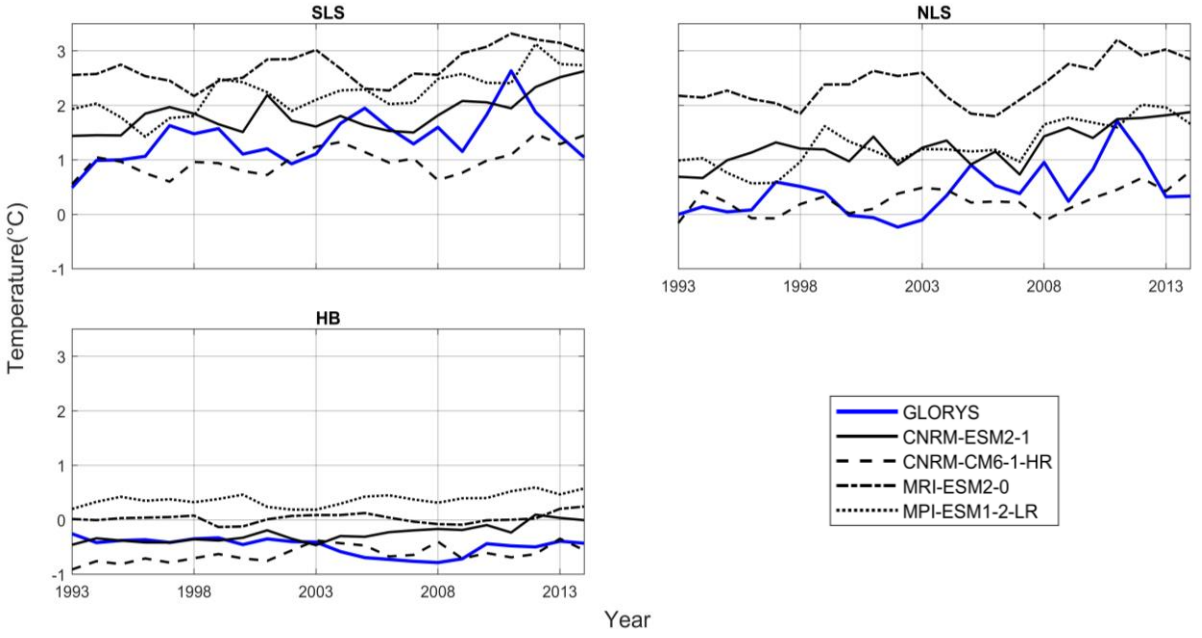

Figure 14. Regionally-averaged annual time series of bottom temperature from top four CMIP6 models (black) for three Atlantic regions, SSL, NLS, and HB against GLORYS12 (blue) for the period 1993-2014.

3.2.1.1 Comparison against observations

In this section, the time series of the top four models' BTs are plotted against the available seasonal observations and GLORYS12. However, in some instances, the observations are snapshot data. The

breakdown of the three statistical measures for the top four models against the observation is provided in the Appendix in Tables S 39, S 40, and S 41. In general, GLORYS12 shows a good agreement in CSS, ESS, WSS, CNS, SNS, SLS, NNS (Figure *15* -17)and closely follow observations, however, in the NLS, while the samples are scarce, there is a higher bias than other regions. Figure 15 shows the BT from the observations (in green) for July on the Scotian Shelf along with the GLORYS12 and the top four models (in black) for the CSS, ESS, and WSS. On the WSS, CNRM-CM6-1-HR is the only model that underestimates BT (-0.5 °C) while the other three models show warm biases (Figure 15c). In the ESS (Figure *15*a), only MPI-ESM1-2-LR shows a cold bias (-0.4 °C) and this is also the only model with a warm bias on the CSS (Figure *15*b) (at 0.6 °C). On the Newfoundland shelf, Figure 16, the four models usually show warm biases in the spring and fall in the CNS and SNS. In this region, MRI-ESM2-0 shows the highest bias (1.6 and 1.4 °C in the CNS, Figure *16*a, b, and 2.6 and 2.6 °C in the SNS, Figure *16*c, d, in fall and spring, respectively). In the NNS and Labrador Shelf (Figure 17), while models still show warm biases, the range of biases is reduced in comparison to the Scotian and Newfoundland Shelves. CNRM-CM6-1-HR, and to a lesser extent CNRM-ESM2-1, show cold biases while the other two models mostly show warm biases.

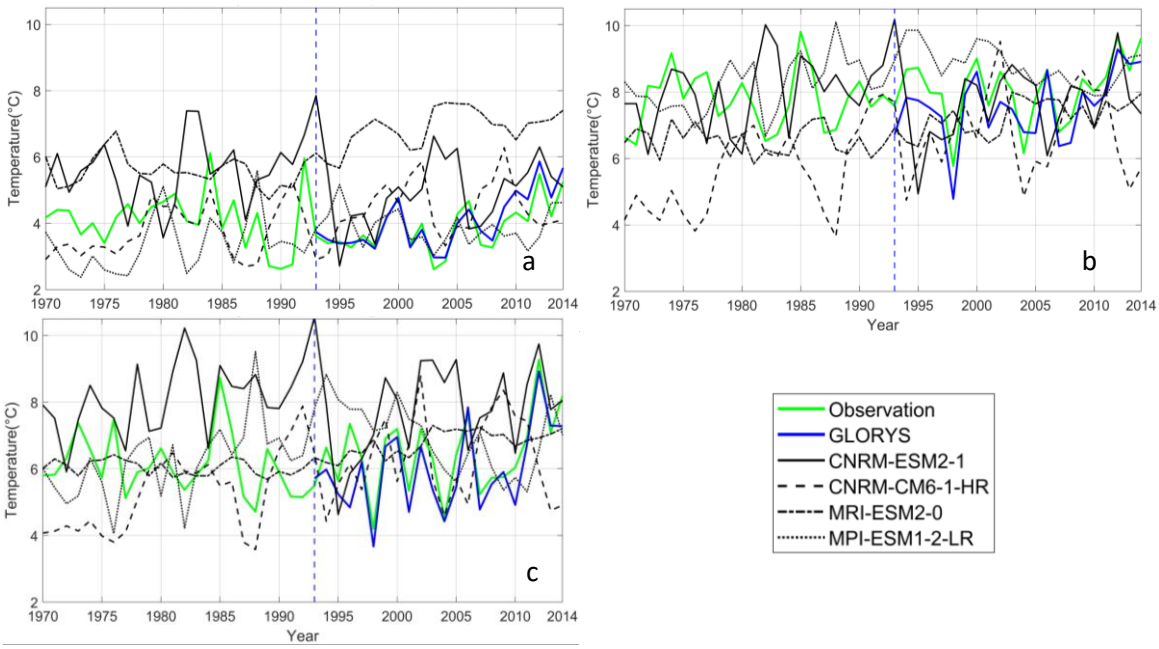

Figure 15. Regionally-averaged time series of bottom temperature from top four CMIP6 models (black) against GLORYS12 (blue) and observation (green) for July for a) ESS, b) CSS, and c) WSS. The vertical dashed line indicates the start of GLORYS12 data in 1993.

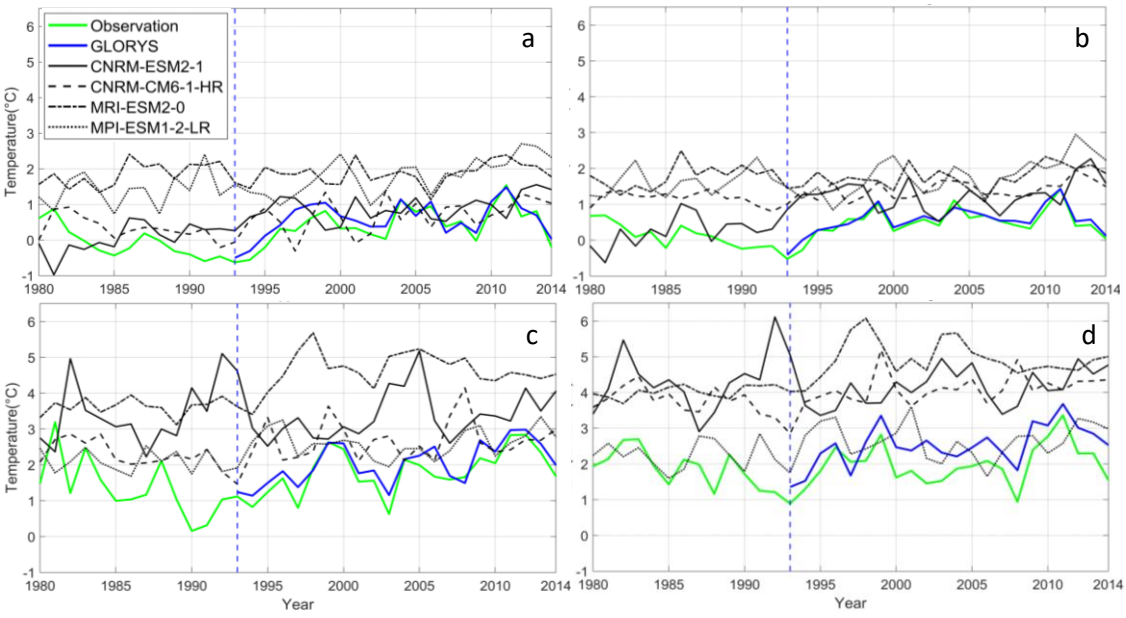

Figure 16. Regionally-averaged time series of bottom temperature from top four CMIP6 models (black) against GLORYS12 (in blue) and observation (green) for spring vs fall for a, b) CNS, and c, d) at SNS. The vertical dashed line is the start of GLORYS12 data in 1993.

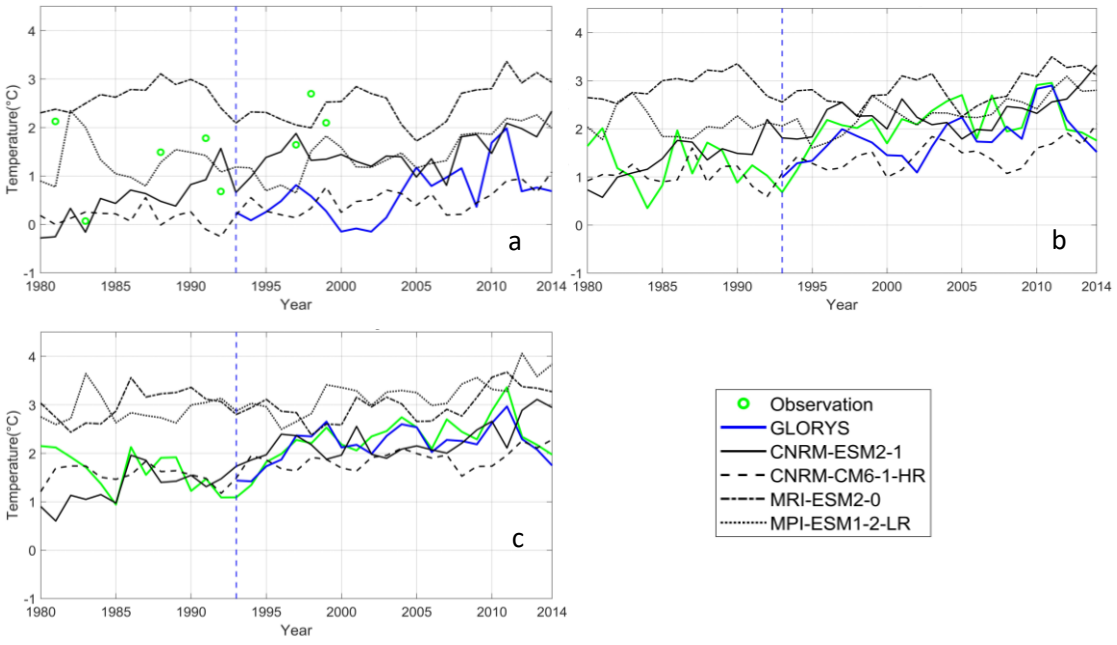

Figure 17. Regionally-averaged time series of bottom temperature from top four CMIP6 models (black) against GLORYS12 (in blue) and observation (green) for fall for a) NLS, b) SLS, and c) NNS. The vertical dashed line is the start of GLORYS12 data in 1993. The green circles represent the spares observations in NLS.

3.2.2 Arctic Shelf Water

In this section, the same statistics and ranking along with the time series plots of bottom temperature are provided for the four regions representing the Arctic shelf. The rankings for all 22 CMIP6 models are shown in Table 11. In this region, three models have the same average ranks (in fourth place), so 6 models are shown instead of 4. The top-ranked model for simulation of the BT across four regions is MPI-ESM1-2-LR with a region average of 6 followed by IPSL-CM6A-LR in the second place with a region average rank of 7. Third place belongs to MPI-ESM1-2-HR with an region average of 8.5, followed closely by UKESM1-0-LL, EC-Earth3, and CESM2-WACCM with region average rank of 8.8. While these models are ranked based on their mean performance in the four regions, some of them are ranked as the top model in a region. For example, IPSL-CM6A-LR, which is ranked second in total, ranked first in SC, and MPI-ESM1-2-HR ranks first in CAA but has a low rank in the other three regions. The lowest ranks for simulating BT in the Arctic regions belong to ACCESS-CM2 and MIROC6, ranking 22nd and 21st, respectively.

Table 11. Arctic Shelf model ranking based on the Modified Kling-Gupta Efficiency (MKGE) for BT. The lower the rank, the better the model performance. The **MPI-ESM1-2-LR**, **IPSL-CM6A-LR**, **MPI-ESM1-2-HR**, and **EC-Earth3** (shared with **CESM2-WACCM**) get the overall rank of 1, 2, 3, and 4, respectively.

Models/Regions	Arctic Shelf					
	BB	CAA	SBS	SC	Region Average	Overall Rank
ACCESS-CM2	21	22	22	20	21.3	22
ACCESS-ESM1-5	10	10	17	6	10.8	13
AWI-CM-1-1-MR	14	11	8	14	11.8	15
CAMS-CSM1-0	15	16	16	9	14.0	17
CanESM5	6	21	15	21	15.8	19
CESM2	17	19	13	2	12.8	16
CESM2-WACCM	**13**	**12**	**7**	**3**	**8.8**	**4**
CMCC-CM2-SR5	2	13	2	22	9.8	8
CNRM-CM6-1	18	6	3	16	10.8	13
CNRM-CM6-1-HR	3	14	18	5	10.0	9
CNRM-ESM2-1	5	20	11	4	10.0	9
EC-Earth3	**8**	**8**	**1**	**18**	**8.8**	**4**
GISS-E2-1-G	7	5	12	17	10.3	12
IPSL-CM6A-LR	**12**	**9**	**6**	**1**	**7.0**	**2**
MIROC6	19	18	19	13	17.3	21
MIROC-ES2L	20	2	21	19	15.5	18
MPI-ESM1-2-HR	**11**	**1**	**10**	**12**	**8.5**	**3**
MPI-ESM1-2-LR	**9**	**4**	**4**	**7**	**6.0**	**1**
MRI-ESM2-0	16	17	20	11	16.0	20
NorESM2-LM	1	15	14	10	10.0	9
TaiESM1	22	3	5	8	9.5	7
UKESM1-0-LL	4	7	9	15	8.8	4

Table 12 shows the statistical measures for the BT for the top six models in the Arctic shelf regions. Other than the BB region, the six models show small biases, mostly less than a absolute magnitude of 0.6 °C with slight over- or under-estimation of BT. There is no specific pattern among the top six models in showing cold or warm biases in these regions; however, UKESM1-0-LL always shows a cold bias in the four regions in the Arctic and two models have near zero bias, CESM2-WACCM in CAA and IPSL-CM6A-LR in SC. The bottom temperature STDs from these six models are very close to that of GLORYS12 in the four regions. As to the trend, the six models show the same trend direction as observation (positive trend) in these regions with one exception, IPSL-CM6A-LR with a small negative trend at about -0.01 °C/decade in BB. In BB, the first and fourth ranked models, MPI-ESM1-2-LR and UKESM1-0-LL, have trends closer (0.05 °C/decade and 0.09 °C/decade) to GLORYS12 (0.04 °C/decade), in comparison to other top models which have higher trends. In other regions, other than one or two models that have a higher trend, the models are in a good agreement with GLORYS12. It is worth mentioning that, GLORYS12 only shows statistically significant trend in BT at CAA, and all top models, other than CESM2-WACCM, are aligned with the GLORYS12 with p-values below 0.05 in this region. In other three regions, while GLORYS12 have p-value above 0.05, the top models show mixed results, some below and some above 0.05.The time series plots (Figure 18) of the six top models in the Arctic regions show that in BB, MPI-ESM1-2-LR, MPI-ESM1-2-HR, and CESM2-WACCM have warm biases, while IPSL-CM6A-LR, EC-Earth3, and UKESM1-0-LL show cold biases.

Table 12. Bottom temperature statistics for regions on the Arctic Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), and trend (Tr; unit: °C/decade, p value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1993-2014 period. The GLORYS12 evaluation dataset is indicated in blue.

Models/Regions	BB					CAA			
	Bs	STD	Tr(p)	MKGE		Bs	STD	Tr(p)	MKGE
GLORYS12	-	0.1	0.04(0.108)	-		-	0.1	0.11(<0.005)	-
MPI-ESM1-2-LR	2.5	0.1	0.05(0.015)	0.52		0.2	0.0	0.05(<0.005)	0.38
IPSL-CM6A-LR	-0.4	0.0	-0.01(0.159)	0.51		0.4	0.0	0.05(<0.005)	0.18
MPI-ESM1-2-HR	0.9	0.1	0.19(<0.005)	0.44		-0.1	0.1	0.08(<0.005)	0.80
UKESM1-0-LL	-1.1	0.1	0.09(<0.005)	0.72		-0.6	0.0	0.06(<0.005)	0.26
EC-Earth3	-0.8	0.1	0.16(<0.005)	0.60		0.2	0.1	0.18(<0.005)	0.24
CESM2-WACCM	3.3	0.1	0.14(<0.005)	0.31		0.0	0.0	0.02(0.089)	0.22
	SBS					SC			
GLORYS12	-	0.2	0.06(0.256)	-		-	0.1	0.05(0.265)	-
MPI-ESM1-2-LR	0.3	0.1	0.11(<0.005)	0.44		-0.4	0.1	0.08(0.051)	0.53
IPSL-CM6A-LR	0.3	0.1	0.05(0.100)	0.43		0.0	0.2	0.02(0.681)	0.88
MPI-ESM1-2-HR	-0.1	0.2	0.18(0.015)	0.30		-0.2	0.2	0.22(<0.005)	0.43
UKESM1-0-LL	-0.6	0.1	0.07(0.057)	0.32		-0.6	0.1	0.04(0.263)	0.31
EC-Earth3	0.1	0.2	0.09(0.110)	0.90		0.3	0.3	0.30(<0.005)	0.13
CESM2-WACCM	0.2	0.1	0.03(0.239)	0.37		-0.2	0.2	0.06(0.240)	0.70

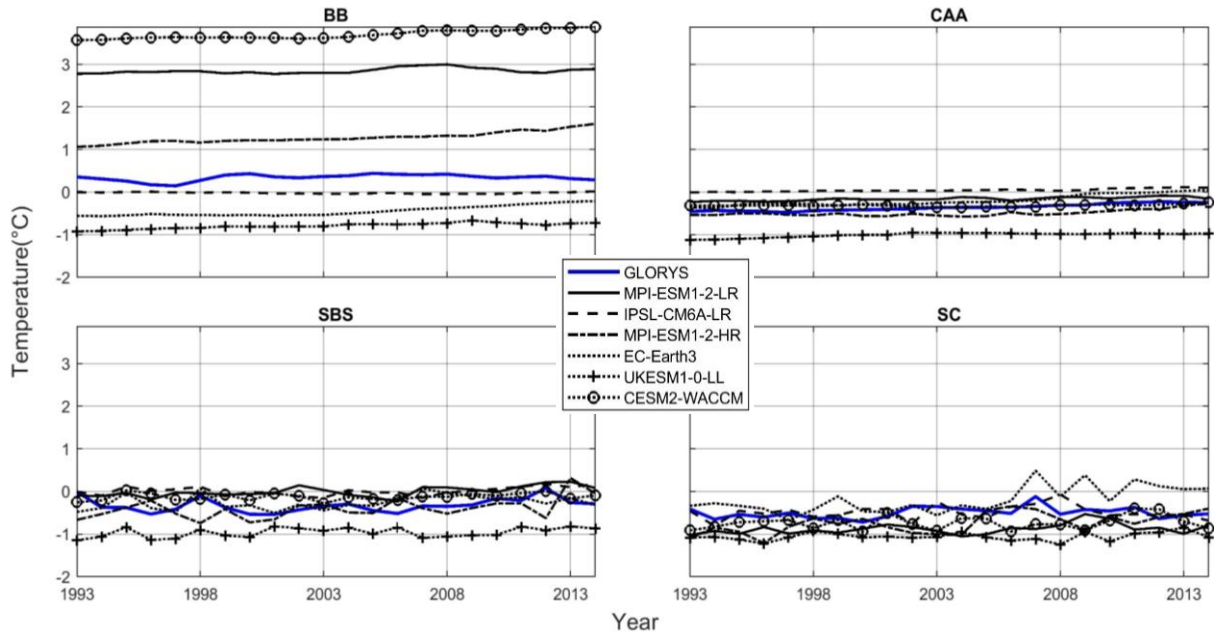

Figure 18. Regionally-averaged annual time series of bottom temperature from top four CMIP6 models (two models shares the fourth place) (in black) for four Arctic regions, BB, CAA, SBS, and SC against GLORYS12 (in blue) for the period of 1993-2014.

In the other three regions, the top six models show some over-or under-estimation of bottom temperature but generally with smaller biases than BB. In BB, the region has changed from partially ice-covered to totally open-water surface in October during 1950-2021 in the northeast, and has thinner ice in the northwest (Ballinger et al., 2022) and a large sea-ice reduction during autumn (Onarheim et al., 2018). This ice retreat resulted from transport of warm Atlantic water into this region (Ballinger et al., 2022) could cause warming and larger temperature range biases in the CMIP models in comparison to other three Arctic regions.

3.2.3 Pacific Shelf Water

The Pacific Shelf is divided into four regions consisting of the BS, EBS, AS, and BCS, and the ranking table, the statistical measures, and plots are provided. The region averages of the top four models are very similar (Table 13). The top model for simulation of the BT in this region is ACCESS-CM2 with an average rank of 2.0, followed by AWI-CM-1-1-MR and CNRM-CM6-1-HR, each tied with region averages of 3.0. Fourth place belongs to CESM2 with a region average of 3.5. These four models show consistent ranks in the four regions on the Pacific shelf, while CESM2 shows better rank in AS (1st) and BCS (1st) than in the other two regions. AWI-CM-1-1-MR has a lower rank in AS (6th) than for the other three regions. Regarding the models with the lowest ranks in the Pacific Shelf, CMC-CM2-SR5 and ACCESS-ESM1-5 are among the poorest rankings for BT with region averages of 20.3 and 19.5, respectively.

Table 14 shows the breakdown of the statistical measures. While first (ACCESS-CM2) and fourth (CESM2) do not show a consistent warm or cold bias in these regions, the second best models, CNRM-CM6-1-HR and AWI-CM-1-1-MR, always show warm and cold biases in the four regions, respectively. All four top models show a small bottom temperature STD close to that of GLORYS12. As for the trend, the top model, ACCESS-CM2, closely captures the negative trends in the bottom temperature on the Pacific shelf, except on the BCS (0.04 °C/decade) while GLORYS12 has a strong negative trend of -0.16 °C/decade. In the BCS region, CNRM-CM6-1-1-HR and CESM2 closely follow the trend, almost the same as the GLORYS12 (-0.16 °C/decade) at -0.13 and -0.14 °C/decade, respectively. In the North Pacific, GLORYS12 has p-value below 0.05 only at BCS, and CNRM-CM6-1-HR is the only model in this region that is aligned with GLORYS12 in showing statistically significant trend in BT in this region. In other three regions, while GLORYS12 have p-values above 0.05, not showing any significant trend, the top four models also do not show any statistically significant trend in BT.

Table 13. Pacific Shelf ranking is based on the Modified Kling-Gupta Efficiency (MKGE) for BT. The lower the rank, the better the model performance. The **ACCESS-CM2**, **AWI-CM-1-1-MR**, **CNRM-CM6-1-HR**, and **CESM2** get the overall rank of 1, 2, 2, and 4, respectively.

Models/Regions	Pacific Shelf					
	BS	EBS	AS	BCS	Region Average	Overall Rank
ACCESS-CM2	**1**	**1**	**3**	**3**	**2.0**	**1**
ACCESS-ESM1-5	18	19	22	19	19.5	21
AWI-CM-1-1-MR	**2**	**2**	**6**	**2**	**3.0**	**2**
CAMS-CSM1-0	13	14	12	5	11.0	9
CanESM5	14	15	21	18	17.0	18
CESM2	**6**	**6**	**1**	**1**	**3.5**	**4**
CESM2-WACCM	5	4	4	7	5.0	5
CMCC-CM2-SR5	22	22	20	17	20.3	22
CNRM-CM6-1	11	13	19	15	14.5	16
CNRM-CM6-1-HR	**3**	**3**	**2**	**4**	**3.0**	**2**
CNRM-ESM2-1	7	9	5	14	8.8	7
EC-Earth3	21	21	13	6	15.3	17
GISS-E2-1-G	19	18	18	13	17.0	18
IPSL-CM6A-LR	15	16	15	10	14.0	15
MIROC6	12	10	10	20	13.0	11
MIROC-ES2L	9	11	7	21	12.0	10
MPI-ESM1-2-HR	10	8	9	11	9.5	8
MPI-ESM1-2-LR	4	7	8	9	7.0	6
MRI-ESM2-0	16	17	11	8	13.0	11
NorESM2-LM	8	5	17	22	13.0	11
TaiESM1	17	12	14	12	13.8	14
UKESM1-0-LL	20	20	16	16	18.0	20

The time series plots for the top four models for the Pacific Shelf are plotted in Figure 19. CNRM-CM6-1-HR consistently overestimates BT and shows a warm bias in all four regions, while AWI-CM-1-1-MR shows the opposite with a cold bias on the North Pacific Shelf. While the models capture the variability of temperature, with STD close to GLORYS12, the bias gets larger, for most models, from the BS to the BCS.

Table 14. Bottom Temperature statistics for regions on the Pacific Shelf. Model Bias (Bs; unit: °C), standard deviation (STD; unit: °C), and trend (Tr; unit: °C/decade, p value in parentheses), and Modified Kling-Gupta Efficiency (MKGE) for the 1993-2014 period. The GLORYS12 evaluation dataset is indicated in blue.

Models/Regions	BS					EBS			
	Bs	STD	Tr(p)	MKGE		Bs	STD	Tr(p)	MKGE
GLORYS12	-	0.4	-0.21(0.108)	-		-	0.5	-0.27(0.132)	-
ACCESS-CM2	-0.1	0.4	-0.18(0.130)	0.93		0.4	0.5	-0.27(0.069)	0.71
CNRM-CM6-1-HR	0.3	0.4	0.05(0.755)	0.47		0.4	0.5	0.05(0.805)	0.47
AWI-CM-1-1-MR	-0.3	0.4	-0.10(0.408)	0.61		-0.5	0.5	-0.13(0.399)	0.50
CESM2	0.7	0.5	0.20(0.234)	0.29		0.9	0.6	0.22(0.277)	0.29
	AS					BCS			
GLORYS12	-	0.3	-0.15(0.170)	-		-	0.3	-0.16(0.052)	-
ACCESS-CM2	0.1	0.5	-0.11(0.532)	0.62		1.2	0.2	0.04(0.591)	0.39
CNRM-CM6-1-HR	1.2	0.4	-0.12(0.348)	0.68		3.0	0.2	-0.13(0.041)	0.33
AWI-CM-1-1-MR	-1.3	0.4	0.00(0.979)	0.38		-0.4	0.2	0.02(0.765)	0.46
CESM2	-0.2	0.3	-0.04(0.729)	0.70		0.4	0.3	-0.14(0.097)	0.92

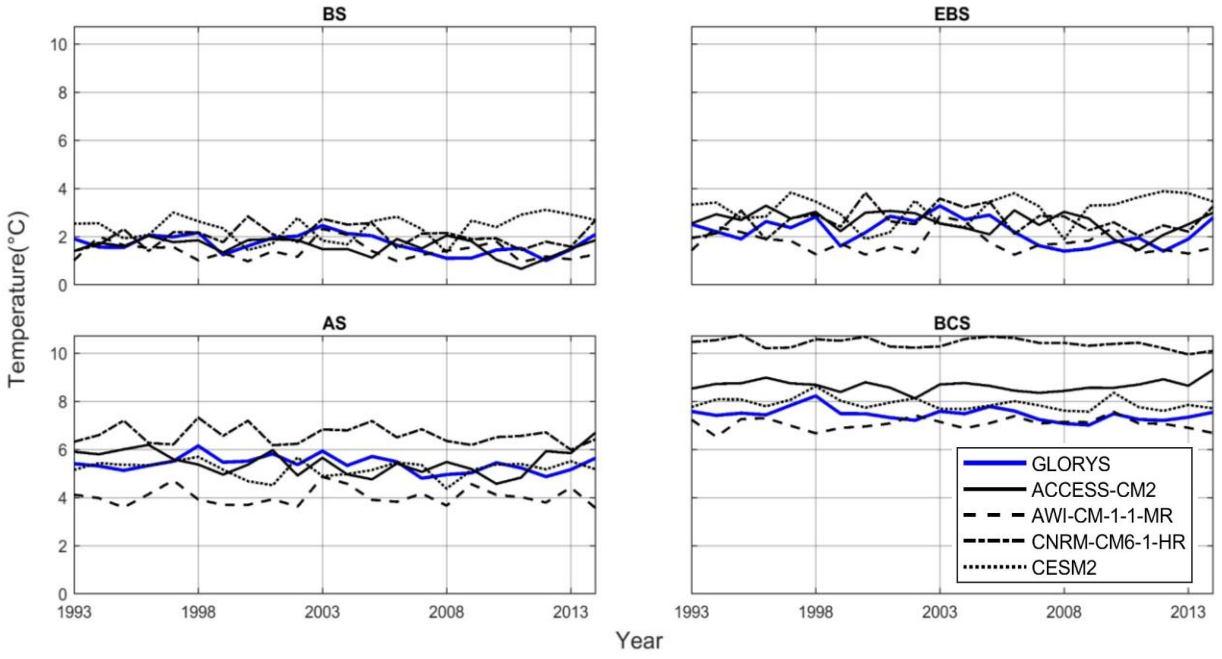

Figure 19. Regionally-averaged annual time series of bottom temperature from top four CMIP6 models (in black) for four Pacific regions, BS, EBS, AS, and BCS against GLORYS12 (in blue) for the period of 1993-2014.

3.2.3.1 Comparison Against Observations

The top four models' BT time series are plotted against the observations and GLORYS12 at the GAK1, NV, and SV stations, and in the EBS and BCS regions (Figure 20 and Figure 21). The breakdown of the three statistical measures for the top four models against the observation is provided in the Appendix (Table S *42* and Table S 43).

In the North Pacific, GLORYS12 shows higher bias than North Atlantic and is region dependent. GLORYS12 has a good agreement against observations in the EBS and GAK1 (Figure 20a-b). Reaching to the middle latitude, around Vancouver at NV and it shows a cool bias during the whole period, 1998-2014, and the cold bias increases to almost 2 degree between 2009 and 2011 (Figure 20c-d). On the other hand, in SV, the bias becomes warm for the whole period, with an unresolved spike by GLORYS12 around 2000 (Figure 21b). A colder bias exists in GLORYS12 in the BCS region for the entire period (Figure 21c). It seems that the bias increases in the bottom temperature in GLORYS12 from high to middle latitude in the Canadian North Pacific Shelf.

In the EBS (Figure 20a) only AWI-CM-1-1-MR shows a cold bias (-0.5 °C) and the other three models have warm biases with the maximum bias from ACCESS-CM2 and CESM2 (1.4 °C). In GAK1 (Figure 20b) ACCESS-CM2 and AWI-CM-1-1-MR have cold biases (-0.8 and -0.1 °C, respectively) and the other two models show warm biases with the maximum bias in CNRM-CM6-1-HR (1.2 °C). While the STD of the observations and

GLORYS12 is lower than the EBS region in the summer (0.3 vs. 0.8 °C), the models show a higher STD than the observations in GAK1. In NV (Figure 20c and Figure 20d) ACCESS-CM2 and AWI-CM-1-1-MR show cold biases both for spring (-0.2 and -0.9 °C, respectively) and fall (-0.7 and -1.2 °C, respectively) while the latter model has higher biases. The other two models, however, show warm biases for both seasons with higher bias in CNRM-CM6-1-HR (1.9 °C). At SV both in both fall and spring (Figure 21a and Figure 21b), the top four models have warm biases and CNRM-CM6-1-HR (4.3 °C) shows the highest bias. In comparison to the GLORYS12 STD, the four models show lower STDs in fall than in spring. Finally, in BCS (Figure 21c), while CNRM-CM6-1-HR still has the highest bias (3.7 °C) of the top models, they all show a warm bias in this region in spring and closely capture the observed STD. It is worth noting that, generally, GLORYS12 shows decent agreement with observations.

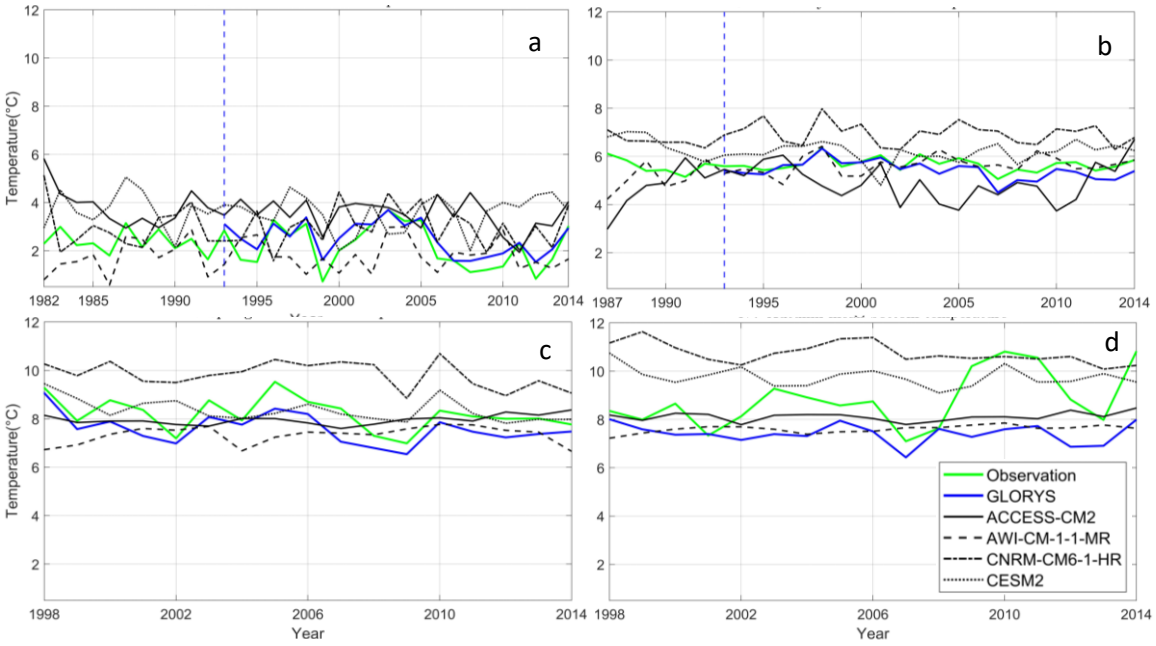

Figure 20. Regionally-averaged time series of bottom temperature from top four CMIP6 models (in black) against GLORYS12 (in blue) and observation (in green) for a) summer at EBS, b) yearly mean at GAK1 station, c) and d) spring and fall at NV. The vertical dashed blue line is the start of GLORYS12.

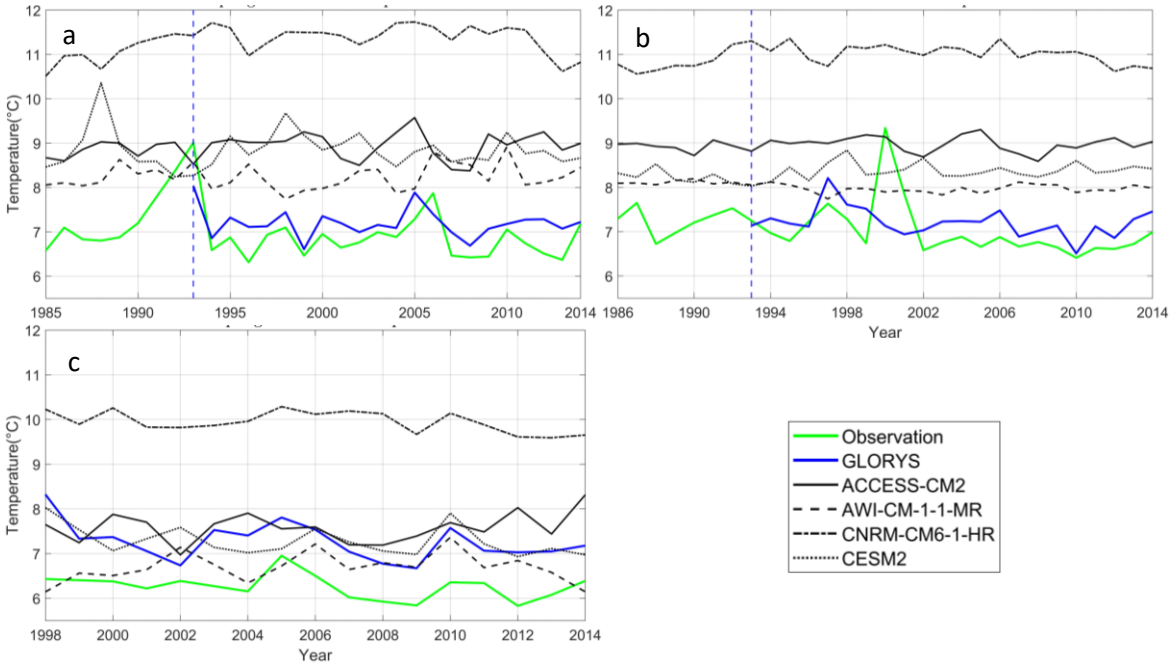

Figure 21. Regionally-averaged time series of bottom temperature from top four CMIP6 models (in black) against GLORYS12 (in blue) and observation (in green) for a and b) spring and fall at SV, and c) spring for BCS. The vertical dashed line is the start of GLORYS12 data in 1993.

4 Discussion

Our assessment of 22 CMIP6 models revealed that the models generally capture the observational STD, trend magnitude, and direction, while mostly having cold/warm systematic biases. The top four models are suggested based on a scoring metric which includes the normalized bias, standard deviation, and trend.

The top four models for SST on the North Atlantic shelf (Table 15) are: 1) CNRM-CM6-1-HR, 2) MRI-ESM2-0, 3) NorESM2-LM, and 4) MIROC6; on the Arctic shelf, 1) CESM2, 2) CESM2-WACCM, 3) MPI-ESM1-2-LR, and 4) MRI-ESM2-0; and on the North Pacific Shelf 1) ACCESS-CM2, 2) TaiESM1, 3) CESM2-WACCM, and 4) CNRM-CM6-1-HR. For the bottom temperature, the top four models on the North Atlantic shelf are 1) CNRM-ESM2-1, 2) CNRM-CM6-1-HR, 3) MRI-ESM2-0, and 4) MPI-ESM1-2-LR; in the Arctic shelf 1) MPI-ESM1-2-LR, 2) IPSL-CM6A-LR, 3) MPI-ESM1-2-HR, and 4) UKESM1-0-LL (tied with EC-Earth3, and CESM2-WACCM); and in the Pacific shelf 1) ACCESS-CM2, 2) CNRM-CM6-1-HR (tied with AWI-CM-1-1-MR), and 4) CESM2.

In the majority of the 11 regions in the North Atlantic, the 22 CIMP6 models generate a warm SST bias, around 63% of the time (Table S 45), but in the Arctic and Pacific, models mostly show cold SST biases,

with a comparable percentage of instances, around 58% and 59% (Table S 45), respectively. The only two models that show warm biases in all 11 regions in the North Atlantic are CNRM-ESM2-1 and MIROC-ES2L, with no model having a uniformly cold SST bias in these regions. In the Arctic, seven models (ACCESS-CM2, CanESM5, CESM2-WACCM, GISS-E2-1-G, MIROC6, TaiESM1, and UKESM1-0-LL) show cold SST biases in all four regions, and four models show warm biases (AWI-CM-1-1MR, CMCC-CM2-SR5, CNRM-CM6-1, and CNRM-ESM2-1) in all regions. In the North Atlantic, there are still six models (CNRM-CM6-1-, CNRM-CM6-1-HR, CNRM-ESM2-1, EC-Earth3, MIROC6, MRI-ESM2-0) with totally warm SST biases in all three regions, while nine models (ACCESS-CM2, AWI-CM-1-1-MR, CAMS-CSM1-0, CanESM5, GISS-E2-1-G, MPI-ESM1-2-HR, MPI-ESM1-2-LR, NorESM2-LM, and TaiESM1) have cold SST biases in all of these regions.

Regarding BT, the CMIP6 models mostly have warm biases in the North Atlantic and North Pacific; in 93% of cases for the North Atlantic and 61% for the North Pacific (Table S 46). Also, in the Atlantic, other than seven models (CAMS-CSM1-0, CNRM-CM6-1, CNRM-CM6-1-HR, GISS-E2-1-G, MRI-ESM2-0, NorESM2-LM, and TaiESM1) that show cold/warm biases in different regions, the remaining 15 models have consistent warm biases in all 11 regions. In the North Pacific, the majority of models show warm BT biases, around 61% of the time. Only four models (AWI-CM-1-1-MR, CAMS-CSM1-0, GISS-E2-1-G, and TaiESM1) have cold BT biases, while nine models (ACCESS-ESM1-5, CMCC-CM2-SR5, CNRM-CM6-1-HR, CNRM-ESM2-1, EC-Earth3, MPI-ESM1-2-HR, MRI-ESM2-0, and UKESM1-0-LL) show warm BT biases in all four regions. However, in the Arctic, the models show a greater number of instances with cold biases, around 49% of the time. Only two models have warm BT biases in the four regions in the Arctic (CESM2 and MRI-ESM2-0), while no model shows a uniformly cold BT bias.

It is worth mentioning that CNRM-ESM2-1 is the only model that has a warm SST bias in all three oceans, while there is no model with a consistent cold bias. Also, there is no model with a consistent cold or warm BT bias in the North Atlantic, Arctic, or North Pacific.

Taking into account of the performance for both the SST and BT, no model shows consistently high ranking (being in the top four models) for all three oceans. Table 15 shows the number of instances where the models rank the top four for both SST and BT in the three oceans. CNRM-CM6-1-HR was in the top four in four of six cases, while MPI-ESM1-2-LR, MPI-ESM2-0, and CESM2-WACCM were each in the top four in three of six cases.

Table 15. Number of instances that a model ranks in the top four models for the North Atlantic, Arctic, and North Pacific for the SST and BT. A model could be at the top four models for each ocean (3 instances) for SST and also for BT (3 instances) which ends up at 6 instances in total. The colored model names are those which have greater than one instance as the top four models.

Top four models for SST					Instances					
	Pacific	**Arctic**	**Atlantic**		CNRM-CM6-1-HR	MPI-ESM1-2-LR	MRI-ESM2-0	CESM2-WACCM	ACCESS-CM2	CESM2
1	ACCESS-CM2	CESM2	CNRM-CM6-1-HR		1				1	1
2	TaiESM1	CESM2-WACCM	MRI-ESM2-0				1	1		
3	CESM2-WACCM	MPI-ESM1-2-LR	NorESM2-LM			1		1		
4	CNRM-CM6-1-HR	MRI-ESM2-0	MIROC6		1		1			
Top four models for BT										
1	ACCESS-CM2	MPI-ESM1-2-LR	CNRM-ESM2-1			1			1	
2	AWI-CM-1-1-MR	IPSL-CM6A-LR	CNRM-CM6-1-HR		1					
3	CNRM-CM6-1-HR	MPI-ESM1-2-HR	MRI-ESM2-0		1		1			
4	CESM2	UKESM1-0-LL, CESM2-WACCM, EC-Earth3	MPI-ESM1-2-LR			1		1		1
Total number of instances for both SST and BT in three oceans					**4/6**	**3/6**	**3/6**	**3/6**	**2/6**	**2/6**

Table 16, Table 17, Table 18 provide the mean of both SST and BT ranks along with the regional average and final rank for North Atlantic, Arctic, and North Pacific shelves. For the North Atlantic, CNRM-CM6-1-HR gets the first rank (with the regional average rank of 3.8), and MRI-ESM2-0, MPI-ESM1-2-LR, and CNRM-ESM2-1 reach the second, third, and fourth, respectively (Table 16). In the Arctic, MPI-ESM1-2-LR, CESM2-WACCM, CESM2, and MPI-ESM1-2-HR rank first to fourth Table 17. The MPI-ESM1-2-LR are among the top four models, considering both SST and BT, at North Atlantic and Arctic. In the North Pacific, ACCESS-CM2, CNRM-CM6-1-HR, CESM2, and CESM2-WACCM are the top four models Table 18. Both CESM2, and CESM2-WACCM rank among the top four model in the Arctic and Pacific considering both SST and BT. For both North Atlantic and Pacific, only CNRM-CM6-1-HR is in the top four models.

Table 16. North Atlantic Shelf ranking based on the Modified Kling-Gupta Efficiency (MKGE) for both surface and bottom temperatures. The ranks for each region is the mean of ranks for SST and BT in that region. The lower the rank, the better the model performance.

Models/Region	North Atlantic Shelf												
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	Region Average	Rank
ACCESS-CM2	20.0	12.5	14.5	12.5	13.5	16.0	18.5	12.0	15.0	13.0	8.5	14.2	18
ACCESS-ESM1-5	21.5	21.0	20.0	17.5	18.0	22.0	14.5	14.5	12.5	8.5	9.0	16.3	21
AWI-CM-1-1-MR	6.5	9.0	11.0	8.5	4.5	8.0	10.5	13.5	9.5	11.5	17.0	10.0	7
CAMS-CSM1-0	8.5	9.0	6.5	6.5	4.5	13.5	11.0	18.0	15.5	10.5	15.0	10.8	9
CanESM5	10.0	14.5	12.5	9.5	12.0	8.5	11.0	10.5	10.0	10.0	17.0	11.4	10
CESM2	9.0	12.0	13.0	13.5	21.5	18.5	11.0	13.5	11.5	11.0	3.5	12.5	13
CESM2-WACCM	8.0	13.0	15.0	15.5	20.0	15.5	13.5	13.5	12.5	10.5	7.0	13.1	16
CMCC-CM2-SR5	10.0	11.0	11.0	9.0	11.0	8.5	9.5	4.5	6.5	6.0	12.5	9.0	5
CNRM-CM6-1	18.5	15.5	12.5	7.5	8.5	6.5	11.0	13.5	12.5	15.5	15.0	12.4	11
CNRM-CM6-1-HR	**1.5**	**2.0**	**2.0**	**4.0**	**4.0**	**3.5**	**3.0**	**3.0**	**6.0**	**9.0**	**4.0**	**3.8**	**1**
CNRM-ESM2-1	**8.0**	**5.0**	**1.5**	**3.0**	**10.0**	**7.5**	**9.5**	**11.0**	**12.5**	**13.0**	**14.0**	**8.6**	**4**
EC-Earth3	12.5	19.0	18.0	15.5	19.5	20.5	18.5	12.0	12.0	13.0	21.5	16.5	22
GISS-E2-1-G	10.5	9.0	10.0	14.5	13.5	18.0	21.0	13.5	12.0	10.5	11.5	13.1	16
IPSL-CM6A-LR	10.0	11.0	11.0	14.0	8.5	6.5	2.5	5.0	6.0	9.5	18.0	9.3	6
MIROC6	16.5	9.0	10.5	9.0	19.0	8.5	12.5	13.0	15.0	16.0	10.0	12.6	14
MIROC-ES2L	10.5	14.5	14.5	17.5	17.0	17.0	14.5	16.0	15.5	18.0	6.5	14.7	20
MPI-ESM1-2-HR	8.5	3.5	8.0	15.5	8.0	9.0	9.5	10.0	9.5	16.5	12.5	10.0	8
MPI-ESM1-2-LR	**12.5**	**9.5**	**11.5**	**9.0**	**5.0**	**8.0**	**6.0**	**8.0**	**6.5**	**6.5**	**11.5**	**8.5**	**3**
MRI-ESM2-0	**12.5**	**6.5**	**7.0**	**5.5**	**11.0**	**4.5**	**4.0**	**4.5**	**4.0**	**5.0**	**4.5**	**6.3**	**2**
NorESM2-LM	13.0	12.5	12.0	10.0	6.0	13.0	13.5	13.5	16.5	13.0	14.5	12.5	12
TaiESM1	12.5	12.5	13.0	17.5	10.5	12.0	13.0	14.0	15.5	11.0	8.0	12.7	15
UKESM1-0-LL	12.5	21.5	18.0	18.0	7.5	8.0	15.0	16.0	16.5	15.5	12.0	14.6	19

It is worth noting that the models are assessed against the bottom and sea surface temperatures and ranking the model performance is based on these two variables alone, hence it does not mean that the rankings of these models can be applied to other quantities. Also, the evaluation of model performance based just on a single ranking metric might only quantify a specific aspect of a model's skill (Hejazi & Moglen, 2008). As the performance metric is case-dependent and depends on the characteristics of the models, data, and the field of application, other performance indicators or metrics that encompass various aspects of the system could be used to generate a composite performance metric (Bennett et al., 2013).

Table 17. Arctic ranking based on the Modified Kling-Gupta Efficiency (MKGE) for both surface and bottom temperatures. The ranks for each region is the mean of ranks for SST and BT in that region. The lower the rank, the better the model performance.

Models/Regions	Arctic Shelf					
	BB	CAA	SBS	SC	Region Average	Rank
ACCESS-CM2	17.0	13.5	13.0	16.5	15.0	20
ACCESS-ESM1-5	7.5	10.5	15.0	3.5	9.1	7
AWI-CM-1-1-MR	10.0	15.5	10.0	9.5	11.3	10
CAMS-CSM1-0	13.0	17.0	12.0	12.5	13.6	16
CanESM5	12.0	15.0	11.0	15.0	13.3	14
CESM2	**12.0**	**10.5**	**7.0**	**3.0**	**8.1**	**3**
CESM2-WACCM	**11.0**	**9.0**	**4.5**	**2.5**	**6.8**	**2**
CMCC-CM2-SR5	7.0	17.5	11.5	22.0	14.5	18
CNRM-CM6-1	19.5	13.5	12.5	17.5	15.8	22
CNRM-CM6-1-HR	10.0	13.5	17.5	9.5	12.6	13
CNRM-ESM2-1	12.5	19.5	15.5	12.0	14.9	19
EC-Earth3	15.0	5.5	8.5	18.0	11.8	12
GISS-E2-1-G	13.0	11.0	10.5	19.0	13.4	15
IPSL-CM6A-LR	10.0	10.5	12.5	9.0	10.5	9
MIROC6	16.5	17.0	17.0	11.5	15.5	21
MIROC-ES2L	12.0	8.5	17.5	17.0	13.8	17
MPI-ESM1-2-HR	**6.5**	**4.5**	**14.0**	**9.0**	**8.5**	**4**
MPI-ESM1-2-LR	**5.0**	**4.0**	**7.5**	**9.0**	**6.4**	**1**
MRI-ESM2-0	9.5	12.0	15.0	9.5	11.5	11
NorESM2-LM	5.5	14.5	8.5	8.5	9.3	8
TaiESM1	19.0	6.5	5.0	5.5	9.0	6
UKESM1-0-LL	9.5	4.0	7.5	13.5	8.6	5

The last challenge in the ranking of model performance is to quantify why some models have better ranks than others. The Earth System Models differ in several key aspects, which could differentiate their performances. Most differences arise from atmospheric, land/surface, ocean, biogeochemical, coupling and feedback mechanisms. Each of these components has their own constituents and quantifying the source of uncertainty and errors is a demanding task. These uncertainties could result from model initialization, data assimilation technique (Weigel et al., 2008), and from the model itself for applying different approximation, discretization or imperfect open boundary conditions (Knutti et al., 2010; Schwierz et al., 2006). In the case of ocean model components, some of important constituents are the horizontal and vertical resolution of models, boundary condition and forcing, model coupling, and the initial condition and model spin-up.

While there is no correlation between the model performance and model resolution, we tried to understand what factors contribute to the significant difference in the model performance among these 22 models. However, it is challenging to make any suggestions on this matter due to the enormous volume

of components that are involved in each of these models. For example, we looked into the vertical and horizontal resolutions of the 22 models and thought the differences in the resolutions should be able to provide some information on the difference in the models' performances. Unfortunately, we could not draw any solid conclusions on this.

Table 18. Pacific Shelf ranking is based on the Modified Kling-Gupta Efficiency (MKGE) for both surface and bottom temperatures. The ranks for each region is the mean of ranks for SST and BT in that region. The lower the rank, the better the model performance.

Models/Region	Pacific Shelf				
	BS	AS	BCS	Region Average	Rank
ACCESS-CM2	**1.5**	**2.0**	**2.0**	**1.8**	**1**
ACCESS-ESM1-5	19.0	21.0	20.0	20.0	21
AWI-CM-1-1-MR	4.5	7.0	3.5	5.0	5
CAMS-CSM1-0	11.0	13.5	10.5	11.7	11
CanESM5	14.5	19.0	15.0	16.2	18
CESM2	**5.0**	**3.5**	**4.5**	**4.3**	**3**
CESM2-WACCM	**4.0**	**4.0**	**5.5**	**4.5**	**4**
CMCC-CM2-SR5	22.0	21.0	19.5	20.8	22
CNRM-CM6-1	13.5	14.5	16.5	14.8	17
CNRM-CM6-1-HR	**2.0**	**2.5**	**7.0**	**3.8**	**2**
CNRM-ESM2-1	8.5	6.0	12.5	9.0	7
EC-Earth3	21.0	12.0	10.5	14.5	16
GISS-E2-1-G	19.0	18.0	16.0	17.7	20
IPSL-CM6A-LR	13.0	10.0	6.5	9.8	9
MIROC6	10.0	11.5	17.0	12.8	13
MIROC-ES2L	13.0	13.0	17.0	14.3	15
MPI-ESM1-2-HR	11.0	9.0	8.5	9.5	8
MPI-ESM1-2-LR	8.5	14.5	14.5	12.5	12
MRI-ESM2-0	11.0	13.5	7.5	10.7	10
NorESM2-LM	11.0	15.5	15.5	14.0	14
TaiESM1	11.0	8.0	7.0	8.7	6
UKESM1-0-LL	19.0	14.0	16.5	16.5	19

In the case of SST, in the North Atlantic region, the statistically significant alignment in SST trend across top four models and HadISST enhances the reliability of these results, confirming that the HadISST and models trends are not due to random variation but represent a true underlying pattern. At the Arctic shelf, the top four ocean models exhibit p-values below 0.05, indicating statistically significant trends consistent with the HadISST data, which also has a p-value below 0.05. However, MRI-ESM2-0 and MPI-ESM1-2-LR show a p-value higher than 0.05 in BB and SBS, suggesting that the trend in these models are not statistically significant in those regions and may be due to random variation. Overall, the consistency among the majority of models and the observed data strengthens the evidence for a significant trend, but

the outlier models warrants additional scrutiny to understand its divergence. When comparing the top four models p-values trend against HadISST in North Pacific, the results show a mixed situation: some models have p-values below 0.05, indicating significant trends, while others have p-values above 0.05, indicating non-significant trends. This variability suggests that while some models are able to capture significant trends that are not evident in the HadISST observations, others do not align as closely with the HadISST data. This discrepancy highlights the variability in model performance and suggests that while the majority of models align well with the HadISST trend, there may be factors or assumptions in the outlier models that need further investigation.

For BT, in general, the top models show mixed results for the p-values in the North Atlantic region, some models have p-values below 0.05, while others have p-values above 0.05. In the Arctic, CAA is the only region where GLORYS12 and top models have statistically significant trend. In other three regions, while GLLORYS12 has p-value above 0.05, the top models show mixed results. In the North Pacific, however, BCS is the only region that GLORYS12 has statistically significant trend and only two of top models show this significant trend. In other three regions, the models are consistent with GLORYS12 in showing that there is no statistically significant trend in these regions.

For models with inconsistent p-value, a diagnostic test should be performed to understand the source of this discrepancy, whether it comes from model assumptions, or model complexity (too complex or too simple). A practical step is to plot the trends for both the observation and models to reveal any pattern or discrepancies. Also using other statistical test and cross validation techniques to assess the model robustness could be helpful. It is worth mentioning that there are some potential reasons where models or observation show high p-value, including true lack of trend, sample size issue (large o small), data quality (measurement errors or outliers), and variability in data. These factors need more assessments and reviewing the data and methods and conducting different diagnostic and considering alternative statistical models.

Acknowledgements

This work is supported by a Competitive Science Research Fund (CSFR; CC-22-05-01), funded by Fisheries and Oceans Canada (DFO), titled “The Performance and Projections of the CMIP6 Earth System Models (ESM) for Canada’s three Oceans”.

DFO colleagues Catherine Brennan and Rick Danielson internally reviewed the report, which helped improve the quality of the report.

References

Amante, C., & Eakins, B. (2009). ETOPO1 1 Arc-Minute Global Relief Model: procedures, data sources and analysis. https://doi.org/10.7289/V5C8276M

Amaya, D. J., Alexander, M. A., Scott, J. D., & Jacox, M. G. (2023). An evaluation of high-resolution ocean reanalyses in the California current system. *Progress in Oceanography*, *210*, 102951. https://doi.org/10.1016/j.pocean.2022.102951

Ballinger, T. J., Moore, G. W. K., Garcia-Quintana, Y., Myers, P. G., Imrit, A. A., Topál, D., & Meier, W. N. (2022). Abrupt Northern Baffin Bay Autumn Warming and Sea-Ice Loss Since the Turn of the Twenty-First Century. *Geophysical Research Letters*, *49*(21), e2022GL101472. https://doi.org/10.1029/2022GL101472

Bayr, T., Wengel, C., Latif, M., Dommenget, D., Lübbecke, J., & Park, W. (2019). Error compensation of ENSO atmospheric feedbacks in climate models and its influence on simulated ENSO dynamics. *Climate Dynamics*, *53*(1), 155-172. https://doi.org/10.1007/s00382-018-4575-7

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1-20. https://doi.org/10.1016/j.envsoft.2012.09.011

Buckley, T. W., Greig, A., & Boldt, J. L. (2009). *Describing summer pelagic habitat over the continental shelf in the eastern Bering Sea, 1982-2206.* U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-196, 49 p., Issue.

Castillo-Trujillo, A. C., Kwon, Y.-O., Fratantoni, P., Chen, K., Seo, H., Alexander, M. A., & Saba, V. S. (2023). An evaluation of eight global ocean reanalyses for the Northeast U.S. Continental shelf. *Progress in Oceanography*, *219*, 103126. https://doi.org/10.1016/j.pocean.2023.103126

Chelton, D. B., deSzoeke, R. A., Schlax, M. G., El Naggar, K., & Siwertz, N. (1998). Geographical Variability of the First Baroclinic Rossby Radius of Deformation. *Journal of Physical Oceanography*, *28*(3), 433-460. https://doi.org/https://doi.org/10.1175/1520-0485(1998)028<0433:GVOTFB>2.0.CO;2

Chelton, D. B., & Risien, C. (2016). *Zonal and Meridional Discontinuities and Other Issues with the HadISST1.1 Dataset*. https://doi.org/10.13140/RG.2.1.4503.0168

Claytor, R., Clark, D., Mclntyre, T., Stone, H., Cook, A., Harris, L., Simon, J., Emery, P., & Hurley, P. (2014). *Review of surveys contributing to groundfish assessments with recommendations for an ecosystem survey program in the Maritimes Region*. Can. Tech. Rep. Fish. Aquat. Sci. 3083: x +82 p., Issue.

Coyne, J., Cyr, F., Donnet, S., Galbraith, P., Geoffroy, M., Hebert, D., Layton, C., Ratsimandresy, A., Snook, S., Soontiens, N., & Walkusz, W. (2023). *Canadian Atlantic Shelf Temperature-Salinity (CASTS)*. https://doi.org/10.20383/102.0739

Cyr, F., Colbourne, E., Holden, J., Snook, S., G, H., Chen, N., Baily, W., Higdon, J., Lewis, S., Pye, B., & Senciall, D. (2019). *Physical oceanographic Conditions on the Newfoundland and Labrador Shelf during 2017*. Canadian Science Advisory Secretariat Research Document,2019/051, iv + 58 p., Issue.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., & Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553-597. https://doi.org/10.1002/qj.828

Deng, R., Qiao, S., Zhu, X., Dong, T., Feng, G., & Dong, W. (2023). The improvements of sea surface temperature simulation over China Offshore Sea in present climate from CMIP5 to CMIP6 models. *Climate Dynamics*. https://doi.org/10.1007/s00382-023-06843-2

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Ronald, S., & Taylor, K. (2015). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, *8*, 10539-10583. https://doi.org/10.5194/gmdd-8-10539-2015

Folland, C., Karl, T., Christy, J., Clarke, R., Gruza, G., Jouzel, J., Mann, M., Oerlemans, J., Salinger, M., & Wang, S. (2001). Observed climate variability and change. *Climate change*, *2001*, 99.

Greenan, B. J. W., Shackell, N. L., Ferguson, K., Greyson, P., Cogswell, A., Brickman, D., Wang, Z., Cook, A., Brennan, C. E., & Saba, V. S. (2019). Climate Change Vulnerability of American Lobster Fishing Communities in Atlantic Canada [Original Research]. *Frontiers in Marine Science*, *6*. https://doi.org/10.3389/fmars.2019.00579

Gregory, J. M., Stott, P., Cresswell, D., Rayner, N., Gordon, C., & Sexton, D. (2002). Recent and future changes in Arctic sea ice simulated by the HadCM3 AOGCM. *Geophysical Research Letters*, *29*(24), 28-21-28-24.

Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour, C. O., Dunne, J. P., Goddard, P., Morrison, A. K., Rosati, A., Wittenberg, A. T., Yin, J., & Zhang, R. (2015). Impacts on Ocean Heat from Transient Mesoscale Eddies in a Hierarchy of Climate Models. *Journal of Climate*, *28*(3), 952-977. https://doi.org/10.1175/JCLI-D-14-00353.1

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*, 80-91.

Hebert, D., Layton, C., Brickman, D., & Galbraith, P. S. (2021). *Physical Oceanographic Conditions on the Scotian Shelf and in the Gulf of Maine during 2020*. DFO Can. Sci. Advis. Sec. Res. Doc. 2021/070. iv + 55 p., Issue.

Hebert, D., Layton, C., Brickman, D., & Galbraith, P. S. (2023). *Physical Oceanographic Conditions on the Scotian Shelf and in the Gulf of Maine During 2022* (9780660496092). DFO Can. Tech. Rep. Hydrogr. Ocean. Sci. 359: vi + 81 p., Issue.

Hejazi, M. I., & Moglen, G. E. (2008). The effect of climate and land use change on flow duration in the Maryland Piedmont region. *Hydrological Processes: An International Journal*, *22*(24), 4710-4722.

ICES. (2004). *International Council for the Exploration of the Sea.Report of the Workshop on Survey Design and Data Analysis (WKSAD)*. https://ices-library.figshare.com/articles/report/Report_of_the_Workshop_on_Survey_Design_and_Data_Analysis_WKSAD_/19256492

ICES. (2005). *International Council for the Exploration of the Sea.Report of the Workshop on Survey Design and Data Analysis (WKSAD)*. https://ices-library.figshare.com/articles/report/Report_of_the_Workshop_on_Survey_Design_and_Data_Analysis_WKSAD_/19256519

Kearney, K. (2021). *Temperature Data from the Eastern Bering Sea Continental Shelf Bottom Trawl Survey as Used for Hydrodynamic Model Validation and Comparison*. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Alaska Fisheries Science Center. https://doi.org/10.25923/e77k-gg40

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, *23*(10), 2739-2758. https://doi.org/https://doi.org/10.1175/2009JCLI3361.1

Koelling, J., Atamanchuk, D., Wallace, D., & Karstensen, J. (2023). Decadal variability of oxygen uptake, export, and storage in the Labrador Sea from observations and CMIP6 models. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1202299

Lauth, R. R., Dawson, E. J., & Conner, J. (2019). *Results of the 2017 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna* [Technical Memorandum](NOAA technical memorandum NMFS AFSC ; 396, Issue. https://doi.org/10.25923/h118-nw41

Lellouche, J.-M., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L. G., Charly, R., Tony, C., Charles-Emmanuel, T., Florent, G., Giovanni, R., Mounir, B., Yann, D., & Pierre-Yves, L. T. (2021). The Copernicus Global 1/12° Oceanic and Sea Ice GLORYS12 Reanalysis [Original Research]. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.698876

Liu, H., Song, Z., Wang, X., & Misra, V. (2022). An ocean perspective on CMIP6 climate model evaluations. *Deep Sea Research Part II: Topical Studies in Oceanography*, *201*, 105120. https://doi.org/10.1016/j.dsr2.2022.105120

Loder, J. W., & Wang, Z. (2015). Trends and Variability of Sea Surface Temperature in the Northwest Atlantic from Three Historical Gridded Datasets. *Atmosphere Ocean*, *53*, 510-528. https://doi.org/10.1080/07055900.2015.1071237

Madec, G., Bourdallé-Badie, R., Jérôme, C., Clementi, E., Coward, A., Ethé, C., Iovino, D., Lea, D., Lévy, C., Lovato, T., Martin, N., Masson, S., Mocavero, S., Rousset, C., Storkey, D., Vancoppenolle, M., Müeller, S., Nurser, G., Bell, M., & Samson, G. (2019). *NEMO ocean engine*. Zenodo.

McKee, E., Wang, Z., & DeTracey, B. (2023). *Evaluation of bottom temperature from GLORYS12 and EN4 for North American continental shelf waters : from the north Atlantic, to the Arctic, to the north Pacific Oceans*. DFO Can. Tech. Rep. Hydrogr. Ocean. Sci. 355: vii + 34 p., Issue.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The Coupled Model Intercomparison Project (CMIP). *Bulletin of the American Meteorological Society*, *81*, 313-318.

Onarheim, I. H., Eldevik, T., Smedsrud, L. H., & Stroeve, J. C. (2018). Seasonal and Regional Manifestation of Arctic Sea Ice Loss. *Journal of Climate*, *31*(12), 4917-4932. https://doi.org/10.1175/JCLI-D-17-0427.1

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., & Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, *108*(D14). https://doi.org/10.1029/2002JD002670

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J. C., Kc, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L. A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J. C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., & Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, *42*, 153-168. https://doi.org/10.1016/j.gloenvcha.2016.05.009

Rickard, G. J., Behrens, E., Bahamondes Dominguez, A. A., & Pinkerton, M. H. (2023). An Assessment of the Oceanic Physical and Biogeochemical Components of CMIP5 and CMIP6 Models for the Ross Sea Region. *Journal of Geophysical Research (Oceans)*, *128*, e2022JC018880. https://doi.org/10.1029/2022jc018880

Schwierz, C., Appenzeller, C., Davies, H. C., Liniger, M. A., Müller, W., Stocker, T. F., & Yoshimori, M. (2006). Challenges posed by and approaches to the study of seasonal-to-decadal climate variability. *Climatic Change*, *79*(1), 31-63. https://doi.org/10.1007/s10584-006-9076-8

Stanley, R. R. E., DiBacco, C., Lowen, B., Beiko, R. G., Jeffery, N. W., Van Wyngaarden, M., Bentzen, P., Brickman, D., Benestan, L., Bernatchez, L., Johnson, C., Snelgrove, P. V. R., Wang, Z., Wringe, B. F., & Bradbury, I. R. (2018). A climate-associated multispecies cryptic cline in the northwest Atlantic. *Sci Adv*, *4*(3), eaaq0929. https://doi.org/10.1126/sciadv.aaq0929

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485-498. https://doi.org/10.1175/BAMS-D-11-00094.1

Wang, Y., Heywood, K. J., Stevens, D. P., & Damerell, G. M. (2022). Seasonal extrema of sea surface temperature in CMIP6 models. *Ocean Sci.*, *18*(3), 839-855. https://doi.org/10.5194/os-18-839-2022

Wang, Z., Brickman, D., Greenan, B., Christian, J., DeTracey, B., & Gilbert, D. (2023). Assessment of Ocean Temperature Trends for the Scotian Shelf and Gulf of Maine Using 22 CMIP6 Earth System Models. *Atmosphere-Ocean*, 1-11. https://doi.org/10.1080/07055900.2023.2264832

Wang, Z., Horwitz, R., Bowlby, H., Ding, F., & Joyce, W. (2020). Changes in ocean conditions and hurricanes affect porbeagle (Lamna nasus) diving behavior. *Marine Ecology Progress Series*, *654*, 219-224. https://doi.org/10.3354/meps13503

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241-260. https://doi.org/https://doi.org/10.1002/qj.210

Yang, M., Li, X., Shi, W., Zhang, C., & Zhang, J. (2020). The Pacific–Indian Ocean associated mode in CMIP5 models. *Ocean Sci.*, *16*(2), 469-482. https://doi.org/10.5194/os-16-469-2020

Yu, T., Chen, W., Gong, H., Feng, J., & Chen, S. (2023). Comparisons between CMIP5 and CMIP6 models in simulations of the climatology and interannual variability of the east asian summer Monsoon. *Climate Dynamics*, *60*, 2183-2198. https://doi.org/10.1007/s00382-022-06408-9

Appendix

S 1.Number of model grid cells for each region for the Atlantic, Arctic, and Pacific.

Models/Region	North Atlantic											Arctic				North Pacific		
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	BB	CAA	SBS	SC	BS	AS	BCS
ACCESS-CM2	21	3	4	10	36	26	21	21	25	19	227	476	587	150	613	282	57	18
ACCESS-ESM1-5	21	3	4	10	36	26	21	21	25	19	227	476	587	150	613	282	57	18
AWI-CM-1-1-MR	946	138	178	518	2511	1105	1279	2047	2105	988	4576	6786	5254	1648	15194	9542	1501	491
CAMS-CSM1-0	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	223	74	28
CanESM5	19	2	3	10	30	23	16	14	21	11	126	218	473	75	271	140	34	12
CESM2	36	5	8	19	64	50	37	35	43	23	267	562	486	73	335	196	54	19
CESM2-WACCM	36	5	8	19	64	50	37	35	43	23	267	562	486	73	335	196	54	19
CMCC-CM2-SR5	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	222	74	28
CNRM-CM6-1	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	222	74	28
CNRM-CM6-1-HR	420	63	82	215	766	475	396	371	463	248	5587	4131	10558	1640	5612	3531	1175	459
CNRM-ESM2-1	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	222	74	28
EC-Earth3	25	4	4	14	42	32	22	23	27	16	355	246	590	101	346	222	74	28
GISS-E2-1-G	15	2	2	9	23	18	11	11	16	8	99	135	254	39	228	113	31	9
IPSL-CM6A-LR	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	222	74	28
MIROC6	19	4	5	12	24	23	22	19	27	14	186	272	517	138	457	213	35	2
MIROC-ES2L	19	4	5	12	24	23	22	19	27	14	186	272	517	138	457	213	35	2
MPI-ESM1-2-HR	23	2	5	15	52	36	33	36	46	31	227	1452	447	40	142	60	17	9
MPI-ESM1-2-LR	116	16	18	55	162	106	84	73	81	40	967	468	897	145	654	462	132	61
MRI-ESM2-0	34	4	4	15	61	36	35	29	35	18	245	441	979	138	554	290	85	30
NorESM2-LM	23	5	5	13	43	29	25	24	31	18	434	388	771	173	570	294	86	26
TaiESM1	36	5	8	19	64	50	37	35	43	23	267	562	486	73	335	196	54	19
UKESM1-0-LL	25	4	4	14	42	32	22	23	27	16	355	246	557	101	346	222	74	28

S 2. Bias, standard deviation, trend, and MKGE score of SST for GoM.

GoM	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.7	-	0.23(<0.005)	-	-
ACCESS-CM2	1.5	0.2	0.6	0.8	0.09(0.028)	0.76	-0.16
ACCESS-ESM1-5	4.4	0.7	0.6	0.8	0.17(<0.005)	0.29	-0.11
AWI-CM-1-1-MR	-0.4	0.1	0.7	0.2	0.18(<0.005)	0.25	0.68
CAMS-CSM1-0	1.9	0.3	0.6	0.3	0.24(<0.005)	0.05	0.58
CanESM5	2.4	0.4	0.6	0.2	0.25(<0.005)	0.09	0.57
CESM2	5.8	0.9	0.7	0.2	0.23(<0.005)	0.00	0.08
CESM2-WACCM	5.9	0.9	0.7	0.0	0.26(<0.005)	0.14	0.08
CMCC-CM2-SR5	3.7	0.6	0.7	0.0	0.24(<0.005)	0.05	0.43
CNRM-CM6-1	1.7	0.2	0.7	0.1	0.05(0.313)	1.00	-0.04
CNRM-CM6-1-HR	-0.7	0.1	0.7	0.3	0.24(<0.005)	0.05	0.68
CNRM-ESM2-1	2.7	0.4	0.6	0.8	0.17(<0.005)	0.34	0.05
EC-Earth3	2.8	0.4	0.8	1.0	0.20(<0.005)	0.17	-0.10
GISS-E2-1-G	4.8	0.7	0.7	0.1	0.16(<0.005)	0.40	0.15
IPSL-CM6A-LR	2.2	0.3	0.6	0.2	0.15(<0.005)	0.45	0.40
MIROC6	2.5	0.4	0.8	0.9	0.29(<0.005)	0.30	-0.02
MIROC-ES2L	1.3	0.2	0.6	0.8	0.17(<0.005)	0.33	0.13
MPI-ESM1-2-HR	0.1	0.0	0.8	0.6	0.12(0.031)	0.61	0.16
MPI-ESM1-2-LR	-0.9	0.1	0.7	0.3	0.18(<0.005)	0.24	0.57
MRI-ESM2-0	1.0	0.1	0.8	0.9	0.13(0.029)	0.57	-0.08
NorESM2-LM	3.5	0.5	0.7	0.2	0.22(<0.005)	0.03	0.44
TaiESM1	6.5	1.0	0.7	0.2	0.25(<0.005)	0.08	-0.02
UKESM1-0-LL	0.6	0.1	0.7	0.4	0.15(<0.005)	0.44	0.38

S 3. Bias, standard deviation, trend, and MKGE score of SST for WSS.

WSS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.8	-	0.35(<0.005)	-	-
ACCESS-CM2	2.1	0.2	1.1	0.8	0.03(0.715)	0.94	-0.28
ACCESS-ESM1-5	5.5	0.7	0.5	0.9	0.16(<0.005)	0.50	-0.23
AWI-CM-1-1-MR	-0.8	0.0	0.6	0.6	0.14(<0.005)	0.55	0.20
CAMS-CSM1-0	2.7	0.3	0.9	0.1	0.29(<0.005)	0.04	0.67
CanESM5	3.8	0.5	0.9	0.3	0.19(<0.005)	0.38	0.35
CESM2	7.1	0.9	0.8	0.1	0.25(<0.005)	0.19	0.03
CESM2-WACCM	7.1	1.0	0.8	0.1	0.26(<0.005)	0.15	0.03
CMCC-CM2-SR5	4.7	0.6	0.9	0.1	0.23(<0.005)	0.27	0.34
CNRM-CM6-1	1.4	0.1	1.0	0.4	0.04(0.555)	0.89	0.01
CNRM-CM6-1-HR	-1.0	0.0	0.7	0.3	0.24(<0.005)	0.23	0.62
CNRM-ESM2-1	2.5	0.3	0.9	0.3	0.15(0.020)	0.51	0.36
EC-Earth3	3.4	0.4	1.1	1.0	0.19(0.021)	0.38	-0.15
GISS-E2-1-G	5.2	0.7	0.8	0.0	0.16(0.008)	0.49	0.17
IPSL-CM6A-LR	2.5	0.3	0.9	0.2	0.13(0.040)	0.58	0.33
MIROC6	2.6	0.3	0.8	0.0	0.27(<0.005)	0.11	0.69
MIROC-ES2L	1.9	0.2	0.6	0.6	0.13(<0.005)	0.59	0.13
MPI-ESM1-2-HR	-0.7	0.0	0.8	0.1	0.16(0.008)	0.49	0.51
MPI-ESM1-2-LR	-2.0	0.2	0.7	0.3	0.18(<0.005)	0.41	0.48
MRI-ESM2-0	1.1	0.1	0.8	0.0	0.16(0.006)	0.50	0.49
NorESM2-LM	3.5	0.4	0.8	0.0	0.20(<0.005)	0.34	0.47
TaiESM1	7.4	1.0	0.9	0.2	0.30(<0.005)	0.00	-0.02
UKESM1-0-LL	0.9	0.0	1.1	0.9	0.01(0.895)	1.00	-0.34

S 4. Bias, standard deviation, trend, and MKGE score of SST for CSS.

CSS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.8	-	0.34(<0.005)	-	-
ACCESS-CM2	2.0	0.3	1.0	0.6	0.05(0.525)	1.00	-0.22
ACCESS-ESM1-5	5.5	0.9	0.6	0.3	0.19(<0.005)	0.45	-0.04
AWI-CM-1-1-MR	0.0	0.0	0.6	0.4	0.15(<0.005)	0.61	0.30
CAMS-CSM1-0	1.0	0.2	0.9	0.2	0.31(<0.005)	0.00	0.72
CanESM5	1.3	0.2	0.9	0.2	0.21(<0.005)	0.38	0.52
CESM2	6.1	1.0	1.0	0.5	0.28(<0.005)	0.11	-0.08
CESM2-WACCM	6.2	1.0	1.0	0.5	0.28(<0.005)	0.09	-0.11
CMCC-CM2-SR5	2.2	0.3	0.7	0.1	0.19(<0.005)	0.44	0.44
CNRM-CM6-1	0.2	0.0	0.7	0.1	0.15(<0.005)	0.59	0.40
CNRM-CM6-1-HR	-0.9	0.1	0.7	0.2	0.24(<0.005)	0.25	0.64
CNRM-ESM2-1	1.1	0.2	0.8	0.0	0.26(<0.005)	0.17	0.76
EC-Earth3	0.4	0.1	1.2	0.9	0.21(0.010)	0.36	0.03
GISS-E2-1-G	4.2	0.7	1.0	0.5	0.15(0.031)	0.59	-0.01
IPSL-CM6A-LR	0.1	0.0	0.7	0.2	0.21(<0.005)	0.39	0.58
MIROC6	1.9	0.3	0.8	0.1	0.25(<0.005)	0.20	0.64
MIROC-ES2L	1.3	0.2	0.6	0.4	0.11(0.012)	0.75	0.14
MPI-ESM1-2-HR	-0.8	0.1	0.8	0.0	0.16(<0.005)	0.55	0.44
MPI-ESM1-2-LR	-2.0	0.3	0.7	0.3	0.21(<0.005)	0.37	0.45
MRI-ESM2-0	0.3	0.0	0.8	0.0	0.17(<0.005)	0.53	0.46
NorESM2-LM	1.5	0.2	0.7	0.1	0.22(<0.005)	0.34	0.58
TaiESM1	6.3	1.0	1.2	1.0	0.37(<0.005)	0.02	-0.41
UKESM1-0-LL	-1.6	0.3	0.9	0.3	0.08(0.208)	0.86	0.05

S 5. Bias, standard deviation, trend, and MKGE score of SST for ESS.

ESS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.7	-	0.28(<0.005)	-	-
ACCESS-CM2	0.3	0.1	0.8	0.2	0.15(0.008)	0.74	0.24
ACCESS-ESM1-5	3.5	1.0	0.8	0.3	0.28(<0.005)	0.00	-0.04
AWI-CM-1-1-MR	-0.3	0.1	0.6	0.2	0.15(<0.005)	0.73	0.24
CAMS-CSM1-0	0.1	0.0	0.8	0.3	0.30(<0.005)	0.09	0.69
CanESM5	0.6	0.2	0.8	0.1	0.21(<0.005)	0.36	0.57
CESM2	3.3	0.9	1.1	0.9	0.33(<0.005)	0.27	-0.32
CESM2-WACCM	3.4	1.0	1.2	1.0	0.31(<0.005)	0.17	-0.41
CMCC-CM2-SR5	1.0	0.3	0.7	0.1	0.17(<0.005)	0.58	0.35
CNRM-CM6-1	0.2	0.1	0.8	0.2	0.18(<0.005)	0.56	0.41
CNRM-CM6-1-HR	-1.1	0.3	0.6	0.2	0.22(<0.005)	0.35	0.50
CNRM-ESM2-1	1.1	0.3	0.7	0.1	0.26(<0.005)	0.11	0.66
EC-Earth3	-0.2	0.1	1.0	0.6	0.25(<0.005)	0.13	0.40
GISS-E2-1-G	3.4	1.0	1.0	0.6	0.14(<0.005)	0.79	-0.40
IPSL-CM6A-LR	0.0	0.0	0.7	0.1	0.19(<0.005)	0.46	0.53
MIROC6	1.1	0.3	0.7	0.0	0.23(<0.005)	0.27	0.58
MIROC-ES2L	1.3	0.4	0.6	0.2	0.10(0.023)	1.00	-0.08
MPI-ESM1-2-HR	-0.3	0.1	0.7	0.0	0.16(<0.005)	0.68	0.31
MPI-ESM1-2-LR	-2.1	0.6	0.6	0.2	0.20(<0.005)	0.44	0.25
MRI-ESM2-0	-0.7	0.2	0.7	0.1	0.20(<0.005)	0.46	0.48
NorESM2-LM	0.0	0.0	0.7	0.0	0.20(<0.005)	0.43	0.57
TaiESM1	3.2	0.9	1.1	1.0	0.37(<0.005)	0.48	-0.42
UKESM1-0-LL	-1.8	0.5	0.7	0.1	0.15(0.006)	0.73	0.10

S 6. Bias, standard deviation, trend, and MKGE score of SST for GSL.

GSL	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.5	-	0.20(<0.005)	-	-
ACCESS-CM2	-0.4	0.2	0.8	0.7	0.20(<0.005)	0.03	0.31
ACCESS-ESM1-5	1.6	0.6	0.8	0.8	0.30(<0.005)	0.86	-0.32
AWI-CM-1-1-MR	0.5	0.2	0.6	0.1	0.17(<0.005)	0.36	0.57
CAMS-CSM1-0	-0.2	0.1	0.6	0.1	0.21(<0.005)	0.02	0.86
CanESM5	1.1	0.4	0.6	0.4	0.24(<0.005)	0.31	0.35
CESM2	0.3	0.1	0.9	1.0	0.31(<0.005)	1.00	-0.41
CESM2-WACCM	0.2	0.1	0.9	1.0	0.31(<0.005)	0.95	-0.38
CMCC-CM2-SR5	1.5	0.6	0.7	0.4	0.22(<0.005)	0.10	0.29
CNRM-CM6-1	1.1	0.4	0.7	0.6	0.20(<0.005)	0.08	0.26
CNRM-CM6-1-HR	0.5	0.2	0.5	0.1	0.15(<0.005)	0.46	0.50
CNRM-ESM2-1	1.8	0.7	0.7	0.6	0.29(<0.005)	0.83	-0.27
EC-Earth3	-0.7	0.3	0.8	0.9	0.30(<0.005)	0.87	-0.28
GISS-E2-1-G	0.9	0.4	0.7	0.5	0.16(<0.005)	0.43	0.25
IPSL-CM6A-LR	1.0	0.4	0.5	0.0	0.18(<0.005)	0.27	0.54
MIROC6	2.5	1.0	0.7	0.5	0.21(<0.005)	0.00	-0.13
MIROC-ES2L	2.4	0.9	0.6	0.2	0.14(<0.005)	0.61	-0.15
MPI-ESM1-2-HR	0.0	0.0	0.5	0.1	0.14(<0.005)	0.59	0.40
MPI-ESM1-2-LR	-0.9	0.3	0.5	0.0	0.18(<0.005)	0.25	0.58
MRI-ESM2-0	-0.2	0.1	0.5	0.1	0.15(<0.005)	0.52	0.46
NorESM2-LM	-0.6	0.2	0.6	0.2	0.19(<0.005)	0.14	0.65
TaiESM1	0.8	0.3	0.6	0.2	0.19(<0.005)	0.11	0.62
UKESM1-0-LL	-0.8	0.3	0.7	0.5	0.21(<0.005)	0.09	0.39

S 7. Bias, standard deviation, trend, and MKGE score of SST for SNS.

SNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.7	-	0.20(<0.005)	-	-
ACCESS-CM2	0.2	0.0	0.9	0.5	0.15(0.020)	0.30	0.41
ACCESS-ESM1-5	2.6	1.0	1.0	0.8	0.28(<0.005)	0.52	-0.38
AWI-CM-1-1-MR	-1.3	0.5	0.5	0.5	0.09(0.006)	0.63	0.07
CAMS-CSM1-0	-1.5	0.6	0.8	0.2	0.24(<0.005)	0.27	0.35
CanESM5	-0.8	0.3	0.8	0.3	0.16(<0.005)	0.21	0.54
CESM2	2.1	0.8	0.8	0.4	0.31(<0.005)	0.70	-0.15
CESM2-WACCM	2.2	0.8	0.9	0.5	0.29(<0.005)	0.55	-0.12
CMCC-CM2-SR5	0.6	0.2	0.7	0.0	0.14(<0.005)	0.32	0.61
CNRM-CM6-1	0.1	0.0	0.8	0.4	0.23(<0.005)	0.19	0.52
CNRM-CM6-1-HR	-1.0	0.3	0.6	0.0	0.22(<0.005)	0.16	0.61
CNRM-ESM2-1	0.9	0.3	0.8	0.3	0.26(<0.005)	0.38	0.43
EC-Earth3	-0.8	0.3	1.1	1.0	0.28(<0.005)	0.52	-0.17
GISS-E2-1-G	2.0	0.8	0.7	0.2	0.03(0.538)	1.00	-0.28
IPSL-CM6A-LR	0.1	0.0	0.8	0.3	0.23(<0.005)	0.19	0.67
MIROC6	-0.1	0.0	0.7	0.1	0.22(<0.005)	0.12	0.86
MIROC-ES2L	1.2	0.4	0.6	0.2	0.05(0.283)	0.93	-0.05
MPI-ESM1-2-HR	-1.2	0.4	0.6	0.2	0.18(<0.005)	0.08	0.52
MPI-ESM1-2-LR	-1.5	0.5	0.6	0.1	0.19(<0.005)	0.05	0.45
MRI-ESM2-0	-0.7	0.2	0.7	0.0	0.20(<0.005)	0.00	0.77
NorESM2-LM	-0.6	0.2	0.6	0.2	0.14(<0.005)	0.37	0.53
TaiESM1	2.5	1.0	0.6	0.0	0.18(<0.005)	0.09	0.04
UKESM1-0-LL	-1.9	0.7	0.8	0.4	0.19(<0.005)	0.01	0.20

S 8. Bias, standard deviation, trend, and MKGE score of SST for CNS.

CNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.6	-	0.19(<0.005)	-	-
ACCESS-CM2	-0.7	0.2	1.0	1.0	0.22(<0.005)	0.14	0.02
ACCESS-ESM1-5	0.8	0.2	0.9	0.8	0.28(<0.005)	0.42	0.04
AWI-CM-1-1-MR	-1.0	0.3	0.5	0.3	0.13(<0.005)	0.27	0.48
CAMS-CSM1-0	-2.2	0.9	0.6	0.0	0.18(<0.005)	0.02	0.13
CanESM5	-2.5	1.0	0.6	0.1	0.12(<0.005)	0.34	-0.06
CESM2	1.1	0.3	0.7	0.1	0.24(<0.005)	0.27	0.55
CESM2-WACCM	1.1	0.4	0.7	0.2	0.25(<0.005)	0.30	0.49
CMCC-CM2-SR5	-1.1	0.4	0.5	0.3	0.11(<0.005)	0.38	0.39
CNRM-CM6-1	-0.3	0.0	0.9	0.6	0.34(<0.005)	0.71	0.05
CNRM-CM6-1-HR	-0.7	0.2	0.5	0.3	0.19(<0.005)	0.00	0.69
CNRM-ESM2-1	0.5	0.1	0.9	0.7	0.34(<0.005)	0.74	-0.01
EC-Earth3	-1.9	0.7	1.0	1.0	0.31(<0.005)	0.59	-0.36
GISS-E2-1-G	1.8	0.7	0.7	0.2	-0.02(0.634)	1.00	-0.23
IPSL-CM6A-LR	-0.4	0.0	0.6	0.0	0.18(<0.005)	0.05	0.94
MIROC6	-0.3	0.0	0.7	0.2	0.14(<0.005)	0.21	0.74
MIROC-ES2L	0.5	0.1	0.8	0.3	0.07(0.217)	0.56	0.34
MPI-ESM1-2-HR	-1.0	0.3	0.5	0.2	0.16(<0.005)	0.14	0.60
MPI-ESM1-2-LR	-1.2	0.4	0.6	0.0	0.22(<0.005)	0.16	0.55
MRI-ESM2-0	-0.5	0.1	0.6	0.2	0.19(<0.005)	0.01	0.81
NorESM2-LM	-0.9	0.3	0.6	0.0	0.11(0.012)	0.37	0.54
TaiESM1	1.4	0.5	0.5	0.4	0.04(0.187)	0.67	0.08
UKESM1-0-LL	-2.3	0.9	0.9	0.7	0.25(<0.005)	0.31	-0.17

S 9. Bias, standard deviation, trend, and MKGE score of SST for NNS.

NNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.5	-	0.16(<0.005)	-	-
ACCESS-CM2	-0.9	0.4	0.9	0.7	0.27(<0.005)	0.32	0.10
ACCESS-ESM1-5	-0.2	0.1	1.1	0.9	0.37(<0.005)	0.65	-0.14
AWI-CM-1-1-MR	1.2	0.6	0.6	0.2	0.16(<0.005)	0.00	0.41
CAMS-CSM1-0	-1.6	0.8	0.5	0.0	0.12(<0.005)	0.13	0.21
CanESM5	-2.1	1.0	0.6	0.2	0.15(<0.005)	0.04	-0.02
CESM2	0.7	0.3	0.6	0.2	0.20(<0.005)	0.10	0.60
CESM2-WACCM	0.8	0.4	0.6	0.2	0.23(<0.005)	0.20	0.51
CMCC-CM2-SR5	-0.3	0.1	0.6	0.1	0.16(<0.005)	0.00	0.81
CNRM-CM6-1	0.7	0.3	0.8	0.6	0.35(<0.005)	0.59	0.12
CNRM-CM6-1-HR	0.3	0.1	0.4	0.1	0.15(<0.005)	0.03	0.85
CNRM-ESM2-1	1.5	0.7	1.0	0.7	0.38(<0.005)	0.68	-0.24
EC-Earth3	-1.9	0.9	1.1	1.0	0.49(<0.005)	1.00	-0.67
GISS-E2-1-G	1.0	0.5	0.5	0.1	-0.04(0.300)	0.62	0.22
IPSL-CM6A-LR	0.4	0.2	0.6	0.2	0.17(<0.005)	0.02	0.75
MIROC6	0.2	0.1	0.6	0.2	0.12(<0.005)	0.11	0.75
MIROC-ES2L	0.9	0.4	0.8	0.5	0.12(0.034)	0.11	0.32
MPI-ESM1-2-HR	0.5	0.3	0.6	0.2	0.15(<0.005)	0.03	0.67
MPI-ESM1-2-LR	0.0	0.0	0.7	0.3	0.24(<0.005)	0.23	0.61
MRI-ESM2-0	-0.3	0.1	0.5	0.0	0.17(<0.005)	0.01	0.85
NorESM2-LM	-0.7	0.3	0.6	0.2	0.10(0.012)	0.17	0.61
TaiESM1	1.1	0.5	0.4	0.0	0.01(0.803)	0.47	0.28
UKESM1-0-LL	-1.6	0.8	1.0	0.8	0.32(<0.005)	0.49	-0.22

S 10. Bias, standard deviation, trend, and MKGE score of SST for SLS.

SLS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.4	-	0.15(<0.005)	-	-
ACCESS-CM2	-1.0	0.4	0.8	0.8	0.23(<0.005)	0.33	0.04
ACCESS-ESM1-5	-0.5	0.2	0.8	0.8	0.25(<0.005)	0.42	0.09
AWI-CM-1-1-MR	0.6	0.2	0.5	0.2	0.14(<0.005)	0.03	0.65
CAMS-CSM1-0	-1.6	0.8	0.4	0.1	0.05(<0.074)	0.41	0.10
CanESM5	-2.0	1.0	0.6	0.4	0.20(<0.005)	0.21	-0.10
CESM2	0.3	0.1	0.5	0.2	0.14(<0.005)	0.00	0.81
CESM2-WACCM	0.4	0.1	0.6	0.3	0.18(<0.005)	0.13	0.64
CMCC-CM2-SR5	-0.3	0.1	0.4	0.1	0.12(<0.005)	0.11	0.84
CNRM-CM6-1	0.9	0.4	0.7	0.7	0.33(<0.005)	0.75	-0.11
CNRM-CM6-1-HR	0.5	0.2	0.3	0.1	0.10(<0.005)	0.17	0.74
CNRM-ESM2-1	1.6	0.8	0.9	1.0	0.35(<0.005)	0.84	-0.52
EC-Earth3	-1.8	0.9	0.8	0.9	0.38(<0.005)	1.00	-0.59
GISS-E2-1-G	0.6	0.3	0.5	0.1	-0.04(<0.274)	0.78	0.17
IPSL-CM6A-LR	0.4	0.1	0.5	0.3	0.18(<0.005)	0.11	0.67
MIROC6	0.1	0.0	0.6	0.4	0.12(<0.005)	0.12	0.63
MIROC-ES2L	0.8	0.4	0.7	0.5	0.14(<0.005)	0.00	0.36
MPI-ESM1-2-HR	0.3	0.1	0.5	0.3	0.13(<0.005)	0.07	0.72
MPI-ESM1-2-LR	-0.2	0.0	0.6	0.4	0.19(<0.005)	0.17	0.55
MRI-ESM2-0	-0.3	0.1	0.4	0.0	0.13(<0.005)	0.07	0.88
NorESM2-LM	-0.8	0.3	0.5	0.2	0.08(0.025)	0.28	0.53
TaiESM1	0.7	0.3	0.5	0.1	-0.03(0.395)	0.74	0.20
UKESM1-0-LL	-1.5	0.7	0.8	0.7	0.25(<0.005)	0.41	-0.09

S 11. Bias, standard deviation, trend, and MKGE score of SST for NLS.

NLS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.3	-	0.08(<0.005)	-	-
ACCESS-CM2	-0.4	0.2	0.5	0.4	0.15(<0.005)	0.29	0.45
ACCESS-ESM1-5	-0.5	0.3	0.5	0.4	0.13(<0.005)	0.23	0.48
AWI-CM-1-1-MR	0.7	0.4	0.4	0.1	0.11(<0.005)	0.11	0.60
CAMS-CSM1-0	-1.1	0.6	0.3	0.1	0.00(0.934)	0.39	0.30
CanESM5	-1.3	0.7	0.5	0.5	0.21(<0.005)	0.58	-0.06
CESM2	0.0	0.0	0.4	0.1	0.11(<0.005)	0.11	0.85
CESM2-WACCM	0.0	0.0	0.4	0.2	0.14(<0.005)	0.27	0.64
CMCC-CM2-SR5	0.4	0.2	0.4	0.2	0.10(<0.005)	0.08	0.73
CNRM-CM6-1	1.3	0.7	0.7	1.0	0.28(<0.005)	0.90	-0.49
CNRM-CM6-1-HR	0.8	0.4	0.3	0.0	0.08(<0.005)	0.00	0.57
CNRM-ESM2-1	1.8	1.0	0.7	1.0	0.26(<0.005)	0.83	-0.64
EC-Earth3	-0.9	0.5	0.7	1.0	0.30(<0.005)	1.00	-0.49
GISS-E2-1-G	-0.4	0.2	0.4	0.2	0.00(0.960)	0.39	0.54
IPSL-CM6A-LR	1.2	0.6	0.5	0.4	0.14(<0.005)	0.26	0.20
MIROC6	0.3	0.2	0.5	0.5	0.13(<0.005)	0.22	0.43
MIROC-ES2L	1.1	0.6	0.5	0.5	0.14(<0.005)	0.24	0.20
MPI-ESM1-2-HR	0.6	0.3	0.5	0.4	0.12(<0.005)	0.18	0.44
MPI-ESM1-2-LR	0.3	0.2	0.5	0.5	0.13(<0.005)	0.20	0.46
MRI-ESM2-0	0.4	0.2	0.4	0.1	0.09(<0.005)	0.05	0.78
NorESM2-LM	-0.4	0.2	0.4	0.1	0.05(0.033)	0.13	0.73
TaiESM1	0.0	0.0	0.3	0.1	0.02(0.282)	0.28	0.69
UKESM1-0-LL	-0.7	0.4	0.5	0.6	0.18(<0.005)	0.45	0.16

S 12. Bias, standard deviation, trend, and MKGE score of SST for HB.

HB	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.3	-	0.12(<0.005)	-	-
ACCESS-CM2	-0.2	0.1	0.2	0.2	0.06(<0.005)	0.28	0.66
ACCESS-ESM1-5	0.1	0.0	0.2	0.2	0.06(<0.005)	0.29	0.67
AWI-CM-1-1-MR	1.2	0.7	0.4	0.3	0.10(<0.005)	0.00	0.23
CAMS-CSM1-0	-0.3	0.1	0.3	0.1	0.00(<0.870)	0.75	0.23
CanESM5	0.2	0.1	0.5	0.6	0.23(<0.005)	0.66	0.10
CESM2	-0.6	0.3	0.3	0.0	0.13(<0.005)	0.00	0.68
CESM2-WACCM	-0.6	0.3	0.4	0.1	0.15(<0.005)	0.10	0.63
CMCC-CM2-SR5	1.1	0.6	0.5	0.6	0.17(<0.005)	0.23	0.10
CNRM-CM6-1	1.1	0.6	0.6	0.8	0.20(<0.005)	0.48	-0.10
CNRM-CM6-1-HR	0.5	0.3	0.4	0.2	0.09(<0.005)	0.12	0.67
CNRM-ESM2-1	1.5	0.9	0.6	0.9	0.22(<0.005)	0.60	-0.43
EC-Earth3	-0.1	0.0	0.6	1.0	0.28(<0.005)	1.00	-0.41
GISS-E2-1-G	-1.7	1.0	0.2	0.3	0.04(<0.013)	0.48	-0.15
IPSL-CM6A-LR	1.4	0.8	0.5	0.4	0.16(<0.005)	0.18	0.04
MIROC6	0.5	0.3	0.5	0.6	0.14(<0.005)	0.07	0.34
MIROC-ES2L	0.9	0.5	0.4	0.1	0.10(<0.005)	0.02	0.46
MPI-ESM1-2-HR	0.6	0.3	0.4	0.2	0.09(<0.005)	0.12	0.60
MPI-ESM1-2-LR	0.6	0.3	0.4	0.3	0.14(<0.005)	0.08	0.55
MRI-ESM2-0	0.1	0.0	0.4	0.2	0.10(<0.005)	0.03	0.76
NorESM2-LM	-0.6	0.3	0.4	0.1	0.14(<0.005)	0.04	0.65
TaiESM1	-0.6	0.3	0.3	0.0	0.10(<0.005)	0.00	0.70
UKESM1-0-LL	-0.5	0.2	0.4	0.2	0.14(<0.005)	0.02	0.67

S 13. Bias, standard deviation, trend, and MKGE score of SST for BB.

BB	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.2	-	0.07(<0.005)	-	-
ACCESS-CM2	-0.7	0.5	0.3	0.2	0.07(<0.005)	0.00	0.43
ACCESS-ESM1-5	-0.4	0.3	0.3	0.2	0.10(<0.005)	0.13	0.61
AWI-CM-1-1-MR	0.4	0.3	0.3	0.2	0.10(<0.005)	0.12	0.60
CAMS-CSM1-0	-0.6	0.4	0.2	0.0	0.01(0.541)	0.28	0.49
CanESM5	-1.2	1.0	0.2	0.1	0.10(<0.005)	0.12	0.01
CESM2	-0.6	0.4	0.2	0.0	0.07(<0.005)	0.00	0.56
CESM2-WACCM	-0.6	0.5	0.2	0.1	0.07(<0.005)	0.01	0.53
CMCC-CM2-SR5	0.6	0.4	0.3	0.3	0.10(<0.005)	0.13	0.46
CNRM-CM6-1	0.6	0.5	0.6	0.8	0.24(<0.005)	0.79	-0.24
CNRM-CM6-1-HR	0.9	0.7	0.4	0.3	0.09(<0.005)	0.09	0.23
CNRM-ESM2-1	0.9	0.7	0.6	0.8	0.18(<0.005)	0.47	-0.12
EC-Earth3	-0.8	0.7	0.7	1.0	0.29(<0.005)	1.00	-0.56
GISS-E2-1-G	-1.2	1.0	0.1	0.2	0.01(0.252)	0.29	-0.07
IPSL-CM6A-LR	-0.4	0.3	0.4	0.3	0.07(<0.005)	0.01	0.56
MIROC6	-0.8	0.6	0.2	0.0	0.03(0.016)	0.16	0.38
MIROC-ES2L	0.1	0.0	0.4	0.3	0.05(<0.045)	0.08	0.69
MPI-ESM1-2-HR	-0.2	0.1	0.3	0.2	0.11(<0.005)	0.17	0.74
MPI-ESM1-2-LR	0.2	0.1	0.3	0.1	0.04(<0.054)	0.14	0.78
MRI-ESM2-0	-0.2	0.1	0.3	0.2	0.03(<0.191)	0.20	0.71
NorESM2-LM	-0.6	0.5	0.2	0.0	0.03(<0.065)	0.18	0.50
TaiESM1	-0.9	0.7	0.2	0.0	0.04(<0.005)	0.13	0.31
UKESM1-0-LL	-0.7	0.5	0.3	0.2	0.13(<0.005)	0.26	0.37

S 14. Bias, standard deviation, trend, and MKGE score of SST for CAA.

CAA	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.1	-	0.05(<0.005)	-	-
ACCESS-CM2	0.0	0.0	0.1	0.1	0.03(<0.005)	0.30	0.68
ACCESS-ESM1-5	0.2	0.2	0.1	0.0	0.02(0.031)	0.44	0.53
AWI-CM-1-1-MR	0.7	0.8	0.2	0.4	0.05(<0.005)	0.00	0.06
CAMS-CSM1-0	0.1	0.1	0.0	0.4	0.00(0.890)	0.74	0.17
CanESM5	0.0	0.0	0.1	0.1	0.02(<0.005)	0.37	0.60
CESM2	-0.1	0.1	0.1	0.2	0.04(<0.005)	0.14	0.78
CESM2-WACCM	-0.1	0.1	0.1	0.2	0.03(<0.005)	0.26	0.66
CMCC-CM2-SR5	0.8	1.0	0.3	1.0	0.11(<0.005)	1.00	-0.73
CNRM-CM6-1	0.5	0.6	0.3	0.6	0.09(<0.005)	0.60	-0.05
CNRM-CM6-1-HR	0.2	0.3	0.1	0.1	0.02(0.087)	0.47	0.46
CNRM-ESM2-1	0.6	0.7	0.2	0.5	0.06(<0.005)	0.18	0.10
EC-Earth3	0.1	0.1	0.1	0.0	0.03(<0.005)	0.27	0.71
GISS-E2-1-G	-0.2	0.2	0.0	0.4	0.01(0.008)	0.62	0.25
IPSL-CM6A-LR	0.4	0.4	0.2	0.3	0.06(<0.005)	0.07	0.50
MIROC6	-0.1	0.1	0.1	0.2	0.00(0.760)	0.71	0.26
MIROC-ES2L	-0.1	0.1	0.1	0.1	0.00(0.642)	0.69	0.29
MPI-ESM1-2-HR	0.2	0.2	0.2	0.3	0.06(<0.005)	0.06	0.61
MPI-ESM1-2-LR	0.0	0.0	0.1	0.1	0.03(<0.005)	0.29	0.70
MRI-ESM2-0	0.3	0.3	0.2	0.2	0.04(<0.005)	0.10	0.64
NorESM2-LM	-0.1	0.1	0.0	0.3	0.01(<0.005)	0.54	0.36
TaiESM1	-0.2	0.2	0.1	0.2	0.02(<0.005)	0.34	0.56
UKESM1-0-LL	-0.1	0.1	0.1	0.1	0.04(<0.005)	0.17	0.80

S 15. Bias, standard deviation, trend, and MKGE score of SST for SBS.

SBS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.4	-	0.07(0.006)	-	-
ACCESS-CM2	-0.1	0.0	0.5	0.3	0.06(0.097)	0.06	0.73
ACCESS-ESM1-5	0.1	0.1	0.7	0.8	0.04(0.482)	0.26	0.17
AWI-CM-1-1-MR	0.1	0.1	0.7	0.8	0.09(0.067)	0.10	0.19
CAMS-CSM1-0	-0.6	0.4	0.3	0.1	0.02(0.370)	0.39	0.40
CanESM5	-0.2	0.2	0.5	0.3	0.03(0.433)	0.30	0.52
CESM2	-0.2	0.2	0.4	0.0	0.10(<0.005)	0.13	0.78
CESM2-WACCM	-0.3	0.2	0.4	0.1	0.10(<0.005)	0.17	0.74
CMCC-CM2-SR5	1.1	0.8	0.6	0.7	0.18(<0.005)	0.85	-0.32
CNRM-CM6-1	0.8	0.6	0.8	1.0	0.17(<0.005)	0.76	-0.39
CNRM-CM6-1-HR	-0.1	0.0	0.5	0.4	-0.05(0.181)	1.00	-0.09
CNRM-ESM2-1	1.4	1.0	0.7	0.7	0.10(0.032)	0.20	-0.26
EC-Earth3	-0.2	0.1	0.7	0.7	0.16(<0.005)	0.71	-0.02
GISS-E2-1-G	-0.6	0.4	0.3	0.2	0.02(0.425)	0.42	0.38
IPSL-CM6A-LR	0.4	0.3	0.8	1.0	0.16(<0.005)	0.67	-0.22
MIROC6	-0.3	0.2	0.6	0.5	-0.03(0.477)	0.80	0.05
MIROC-ES2L	0.0	0.0	0.6	0.6	-0.01(0.855)	0.62	0.10
MPI-ESM1-2-HR	-0.5	0.3	0.6	0.7	0.18(<0.005)	0.81	-0.11
MPI-ESM1-2-LR	-0.6	0.4	0.6	0.6	0.06(0.207)	0.09	0.31
MRI-ESM2-0	0.0	0.0	0.6	0.5	0.14(<0.005)	0.47	0.33
NorESM2-LM	-0.2	0.2	0.5	0.2	0.05(0.145)	0.15	0.73
TaiESM1	-0.5	0.3	0.5	0.2	0.08(0.021)	0.00	0.59
UKESM1-0-LL	-0.6	0.4	0.3	0.0	0.09(<0.005)	0.06	0.55

S 16. Bias, standard deviation, trend, and MKGE score of SST for SC.

SC	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.3	-	0.12(<0.005)	-	-
ACCESS-CM2	-0.4	0.4	0.2	0.5	0.04(<0.005)	0.70	0.06
ACCESS-ESM1-5	0.1	0.1	0.3	0.0	0.09(<0.005)	0.20	0.77
AWI-CM-1-1-MR	0.1	0.1	0.3	0.1	0.07(<0.005)	0.39	0.60
CAMS-CSM1-0	-0.4	0.4	0.1	0.7	0.03(<0.005)	0.75	-0.09
CanESM5	-0.3	0.3	0.2	0.3	0.05(<0.005)	0.58	0.28
CESM2	0.1	0.1	0.4	0.2	0.15(<0.005)	0.31	0.61
CESM2-WACCM	-0.2	0.2	0.3	0.0	0.09(<0.005)	0.19	0.71
CMCC-CM2-SR5	1.1	1.0	0.5	0.8	0.19(<0.005)	0.67	-0.46
CNRM-CM6-1	0.7	0.6	0.5	0.9	0.17(<0.005)	0.49	-0.19
CNRM-CM6-1-HR	0.2	0.1	0.3	0.1	0.01(0.805)	1.00	-0.01
CNRM-ESM2-1	1.0	0.9	0.5	0.8	0.17(<0.005)	0.44	-0.30
EC-Earth3	0.3	0.3	0.5	0.9	0.20(<0.005)	0.72	-0.17
GISS-E2-1-G	-0.6	0.6	0.1	0.9	0.02(<0.036)	0.91	-0.39
IPSL-CM6A-LR	0.6	0.5	0.6	1.0	0.15(<0.005)	0.25	-0.15
MIROC6	-0.3	0.3	0.3	0.2	0.05(<0.013)	0.63	0.27
MIROC-ES2L	-0.5	0.5	0.2	0.5	0.03(<0.019)	0.77	-0.05
MPI-ESM1-2-HR	-0.4	0.4	0.4	0.2	0.14(<0.005)	0.23	0.53
MPI-ESM1-2-LR	-0.7	0.6	0.2	0.3	0.06(<0.005)	0.53	0.14
MRI-ESM2-0	0.3	0.3	0.4	0.5	0.13(<0.005)	0.14	0.40
NorESM2-LM	0.0	0.0	0.4	0.5	0.13(<0.005)	0.08	0.49
TaiESM1	-0.4	0.3	0.3	0.0	0.11(<0.005)	0.00	0.68
UKESM1-0-LL	-0.4	0.4	0.2	0.6	0.05(<0.005)	0.55	0.12

S 17. Bias, standard deviation, trend, and MKGE score of SST for BS.

BS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.4	-	0.06(0.036)	-	-
ACCESS-CM2	-1.3	0.4	0.5	0.1	0.01(0.691)	0.13	0.58
ACCESS-ESM1-5	-0.9	0.3	0.9	0.7	0.34(<0.005)	0.79	-0.08
AWI-CM-1-1-MR	-1.4	0.4	0.7	0.4	0.19(<0.005)	0.38	0.32
CAMS-CSM1-0	-2.3	0.7	0.4	0.0	0.06(0.018)	0.00	0.28
CanESM5	-1.2	0.4	0.8	0.6	0.18(<0.005)	0.34	0.19
CESM2	-0.1	0.0	0.7	0.4	0.20(<0.005)	0.39	0.44
CESM2-WACCM	-0.5	0.1	0.7	0.4	0.12(0.014)	0.16	0.55
CMCC-CM2-SR5	0.7	0.2	1.1	1.0	0.41(<0.005)	1.00	-0.43
CNRM-CM6-1	0.8	0.2	0.8	0.6	0.25(<0.005)	0.53	0.19
CNRM-CM6-1-HR	0.7	0.2	0.6	0.3	0.05(0.294)	0.04	0.66
CNRM-ESM2-1	1.7	0.5	0.7	0.4	0.16(<0.005)	0.29	0.27
EC-Earth3	1.1	0.3	1.0	0.9	0.35(<0.005)	0.83	-0.26
GISS-E2-1-G	-3.2	1.0	0.4	0.0	0.01(0.777)	0.15	-0.01
IPSL-CM6A-LR	1.0	0.3	0.8	0.5	0.21(<0.005)	0.44	0.26
MIROC6	0.1	0.0	0.9	0.7	0.03(0.633)	0.08	0.31
MIROC-ES2L	-1.4	0.4	0.9	0.7	0.16(0.011)	0.27	0.16
MPI-ESM1-2-HR	-1.4	0.4	0.7	0.4	0.22(<0.005)	0.45	0.25
MPI-ESM1-2-LR	-2.2	0.7	0.6	0.3	0.16(<0.005)	0.28	0.22
MRI-ESM2-0	0.7	0.2	0.7	0.4	0.21(<0.005)	0.41	0.37
NorESM2-LM	-0.4	0.1	0.8	0.6	0.22(<0.005)	0.46	0.21
TaiESM1	-1.0	0.3	0.7	0.4	0.18(<0.005)	0.33	0.41
UKESM1-0-LL	-1.3	0.4	0.8	0.6	0.24(<0.005)	0.53	0.12

S 18. Bias, standard deviation, trend, and MKGE score of SST for AS.

AS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.5	-	0.03(0.470)	-	-
ACCESS-CM2	-1.0	0.2	0.6	0.2	0.05(0.335)	0.04	0.77
ACCESS-ESM1-5	-0.3	0.0	1.0	0.7	0.35(<0.005)	0.98	-0.22
AWI-CM-1-1-MR	-3.4	0.6	0.6	0.2	0.11(0.023)	0.23	0.35
CAMS-CSM1-0	-4.3	0.8	0.5	0.1	0.06(0.069)	0.09	0.24
CanESM5	-3.9	0.7	0.7	0.3	0.16(<0.005)	0.39	0.15
CESM2	-0.6	0.1	0.7	0.3	0.16(<0.005)	0.39	0.48
CESM2-WACCM	-0.8	0.1	0.7	0.3	0.05(0.376)	0.04	0.67
CMCC-CM2-SR5	-0.7	0.1	1.1	0.9	0.36(<0.005)	1.00	-0.38
CNRM-CM6-1	0.4	0.0	0.8	0.4	0.21(<0.005)	0.55	0.33
CNRM-CM6-1-HR	1.2	0.2	0.5	0.0	0.12(<0.005)	0.27	0.67
CNRM-ESM2-1	1.2	0.2	0.6	0.1	0.19(<0.005)	0.47	0.48
EC-Earth3	0.2	0.0	0.8	0.4	0.22(<0.005)	0.57	0.33
GISS-E2-1-G	-5.7	1.0	0.5	0.0	-0.09(0.018)	0.35	-0.06
IPSL-CM6A-LR	-0.7	0.1	0.6	0.2	0.14(<0.005)	0.33	0.61
MIROC6	0.2	0.0	1.0	0.7	0.13(0.064)	0.30	0.26
MIROC-ES2L	-0.2	0.0	1.2	1.0	0.19(0.021)	0.49	-0.12
MPI-ESM1-2-HR	-2.3	0.4	0.7	0.2	0.19(<0.005)	0.49	0.35
MPI-ESM1-2-LR	-4.1	0.7	1.0	0.7	0.27(<0.005)	0.74	-0.27
MRI-ESM2-0	0.7	0.1	0.7	0.3	0.26(<0.005)	0.70	0.24
NorESM2-LM	-1.6	0.3	0.8	0.5	0.19(<0.005)	0.50	0.24
TaiESM1	-1.4	0.2	0.7	0.2	0.02(0.647)	0.00	0.71
UKESM1-0-LL	-1.5	0.2	0.9	0.5	0.17(<0.005)	0.44	0.28

S 19. Bias, standard deviation, trend, and MKGE score of SST for BCS.

BCS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
HadISST	-	-	0.5	-	0.02(0.555)	-	-
ACCESS-CM2	0.0	0.0	0.5	0.0	0.09(0.008)	0.14	0.86
ACCESS-ESM1-5	0.6	0.1	0.8	0.7	0.30(<0.005)	1.00	-0.20
AWI-CM-1-1-MR	-1.3	0.3	0.6	0.3	0.06(0.172)	0.00	0.61
CAMS-CSM1-0	-2.9	0.7	0.6	0.3	0.09(<0.031)	0.15	0.25
CanESM5	-1.3	0.3	0.7	0.4	0.16(<0.005)	0.43	0.31
CESM2	0.5	0.1	0.7	0.4	0.14(<0.005)	0.33	0.43
CESM2-WACCM	0.2	0.1	0.6	0.3	0.09(0.034)	0.14	0.66
CMCC-CM2-SR5	0.9	0.2	0.9	1.0	0.23(<0.005)	0.71	-0.24
CNRM-CM6-1	2.0	0.5	0.7	0.5	0.20(<0.005)	0.61	0.10
CNRM-CM6-1-HR	2.3	0.5	0.5	0.0	0.14(<0.005)	0.33	0.38
CNRM-ESM2-1	2.5	0.6	0.6	0.2	0.13(<0.005)	0.29	0.32
EC-Earth3	1.2	0.3	0.7	0.4	0.18(<0.005)	0.51	0.29
GISS-E2-1-G	-4.3	1.0	0.5	0.0	-0.11(<0.005)	0.36	-0.06
IPSL-CM6A-LR	0.3	0.1	0.5	0.1	0.12(<0.005)	0.27	0.69
MIROC6	1.8	0.4	0.7	0.5	0.12(0.013)	0.28	0.30
MIROC-ES2L	1.4	0.3	0.7	0.6	0.09(0.082)	0.15	0.30
MPI-ESM1-2-HR	-1.1	0.3	0.6	0.2	0.15(<0.005)	0.39	0.49
MPI-ESM1-2-LR	-2.3	0.5	0.9	0.9	0.20(<0.005)	0.58	-0.19
MRI-ESM2-0	1.3	0.3	0.5	0.1	0.16(<0.005)	0.44	0.45
NorESM2-LM	-0.7	0.2	0.7	0.6	0.09(0.076)	0.15	0.41
TaiESM1	-0.1	0.0	0.6	0.2	-0.04(0.336)	0.09	0.74
UKESM1-0-LL	0.3	0.1	0.8	0.7	0.21(<0.005)	0.62	0.10

S 20. Bias, standard deviation, trend, and MKGE score of BT for GoM.

GoM	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.6	-	0.39(0.060)	-	-
ACCESS-CM2	3.3	0.5	0.4	0.5	-0.01(0.940)	0.50	0.14
ACCESS-ESM1-5	6.6	1.0	0.4	0.5	0.01(0.963)	0.40	-0.18
AWI-CM-1-1-MR	0.1	0.0	0.4	0.6	0.19(0.135)	0.20	0.41
CAMS-CSM1-0	2.0	0.3	0.4	0.6	0.04(0.747)	0.36	0.27
CanESM5	3.5	0.5	0.4	0.5	0.27(0.040)	0.12	0.26
CESM2	2.1	0.3	0.5	0.4	0.41(<0.005)	0.01	0.52
CESM2-WACCM	2.0	0.3	0.5	0.4	0.50(<0.005)	0.11	0.52
CMCC-CM2-SR5	3.6	0.5	0.7	0.2	0.74(<0.005)	0.37	0.32
CNRM-CM6-1	0.0	0.0	0.6	0.0	-0.61(<0.005)	0.93	0.07
CNRM-CM6-1-HR	-0.6	0.1	0.7	0.0	0.08(0.722)	0.32	0.66
CNRM-ESM2-1	0.9	0.1	0.6	0.0	0.41(0.043)	0.01	0.86
EC-Earth3	3.0	0.5	0.6	0.2	0.45(0.009)	0.06	0.51
GISS-E2-1-G	3.6	0.6	0.7	0.0	0.24(0.275)	0.15	0.43
IPSL-CM6A-LR	3.7	0.6	0.6	0.1	0.56(<0.005)	0.17	0.40
MIROC6	4.0	0.6	0.4	0.6	0.38(<0.005)	0.00	0.16
MIROC-ES2L	2.4	0.4	0.5	0.4	0.23(0.125)	0.16	0.43
MPI-ESM1-2-HR	2.4	0.4	0.6	0.0	0.02(0.924)	0.38	0.47
MPI-ESM1-2-LR	2.4	0.4	0.7	0.2	-0.72(<0.005)	1.00	-0.08
MRI-ESM2-0	-1.1	0.2	0.4	0.5	0.28(0.034)	0.10	0.47
NorESM2-LM	-1.4	0.2	0.2	1.0	0.24(<0.005)	0.15	-0.03
TaiESM1	2.0	0.3	0.4	0.5	0.45(<0.005)	0.06	0.45
UKESM1-0-LL	2.0	0.3	0.5	0.4	-0.10(0.536)	0.56	0.26

S 21. Bias, standard deviation, trend, and MKGE score of BT for WSS.

WSS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	1.0	-	0.58(0.079)	-	-
ACCESS-CM2	4.1	0.5	1.2	0.1	0.14(0.730)	0.36	0.37
ACCESS-ESM1-5	7.9	1.0	0.4	0.7	0.10(0.420)	0.39	-0.30
AWI-CM-1-1-MR	3.1	0.4	0.6	0.5	0.08(0.679)	0.40	0.28
CAMS-CSM1-0	5.4	0.7	0.5	0.6	-0.01(0.928)	0.50	-0.03
CanESM5	7.2	0.9	0.6	0.5	0.17(0.358)	0.33	-0.09
CESM2	4.7	0.6	0.6	0.5	0.62(<0.005)	0.01	0.21
CESM2-WACCM	4.6	0.6	0.5	0.6	0.52(<0.005)	0.04	0.15
CMCC-CM2-SR5	7.4	0.9	0.9	0.1	0.56(0.064)	0.01	0.06
CNRM-CM6-1	1.5	0.2	1.1	0.0	-0.99(<0.005)	0.98	0.01
CNRM-CM6-1-HR	0.2	0.0	1.2	0.2	0.28(0.483)	0.24	0.71
CNRM-ESM2-1	2.7	0.3	1.4	0.4	0.33(0.465)	0.20	0.48
EC-Earth3	7.6	1.0	0.7	0.4	0.16(0.488)	0.34	-0.08
GISS-E2-1-G	4.2	0.5	0.7	0.3	0.22(0.356)	0.29	0.32
IPSL-CM6A-LR	6.7	0.8	1.0	0.0	0.20(0.539)	0.30	0.10
MIROC6	5.1	0.6	0.3	0.8	0.31(<0.005)	0.21	-0.04
MIROC-ES2L	4.0	0.5	0.3	0.8	0.22(0.024)	0.29	0.00
MPI-ESM1-2-HR	4.4	0.5	1.1	0.1	0.57(0.118)	0.00	0.45
MPI-ESM1-2-LR	1.5	0.2	1.0	0.0	-0.89(<0.005)	0.93	0.05
MRI-ESM2-0	0.8	0.1	0.4	0.8	0.41(<0.005)	0.13	0.23
NorESM2-LM	-0.8	0.1	0.2	1.0	0.18(<0.005)	0.32	-0.05
TaiESM1	4.3	0.5	0.6	0.5	0.69(<0.005)	0.08	0.25
UKESM1-0-LL	5.3	0.7	1.0	0.0	-1.03(<0.005)	1.00	-0.20

S 22. Bias, standard deviation, trend, and MKGE score of BT for CSS.

CSS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	1.0	-	0.53(0.081)	-	0.00
ACCESS-CM2	2.5	0.3	1.0	0.0	-0.04(0.903)	0.50	0.41
ACCESS-ESM1-5	6.1	0.8	0.3	0.8	0.15(0.166)	0.30	-0.16
AWI-CM-1-1-MR	2.5	0.3	0.8	0.2	-0.09(0.746)	0.52	0.37
CAMS-CSM1-0	2.2	0.3	0.6	0.5	0.04(0.839)	0.39	0.33
CanESM5	6.6	0.9	0.5	0.5	0.18(0.331)	0.28	-0.06
CESM2	3.2	0.4	0.6	0.4	0.66(<0.005)	0.10	0.43
CESM2-WACCM	3.2	0.4	0.5	0.5	0.57(<0.005)	0.02	0.36
CMCC-CM2-SR5	5.4	0.7	1.1	0.2	0.52(0.161)	0.00	0.25
CNRM-CM6-1	-0.8	0.1	1.1	0.1	-0.70(0.034)	0.76	0.23
CNRM-CM6-1-HR	-0.6	0.0	1.2	0.3	0.13(0.738)	0.32	0.56
CNRM-ESM2-1	0.3	0.0	1.3	0.4	0.17(0.677)	0.28	0.53
EC-Earth3	7.4	1.0	0.7	0.3	0.03(0.894)	0.40	-0.11
GISS-E2-1-G	3.0	0.4	0.8	0.2	0.15(0.570)	0.30	0.49
IPSL-CM6A-LR	5.0	0.7	1.1	0.1	-0.14(0.694)	0.54	0.13
MIROC6	3.9	0.5	0.4	0.7	0.34(<0.005)	0.15	0.09
MIROC-ES2L	2.9	0.4	0.3	0.8	0.28(0.007)	0.19	0.14
MPI-ESM1-2-HR	4.0	0.5	1.0	0.1	0.58(0.084)	0.03	0.46
MPI-ESM1-2-LR	1.6	0.2	0.6	0.5	-0.43(0.014)	0.66	0.18
MRI-ESM2-0	0.3	0.0	0.5	0.6	0.50(<0.005)	0.02	0.45
NorESM2-LM	-2.2	0.3	0.1	1.0	0.18(<0.005)	0.28	-0.07
TaiESM1	2.9	0.4	0.7	0.3	0.72(<0.005)	0.14	0.48
UKESM1-0-LL	3.4	0.4	1.3	0.4	-1.30(<0.005)	1.00	-0.16

S 23. Bias, standard deviation, trend, and MKGE score of BT for ESS.

ESS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.7	-	0.81(<0.005)	-	-
ACCESS-CM2	4.3	0.5	1.1	0.5	0.57(0.104)	0.07	0.27
ACCESS-ESM1-5	8.0	1.0	0.8	0.1	0.35(0.196)	0.24	0.01
AWI-CM-1-1-MR	2.9	0.3	0.7	0.1	0.36(0.087)	0.23	0.57
CAMS-CSM1-0	4.9	0.6	0.7	0.0	-0.06(0.798)	0.50	0.23
CanESM5	6.3	0.8	0.7	0.1	0.14(0.555)	0.40	0.14
CESM2	5.8	0.7	0.8	0.1	0.99(<0.005)	0.03	0.29
CESM2-WACCM	5.8	0.7	0.6	0.2	0.67(<0.005)	0.00	0.29
CMCC-CM2-SR5	4.3	0.5	0.8	0.1	0.14(0.604)	0.40	0.34
CNRM-CM6-1	0.4	0.0	0.8	0.0	-0.24(0.341)	0.64	0.36
CNRM-CM6-1-HR	0.4	0.0	0.8	0.0	0.32(0.191)	0.26	0.73
CNRM-ESM2-1	1.0	0.1	1.0	0.4	0.27(0.435)	0.30	0.45
EC-Earth3	8.3	1.0	1.0	0.4	-0.28(0.403)	0.66	-0.26
GISS-E2-1-G	4.9	0.6	0.7	0.0	0.16(0.506)	0.38	0.29
IPSL-CM6A-LR	5.6	0.7	1.4	1.0	-0.45(0.340)	0.79	-0.44
MIROC6	6.7	0.8	0.5	0.4	0.38(0.008)	0.21	0.08
MIROC-ES2L	5.7	0.7	0.3	0.6	0.27(<0.005)	0.30	0.03
MPI-ESM1-2-HR	6.1	0.7	1.2	0.7	0.48(0.222)	0.14	0.00
MPI-ESM1-2-LR	0.0	0.0	0.5	0.3	-0.12(0.504)	0.54	0.38
MRI-ESM2-0	3.2	0.4	0.6	0.3	0.48(<0.005)	0.14	0.51
NorESM2-LM	0.9	0.1	0.1	0.9	0.18(<0.005)	0.37	0.04
TaiESM1	4.7	0.6	1.1	0.5	1.45(<0.005)	0.37	0.18
UKESM1-0-LL	2.8	0.3	1.2	0.7	-0.73(0.053)	1.00	-0.24

S 24. Bias, standard deviation, trend, and MKGE score of BT for GSL.

GSL	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.3	-	0.38(<0.005)	-	-
ACCESS-CM2	1.8	0.2	0.7	0.8	0.47(0.031)	0.05	0.20
ACCESS-ESM1-5	4.4	0.6	0.6	0.6	0.48(0.013)	0.05	0.14
AWI-CM-1-1-MR	1.5	0.2	0.3	0.0	0.44(<0.005)	0.03	0.80
CAMS-CSM1-0	2.4	0.3	0.5	0.3	0.04(0.784)	0.19	0.51
CanESM5	4.7	0.7	0.3	0.1	0.17(0.059)	0.11	0.33
CESM2	3.1	0.4	0.8	1.0	1.07(<0.005)	0.40	-0.16
CESM2-WACCM	3.1	0.4	0.7	0.9	0.91(<0.005)	0.31	0.00
CMCC-CM2-SR5	3.9	0.5	0.3	0.1	0.22(0.012)	0.08	0.46
CNRM-CM6-1	0.5	0.1	0.3	0.1	0.09(0.363)	0.16	0.81
CNRM-CM6-1-HR	0.0	0.0	0.3	0.0	0.33(<0.005)	0.02	0.97
CNRM-ESM2-1	0.8	0.1	0.3	0.0	0.24(0.010)	0.07	0.86
EC-Earth3	4.2	0.6	0.4	0.0	-0.15(0.216)	1.00	-0.16
GISS-E2-1-G	3.4	0.5	0.5	0.4	0.11(0.537)	0.15	0.36
IPSL-CM6A-LR	2.1	0.3	0.3	0.1	-0.12(0.180)	0.50	0.41
MIROC6	7.2	1.0	0.7	0.7	0.84(<0.005)	0.27	-0.27
MIROC-ES2L	6.3	0.9	0.4	0.1	0.32(0.006)	0.02	0.13
MPI-ESM1-2-HR	3.4	0.5	0.4	0.1	0.39(<0.005)	0.00	0.51
MPI-ESM1-2-LR	0.9	0.1	0.1	0.4	0.04(0.451)	0.19	0.56
MRI-ESM2-0	3.2	0.4	0.6	0.5	0.55(<0.005)	0.09	0.32
NorESM2-LM	1.5	0.2	0.6	0.5	0.72(<0.005)	0.20	0.44
TaiESM1	2.1	0.3	0.7	0.8	0.95(<0.005)	0.33	0.09
UKESM1-0-LL	1.9	0.3	0.3	0.1	0.25(<0.005)	0.07	0.70

S 25. Bias, standard deviation, trend, and MKGE score of BT for SNS.

SNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.5	-	0.51(0.001)	-	-
ACCESS-CM2	4.9	0.6	1.0	1.0	0.38(0.257)	0.08	-0.18
ACCESS-ESM1-5	7.5	1.0	1.0	1.0	0.60(0.064)	0.05	-0.39
AWI-CM-1-1-MR	1.3	0.1	0.5	0.1	0.49(<0.005)	0.00	0.87
CAMS-CSM1-0	3.7	0.5	0.8	0.6	0.26(0.338)	0.17	0.25
CanESM5	4.2	0.5	0.5	0.0	0.21(0.230)	0.20	0.44
CESM2	4.6	0.6	0.8	0.5	1.07(<0.005)	0.40	0.11
CESM2-WACCM	4.7	0.6	0.7	0.3	0.78(<0.005)	0.18	0.30
CMCC-CM2-SR5	4.1	0.5	0.5	0.1	0.11(0.475)	0.28	0.40
CNRM-CM6-1	1.4	0.1	0.6	0.2	0.13(0.530)	0.26	0.65
CNRM-CM6-1-HR	1.0	0.1	0.5	0.0	0.25(0.128)	0.18	0.80
CNRM-ESM2-1	1.6	0.2	0.6	0.1	0.24(0.197)	0.18	0.75
EC-Earth3	3.8	0.5	0.9	0.7	-0.79(<0.005)	1.00	-0.34
GISS-E2-1-G	4.4	0.6	0.6	0.1	-0.17(0.396)	0.50	0.24
IPSL-CM6A-LR	2.9	0.3	0.7	0.3	0.10(0.652)	0.28	0.46
MIROC6	5.8	0.8	0.5	0.1	0.11(0.474)	0.28	0.19
MIROC-ES2L	6.1	0.8	0.4	0.3	0.28(0.008)	0.15	0.11
MPI-ESM1-2-HR	1.9	0.2	0.4	0.2	0.03(0.796)	0.34	0.54
MPI-ESM1-2-LR	0.4	0.0	0.4	0.2	0.10(0.434)	0.29	0.62
MRI-ESM2-0	2.5	0.3	0.5	0.1	0.10(0.515)	0.29	0.58
NorESM2-LM	2.6	0.3	0.1	0.9	0.10(<0.005)	0.29	0.05
TaiESM1	3.2	0.4	0.6	0.1	0.77(<0.005)	0.18	0.55
UKESM1-0-LL	1.4	0.1	0.5	0.0	0.54(<0.005)	0.01	0.86

S 26. Bias, standard deviation, trend, and MKGE score of BT for CNS.

CNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.4	-	0.30(0.025)	-	-
ACCESS-CM2	2.2	0.5	1.3	1.0	0.68(0.093)	0.18	-0.13
ACCESS-ESM1-5	2.7	0.6	0.8	0.5	0.34(0.215)	0.00	0.20
AWI-CM-1-1-MR	3.3	0.8	0.5	0.1	0.48(<0.005)	0.08	0.22
CAMS-CSM1-0	1.1	0.2	1.0	0.7	0.40(0.215)	0.03	0.30
CanESM5	0.4	0.1	0.3	0.1	0.22(0.033)	0.02	0.85
CESM2	3.7	0.9	0.3	0.2	0.14(0.134)	0.07	0.11
CESM2-WACCM	3.8	0.9	0.3	0.2	0.22(0.017)	0.02	0.09
CMCC-CM2-SR5	0.9	0.2	0.4	0.1	0.00(0.992)	0.50	0.46
CNRM-CM6-1	0.1	0.0	0.7	0.3	0.71(<0.005)	0.19	0.63
CNRM-CM6-1-HR	0.5	0.1	0.3	0.1	0.27(0.010)	0.00	0.86
CNRM-ESM2-1	0.5	0.1	0.4	0.1	0.24(0.027)	0.01	0.87
EC-Earth3	0.2	0.0	0.5	0.1	-0.18(0.291)	0.87	0.12
GISS-E2-1-G	2.7	0.6	0.5	0.1	-0.21(0.204)	0.95	-0.13
IPSL-CM6A-LR	0.3	0.0	0.6	0.2	0.43(0.019)	0.05	0.80
MIROC6	2.7	0.6	0.4	0.0	-0.24(0.085)	1.00	-0.18
MIROC-ES2L	3.9	0.9	0.4	0.0	0.50(<0.005)	0.09	0.09
MPI-ESM1-2-HR	1.3	0.3	0.5	0.1	-0.14(0.400)	0.80	0.15
MPI-ESM1-2-LR	1.2	0.3	0.5	0.1	0.49(<0.005)	0.08	0.71
MRI-ESM2-0	1.3	0.3	0.3	0.2	0.17(0.043)	0.05	0.66
NorESM2-LM	4.2	1.0	0.2	0.3	0.06(0.293)	0.10	-0.05
TaiESM1	3.3	0.8	0.3	0.2	0.06(0.469)	0.11	0.20
UKESM1-0-LL	0.5	0.1	0.9	0.6	1.10(<0.005)	0.40	0.27

S 27. Bias, standard deviation, trend, and MKGE score of BT for NNS.

NNS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.4	-	0.28(0.020)	-	-
ACCESS-CM2	0.5	0.1	0.7	0.2	0.22(0.334)	0.03	0.71
ACCESS-ESM1-5	0.3	0.1	0.7	0.2	0.59(<0.005)	0.17	0.68
AWI-CM-1-1-MR	2.5	0.8	0.4	0.0	0.40(<0.005)	0.06	0.24
CAMS-CSM1-0	1.6	0.5	1.6	1.0	0.85(0.097)	0.32	-0.16
CanESM5	0.3	0.1	0.5	0.1	0.48(<0.005)	0.11	0.84
CESM2	2.8	0.9	0.2	0.1	0.18(0.007)	0.05	0.11
CESM2-WACCM	2.8	0.9	0.3	0.1	0.27(<0.005)	0.00	0.13
CMCC-CM2-SR5	0.8	0.2	0.3	0.1	0.14(0.111)	0.07	0.74
CNRM-CM6-1	-0.5	0.1	0.6	0.2	0.69(<0.005)	0.22	0.67
CNRM-CM6-1-HR	-0.5	0.2	0.3	0.1	0.20(0.013)	0.04	0.81
CNRM-ESM2-1	0.0	0.0	0.3	0.0	0.35(<0.005)	0.03	0.95
EC-Earth3	0.2	0.1	0.4	0.0	0.08(0.542)	0.11	0.87
GISS-E2-1-G	0.7	0.2	0.4	0.0	-0.10(0.427)	0.50	0.45
IPSL-CM6A-LR	0.4	0.1	0.6	0.2	0.55(<0.005)	0.15	0.75
MIROC6	2.1	0.6	0.4	0.0	-0.28(0.039)	1.00	-0.19
MIROC-ES2L	3.2	1.0	0.4	0.0	0.37(<0.005)	0.04	0.00
MPI-ESM1-2-HR	1.0	0.3	0.4	0.0	-0.12(0.406)	0.55	0.37
MPI-ESM1-2-LR	1.0	0.3	0.4	0.0	0.42(<0.005)	0.07	0.69
MRI-ESM2-0	1.0	0.3	0.3	0.1	0.26(0.007)	0.01	0.69
NorESM2-LM	2.9	0.9	0.1	0.2	0.10(0.012)	0.10	0.08
TaiESM1	2.2	0.7	0.2	0.1	0.12(0.078)	0.08	0.30
UKESM1-0-LL	0.3	0.1	0.8	0.3	1.00(<0.005)	0.40	0.49

S 28. Bias, standard deviation, trend, and MKGE score of BT for SLS.

SLS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.5	-	0.35(0.013)	-	-
ACCESS-CM2	0.3	0.1	0.7	0.7	0.20(0.406)	0.10	0.28
ACCESS-ESM1-5	0.2	0.1	0.7	0.5	0.54(0.008)	0.13	0.46
AWI-CM-1-1-MR	2.8	0.7	0.4	0.2	0.38(<0.005)	0.02	0.29
CAMS-CSM1-0	0.5	0.1	0.8	0.9	0.71(<0.005)	0.24	0.06
CanESM5	0.2	0.1	0.4	0.1	0.40(<0.005)	0.03	0.90
CESM2	3.6	0.9	0.2	0.7	0.11(0.048)	0.16	-0.14
CESM2-WACCM	3.6	0.9	0.2	0.5	0.20(<0.005)	0.10	-0.03
CMCC-CM2-SR5	0.9	0.2	0.2	0.6	0.14(0.049)	0.14	0.38
CNRM-CM6-1	0.0	0.0	0.6	0.2	0.59(<0.005)	0.16	0.72
CNRM-CM6-1-HR	-0.4	0.1	0.3	0.5	0.22(0.006)	0.09	0.51
CNRM-ESM2-1	0.4	0.1	0.3	0.2	0.34(<0.005)	0.00	0.74
EC-Earth3	0.4	0.1	0.5	0.0	0.14(0.367)	0.14	0.84
GISS-E2-1-G	1.0	0.2	0.3	0.2	-0.09(0.406)	0.50	0.39
IPSL-CM6A-LR	0.7	0.2	0.6	0.3	0.59(<0.005)	0.16	0.64
MIROC6	3.2	0.8	0.4	0.1	-0.22(0.078)	1.00	-0.28
MIROC-ES2L	4.1	1.0	0.3	0.4	0.35(<0.005)	0.00	-0.07
MPI-ESM1-2-HR	1.2	0.3	0.4	0.0	-0.18(0.193)	0.85	0.10
MPI-ESM1-2-LR	0.8	0.2	0.4	0.1	0.45(<0.005)	0.06	0.77
MRI-ESM2-0	1.3	0.3	0.3	0.3	0.29(<0.005)	0.04	0.55
NorESM2-LM	3.5	0.9	0.1	1.0	0.05(0.045)	0.20	-0.33
TaiESM1	3.0	0.7	0.2	0.7	0.08(0.147)	0.18	-0.04
UKESM1-0-LL	0.4	0.1	0.8	0.8	0.95(<0.005)	0.40	0.07

S 29. Bias, standard deviation, trend, and MKGE score of BT for NLS.

NLS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.5	-	0.36(0.011)	-	-
ACCESS-CM2	1.0	0.2	0.8	0.9	0.32(0.222)	0.03	0.08
ACCESS-ESM1-5	0.8	0.2	0.6	0.3	0.54(<0.005)	0.14	0.60
AWI-CM-1-1-MR	4.7	1.0	0.4	0.0	0.31(0.023)	0.04	0.00
CAMS-CSM1-0	0.9	0.2	0.3	0.3	0.29(<0.005)	0.06	0.62
CanESM5	0.1	0.0	0.4	0.2	0.33(<0.005)	0.03	0.79
CESM2	4.1	0.9	0.2	0.6	0.18(0.008)	0.14	-0.08
CESM2-WACCM	4.0	0.9	0.3	0.5	0.26(<0.005)	0.08	0.00
CMCC-CM2-SR5	0.4	0.1	0.3	0.4	0.19(0.046)	0.14	0.55
CNRM-CM6-1	0.4	0.1	0.7	0.5	0.68(<0.005)	0.26	0.41
CNRM-CM6-1-HR	-0.2	0.0	0.2	0.6	0.20(0.009)	0.13	0.43
CNRM-ESM2-1	0.8	0.2	0.4	0.2	0.40(<0.005)	0.03	0.71
EC-Earth3	0.2	0.0	0.6	0.4	0.26(0.191)	0.08	0.62
GISS-E2-1-G	0.5	0.1	0.3	0.3	-0.10(0.347)	0.50	0.40
IPSL-CM6A-LR	0.8	0.2	0.5	0.1	0.55(<0.005)	0.15	0.75
MIROC6	3.5	0.8	0.4	0.2	-0.18(0.121)	0.69	-0.05
MIROC-ES2L	4.4	0.9	0.3	0.5	0.29(<0.005)	0.06	-0.07
MPI-ESM1-2-HR	1.6	0.3	0.5	0.2	-0.31(0.075)	1.00	-0.07
MPI-ESM1-2-LR	0.9	0.2	0.4	0.1	0.49(<0.005)	0.10	0.78
MRI-ESM2-0	2.0	0.4	0.4	0.1	0.37(<0.005)	0.00	0.55
NorESM2-LM	4.6	1.0	0.1	1.0	0.06(0.038)	0.24	-0.42
TaiESM1	3.4	0.7	0.2	0.7	0.17(<0.005)	0.16	-0.03
UKESM1-0-LL	0.0	0.0	0.7	0.7	0.86(<0.005)	0.40	0.19

S 30. Bias, standard deviation, trend, and MKGE score of BT for HB.

HB	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.2	-	-0.11(0.024)	-	-
ACCESS-CM2	0.8	0.5	0.1	0.1	0.11(0.017)	0.63	0.21
ACCESS-ESM1-5	1.1	0.7	0.2	0.1	0.06(0.278)	0.56	0.13
AWI-CM-1-1-MR	1.6	1.0	0.2	0.2	0.11(0.065)	0.63	-0.20
CAMS-CSM1-0	-0.4	0.2	0.1	0.8	0.02(0.277)	0.50	0.05
CanESM5	0.7	0.4	0.2	0.3	0.24(<0.005)	0.83	0.01
CESM2	0.1	0.0	0.1	0.2	0.14(<0.005)	0.68	0.28
CESM2-WACCM	0.1	0.0	0.1	0.3	0.15(<0.005)	0.70	0.25
CMCC-CM2-SR5	0.4	0.2	0.2	0.3	0.14(0.022)	0.68	0.22
CNRM-CM6-1	0.4	0.2	0.2	0.1	0.20(<0.005)	0.77	0.20
CNRM-CM6-1-HR	-0.1	0.0	0.1	0.0	0.13(<0.005)	0.66	0.33
CNRM-ESM2-1	0.2	0.1	0.2	0.0	0.20(<0.005)	0.77	0.23
EC-Earth3	1.3	0.8	0.2	0.7	0.30(<0.005)	0.92	-0.42
GISS-E2-1-G	-0.5	0.3	0.2	0.3	0.06(0.356)	0.56	0.31
IPSL-CM6A-LR	1.3	0.8	0.2	0.0	0.18(<0.005)	0.74	-0.11
MIROC6	0.6	0.3	0.1	0.1	0.15(<0.005)	0.70	0.23
MIROC-ES2L	0.2	0.1	0.1	0.3	0.00(0.922)	0.00	0.64
MPI-ESM1-2-HR	1.1	0.7	0.1	0.0	0.12(0.011)	0.64	0.07
MPI-ESM1-2-LR	0.9	0.5	0.1	0.3	0.10(0.006)	0.61	0.13
MRI-ESM2-0	0.5	0.3	0.1	0.5	0.04(0.219)	0.53	0.22
NorESM2-LM	0.6	0.3	0.3	1.0	0.24(<0.005)	0.84	-0.34
TaiESM1	-0.2	0.1	0.1	0.7	0.08(<0.005)	0.58	0.12
UKESM1-0-LL	0.4	0.2	0.3	0.8	0.35(<0.005)	1.00	-0.29

S 31. Bias, standard deviation, trend, and MKGE score of BT for BB.

BB	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.1	-	0.04(0.108)	-	-
ACCESS-CM2	-1.2	0.2	0.2	1.0	0.23(<0.005)	0.32	-0.07
ACCESS-ESM1-5	2.6	0.5	0.1	0.0	0.03(0.285)	0.01	0.51
AWI-CM-1-1-MR	3.5	0.7	0.1	0.2	0.11(<0.005)	0.10	0.31
CAMS-CSM1-0	3.6	0.7	0.1	0.3	0.13(<0.005)	0.14	0.24
CanESM5	-0.7	0.1	0.0	0.3	0.05(<0.005)	0.00	0.69
CESM2	3.7	0.7	0.0	0.4	0.01(0.325)	0.04	0.19
CESM2-WACCM	3.4	0.6	0.1	0.2	0.14(<0.005)	0.16	0.32
CMCC-CM2-SR5	-0.5	0.1	0.1	0.0	0.07(<0.005)	0.04	0.91
CNRM-CM6-1	-1.0	0.2	0.2	0.8	0.24(<0.005)	0.33	0.13
CNRM-CM6-1-HR	-0.4	0.1	0.1	0.1	0.08(<0.005)	0.06	0.87
CNRM-ESM2-1	-1.1	0.2	0.1	0.2	0.12(<0.005)	0.13	0.71
EC-Earth3	-0.8	0.1	0.1	0.3	0.16(<0.005)	0.20	0.60
GISS-E2-1-G	-1.8	0.3	0.1	0.1	0.07(<0.005)	0.05	0.64
IPSL-CM6A-LR	-0.4	0.1	0.0	0.4	-0.01(0.159)	0.50	0.33
MIROC6	4.4	0.9	0.0	0.3	0.03(<0.005)	0.00	0.10
MIROC-ES2L	5.2	1.0	0.0	0.2	0.01(0.538)	0.04	-0.02
MPI-ESM1-2-HR	0.9	0.2	0.1	0.5	0.19(<0.005)	0.25	0.45
MPI-ESM1-2-LR	2.5	0.5	0.1	0.0	0.05(0.015)	0.01	0.52
MRI-ESM2-0	2.5	0.5	0.1	0.6	0.17(<0.005)	0.22	0.24
NorESM2-LM	-0.1	0.0	0.1	0.0	0.01(0.717)	0.04	0.96
TaiESM1	2.4	0.5	0.2	1.0	0.28(<0.005)	0.40	-0.13
UKESM1-0-LL	-1.1	0.2	0.1	0.0	0.09(<0.005)	0.08	0.78

S 32. Bias, standard deviation, trend, and MKGE score of BT for CAA.

CAA	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.1	-	0.11(<0.005)	-	-
ACCESS-CM2	-1.0	1.0	0.0	0.9	-0.02(<0.005)	0.68	-0.49
ACCESS-ESM1-5	-0.5	0.5	0.1	0.4	-0.01(0.864)	0.50	0.19
AWI-CM-1-1-MR	-0.3	0.2	0.0	0.7	0.03(<0.005)	0.28	0.18
CAMS-CSM1-0	-0.1	0.1	0.0	0.9	0.00(0.582)	0.40	0.00
CanESM5	-1.0	0.9	0.0	0.8	0.02(<0.005)	0.30	-0.30
CESM2	0.2	0.1	0.0	1.0	-0.01(<0.005)	0.60	-0.16
CESM2-WACCM	0.0	0.0	0.0	0.8	0.02(0.089)	0.34	0.15
CMCC-CM2-SR5	-0.5	0.5	0.0	0.7	0.04(<0.005)	0.24	0.15
CNRM-CM6-1	-0.4	0.4	0.0	0.5	0.06(<0.005)	0.11	0.31
CNRM-CM6-1-HR	-0.5	0.5	0.0	0.8	0.01(0.215)	0.36	0.03
CNRM-ESM2-1	-0.6	0.5	0.0	1.0	0.01(0.054)	0.38	-0.19
EC-Earth3	0.2	0.1	0.1	0.7	0.18(<0.005)	0.16	0.24
GISS-E2-1-G	-0.5	0.4	0.1	0.4	0.07(<0.005)	0.05	0.41
IPSL-CM6A-LR	0.4	0.4	0.0	0.7	0.05(<0.005)	0.18	0.21
MIROC6	0.1	0.1	0.0	0.6	-0.04(<0.005)	1.00	-0.15
MIROC-ES2L	-0.3	0.2	0.1	0.0	0.08(<0.005)	0.00	0.79
MPI-ESM1-2-HR	-0.1	0.1	0.1	0.0	0.08(<0.005)	0.02	0.91
MPI-ESM1-2-LR	0.2	0.2	0.0	0.5	0.05(<0.005)	0.16	0.41
MRI-ESM2-0	0.4	0.3	0.0	0.8	-0.01(0.140)	0.59	-0.04
NorESM2-LM	-0.1	0.0	0.0	0.9	0.02(<0.005)	0.31	0.01
TaiESM1	-0.2	0.1	0.0	0.5	0.06(<0.005)	0.12	0.45
UKESM1-0-LL	-0.6	0.6	0.1	0.4	0.06(<0.005)	0.10	0.27

S 33. Bias, standard deviation, trend, and MKGE score of BT for SBS.

SBS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.2	-	0.06(0.256)	-	-
ACCESS-CM2	-1.1	1.0	0.1	0.8	0.00(0.835)	0.17	-0.30
ACCESS-ESM1-5	-0.8	0.8	0.1	0.8	0.04(0.044)	0.05	-0.07
AWI-CM-1-1-MR	-0.2	0.2	0.1	0.4	0.00(0.911)	0.50	0.34
CAMS-CSM1-0	0.4	0.4	0.1	0.7	-0.01(0.517)	0.68	-0.07
CanESM5	-0.9	0.8	0.1	0.3	-0.01(0.776)	0.64	-0.05
CESM2	0.3	0.3	0.1	0.5	-0.01(0.674)	0.67	0.12
CESM2-WACCM	0.2	0.2	0.1	0.6	0.03(0.239)	0.08	0.37
CMCC-CM2-SR5	-0.1	0.1	0.2	0.0	0.11(0.036)	0.15	0.82
CNRM-CM6-1	-0.4	0.3	0.1	0.1	0.07(0.137)	0.00	0.63
CNRM-CM6-1-HR	-0.5	0.5	0.1	0.1	-0.03(0.515)	1.00	-0.10
CNRM-ESM2-1	-0.5	0.5	0.1	0.6	0.05(0.042)	0.00	0.23
EC-Earth3	0.1	0.1	0.2	0.0	0.09(0.110)	0.07	0.90
GISS-E2-1-G	-0.1	0.1	0.1	0.8	0.05(<0.005)	0.03	0.15
IPSL-CM6A-LR	0.3	0.3	0.1	0.5	0.05(0.100)	0.01	0.43
MIROC6	0.5	0.5	0.0	0.9	-0.01(0.535)	0.58	-0.17
MIROC-ES2L	-0.3	0.3	0.1	0.8	-0.03(0.125)	0.89	-0.25
MPI-ESM1-2-HR	-0.1	0.1	0.2	0.6	0.18(0.015)	0.40	0.30
MPI-ESM1-2-LR	0.3	0.3	0.1	0.3	0.11(<0.005)	0.15	0.57
MRI-ESM2-0	0.9	0.8	0.0	0.9	0.01(0.431)	0.15	-0.22
NorESM2-LM	-0.1	0.1	0.0	1.0	0.03(<0.005)	0.09	-0.01
TaiESM1	0.0	0.0	0.1	0.5	0.10(<0.005)	0.13	0.50
UKESM1-0-LL	-0.6	0.6	0.1	0.4	0.07(0.057)	0.00	0.32

S 34. Bias, standard deviation, trend, and MKGE score of BT for SC.

SC	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.1	-	0.05(0.265)	-	-
ACCESS-CM2	-0.8	1.0	0.1	0.2	0.04(0.168)	0.01	-0.01
ACCESS-ESM1-5	-0.3	0.4	0.2	0.2	0.19(<0.005)	0.17	0.57
AWI-CM-1-1-MR	-0.1	0.2	0.1	0.1	-0.01(0.692)	0.63	0.34
CAMS-CSM1-0	-0.4	0.5	0.1	0.2	0.03(0.267)	0.02	0.52
CanESM5	-0.6	0.7	0.2	0.2	-0.03(0.657)	1.00	-0.25
CESM2	0.1	0.1	0.2	0.2	0.01(0.891)	0.05	0.79
CESM2-WACCM	-0.2	0.3	0.2	0.1	0.06(0.240)	0.01	0.71
CMCC-CM2-SR5	0.6	0.8	0.4	1.0	0.37(<0.005)	0.40	-0.32
CNRM-CM6-1	-0.1	0.1	0.3	0.5	-0.01(0.934)	0.50	0.27
CNRM-CM6-1-HR	-0.3	0.3	0.2	0.2	0.02(0.756)	0.03	0.62
CNRM-ESM2-1	-0.3	0.3	0.2	0.1	0.10(0.053)	0.06	0.64
EC-Earth3	0.3	0.4	0.3	0.7	0.30(<0.005)	0.31	0.13
GISS-E2-1-G	-0.6	0.8	0.1	0.3	0.02(0.198)	0.03	0.15
IPSL-CM6A-LR	0.0	0.0	0.2	0.1	0.02(0.681)	0.03	0.90
MIROC6	-0.5	0.6	0.1	0.2	0.05(0.043)	0.00	0.42
MIROC-ES2L	-0.7	0.9	0.0	0.3	0.01(0.508)	0.04	0.03
MPI-ESM1-2-HR	-0.2	0.3	0.2	0.4	0.22(<0.005)	0.21	0.44
MPI-ESM1-2-LR	-0.4	0.5	0.1	0.0	0.08(0.051)	0.03	0.54
MRI-ESM2-0	0.3	0.4	0.2	0.4	0.20(<0.005)	0.18	0.45
NorESM2-LM	-0.3	0.3	0.2	0.3	0.23(<0.005)	0.22	0.51
TaiESM1	-0.4	0.4	0.2	0.1	0.19(<0.005)	0.17	0.52
UKESM1-0-LL	-0.6	0.7	0.1	0.1	0.04(0.263)	0.01	0.31

S 35. Bias, standard deviation, trend, and MKGE score of BT for BS.

BS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.4	-	-0.21(0.108)	-	-
ACCESS-CM2	-0.1	0.1	0.4	0.0	-0.18(0.130)	0.00	0.93
ACCESS-ESM1-5	0.9	0.4	0.7	0.5	0.53(0.010)	0.79	-0.06
AWI-CM-1-1-MR	-0.3	0.1	0.4	0.0	-0.10(0.408)	0.36	0.61
CAMS-CSM1-0	-1.4	0.7	0.2	0.4	-0.10(0.137)	0.40	0.11
CanESM5	-0.1	0.1	0.7	0.6	0.43(0.055)	0.73	0.03
CESM2	0.7	0.3	0.5	0.2	0.20(0.234)	0.59	0.29
CESM2-WACCM	0.3	0.2	0.5	0.1	0.29(0.062)	0.64	0.32
CMCC-CM2-SR5	2.1	1.0	0.9	1.0	0.84(<0.005)	0.98	-0.72
CNRM-CM6-1	-0.1	0.0	0.7	0.7	0.07(0.787)	0.51	0.16
CNRM-CM6-1-HR	0.3	0.1	0.4	0.1	0.05(0.755)	0.50	0.47
CNRM-ESM2-1	0.1	0.0	0.6	0.3	0.33(0.069)	0.67	0.25
EC-Earth3	2.0	1.0	0.8	0.8	0.85(<0.005)	0.98	-0.59
GISS-E2-1-G	-2.0	0.9	0.3	0.2	0.07(0.405)	0.52	-0.11
IPSL-CM6A-LR	0.3	0.1	0.7	0.5	0.64(<0.005)	0.85	-0.02
MIROC6	0.5	0.3	0.7	0.6	0.11(0.641)	0.54	0.13
MIROC-ES2L	-0.3	0.2	0.6	0.4	0.40(0.033)	0.71	0.17
MPI-ESM1-2-HR	0.2	0.1	0.5	0.2	0.57(<0.005)	0.81	0.17
MPI-ESM1-2-LR	0.0	0.0	0.4	0.1	0.31(<0.005)	0.66	0.34
MRI-ESM2-0	1.0	0.5	0.7	0.5	0.46(0.025)	0.75	-0.03
NorESM2-LM	0.1	0.0	0.5	0.2	0.50(<0.005)	0.77	0.21
TaiESM1	-0.4	0.2	0.6	0.5	0.72(<0.005)	0.91	-0.05
UKESM1-0-LL	0.1	0.1	0.7	0.7	0.88(<0.005)	1.00	-0.21

S 36. Bias, standard deviation, trend, and MKGE score of BT for EBS.

EBS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.6	-	-0.27(0.132)	-	-
ACCESS-CM2	0.4	0.2	0.5	0.2	-0.27(0.069)	0.00	0.71
ACCESS-ESM1-5	1.4	0.6	0.9	0.8	0.63(0.017)	0.78	-0.27
AWI-CM-1-1-MR	-0.5	0.2	0.5	0.2	-0.13(0.399)	0.40	0.50
CAMS-CSM1-0	-1.3	0.6	0.3	0.8	-0.14(0.087)	0.38	-0.05
CanESM5	0.0	0.0	0.8	0.8	0.54(0.042)	0.73	-0.06
CESM2	0.9	0.4	0.6	0.2	0.22(0.277)	0.58	0.29
CESM2-WACCM	0.6	0.2	0.5	0.1	0.19(0.261)	0.57	0.37
CMCC-CM2-SR5	2.3	1.0	0.9	1.0	0.81(<0.005)	0.86	-0.63
CNRM-CM6-1	0.0	0.0	0.9	0.9	0.12(0.694)	0.53	-0.04
CNRM-CM6-1-HR	0.4	0.2	0.6	0.0	0.05(0.805)	0.50	0.47
CNRM-ESM2-1	0.2	0.1	0.7	0.4	0.35(0.125)	0.64	0.22
EC-Earth3	2.3	1.0	0.8	0.6	0.79(<0.005)	0.85	-0.43
GISS-E2-1-G	-2.0	0.9	0.3	0.6	0.08(0.414)	0.52	-0.19
IPSL-CM6A-LR	0.3	0.1	0.8	0.7	0.72(<0.005)	0.82	-0.09
MIROC6	1.1	0.5	0.8	0.5	0.09(0.715)	0.52	0.11
MIROC-ES2L	0.3	0.1	0.8	0.7	0.49(0.058)	0.71	0.02
MPI-ESM1-2-HR	0.2	0.1	0.6	0.1	0.61(<0.005)	0.77	0.22
MPI-ESM1-2-LR	0.3	0.1	0.4	0.3	0.36(<0.005)	0.65	0.26
MRI-ESM2-0	1.2	0.5	0.8	0.6	0.54(0.032)	0.73	-0.10
NorESM2-LM	0.4	0.2	0.5	0.1	0.42(0.008)	0.67	0.30
TaiESM1	-0.2	0.1	0.7	0.5	0.81(<0.005)	0.86	0.00
UKESM1-0-LL	0.4	0.1	0.9	1.0	1.11(<0.005)	1.00	-0.41

S 37. Bias, standard deviation, trend, and MKGE score of BT for AS.

AS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.3	-	-0.15(0.170)	-	-
ACCESS-CM2	0.1	0.0	0.5	0.4	-0.11(0.532)	0.02	0.62
ACCESS-ESM1-5	1.2	0.3	0.9	1.0	0.81(<0.005)	1.00	-0.45
AWI-CM-1-1-MR	-1.4	0.4	0.4	0.1	0.00(0.979)	0.50	0.38
CAMS-CSM1-0	-1.9	0.5	0.3	0.1	0.20(0.028)	0.62	0.20
CanESM5	-2.5	0.7	0.7	0.8	0.54(0.021)	0.83	-0.31
CESM2	-0.2	0.0	0.3	0.0	-0.04(0.729)	0.30	0.70
CESM2-WACCM	-0.3	0.1	0.4	0.1	0.03(0.839)	0.51	0.48
CMCC-CM2-SR5	0.3	0.1	0.8	0.9	0.59(0.017)	0.86	-0.24
CNRM-CM6-1	-0.4	0.1	0.7	0.7	0.55(0.010)	0.84	-0.08
CNRM-CM6-1-HR	1.2	0.3	0.4	0.1	-0.12(0.348)	0.00	0.68
CNRM-ESM2-1	0.1	0.0	0.5	0.2	0.10(0.525)	0.56	0.39
EC-Earth3	1.2	0.3	0.5	0.3	0.31(0.039)	0.69	0.20
GISS-E2-1-G	-3.7	1.0	0.3	0.1	-0.01(0.895)	0.40	-0.08
IPSL-CM6A-LR	-0.1	0.0	0.6	0.4	0.53(<0.005)	0.83	0.06
MIROC6	0.1	0.0	0.1	0.5	0.04(0.075)	0.52	0.28
MIROC-ES2L	-0.1	0.0	0.2	0.3	0.10(0.032)	0.56	0.34
MPI-ESM1-2-HR	0.1	0.0	0.4	0.0	0.33(<0.005)	0.71	0.29
MPI-ESM1-2-LR	-0.8	0.2	0.4	0.1	0.21(0.087)	0.63	0.33
MRI-ESM2-0	0.6	0.1	0.5	0.2	0.30(0.047)	0.68	0.26
NorESM2-LM	-3.2	0.9	0.2	0.3	0.11(0.034)	0.57	-0.08
TaiESM1	-1.4	0.4	0.4	0.1	0.40(<0.005)	0.74	0.16
UKESM1-0-LL	0.3	0.1	0.6	0.5	0.68(<0.005)	0.92	-0.03

S 38. Bias, standard deviation, trend, and MKGE score of BT for BCS.

BCS	Bias	Norm_BS	STD	Norm_STD	Tr(p)	Norm_Tr	MKGE
GLORYS12	-	-	0.3	-	-0.16(0.052)	-	-
ACCESS-CM2	1.2	0.2	0.2	0.1	0.04(0.591)	0.55	0.39
ACCESS-ESM1-5	1.8	0.4	0.4	0.6	0.38(<0.005)	0.98	-0.22
AWI-CM-1-1-MR	-0.4	0.1	0.3	0.1	0.02(0.765)	0.53	0.46
CAMS-CSM1-0	-1.0	0.2	0.3	0.0	0.13(0.145)	0.66	0.31
CanESM5	0.8	0.2	0.4	0.4	0.36(<0.005)	0.95	-0.06
CESM2	0.4	0.1	0.3	0.0	-0.14(0.097)	0.00	0.92
CESM2-WACCM	0.3	0.0	0.4	0.4	0.07(0.554)	0.59	0.28
CMCC-CM2-SR5	1.0	0.2	0.4	0.5	0.33(<0.005)	0.92	-0.05
CNRM-CM6-1	2.1	0.5	0.3	0.0	0.24(<0.005)	0.80	0.08
CNRM-CM6-1-HR	2.9	0.6	0.2	0.2	-0.13(0.041)	0.02	0.33
CNRM-ESM2-1	2.2	0.5	0.2	0.5	0.05(0.306)	0.56	0.13
EC-Earth3	1.7	0.4	0.2	0.3	0.03(0.648)	0.54	0.30
GISS-E2-1-G	-2.7	0.6	0.4	0.4	-0.06(0.588)	0.40	0.21
IPSL-CM6A-LR	2.0	0.4	0.2	0.2	0.07(0.293)	0.59	0.24
MIROC6	-3.0	0.6	0.0	1.0	0.00(0.828)	0.50	-0.26
MIROC-ES2L	-3.0	0.6	0.0	1.0	0.01(0.162)	0.51	-0.29
MPI-ESM1-2-HR	1.4	0.3	0.2	0.5	0.02(0.633)	0.52	0.24
MPI-ESM1-2-LR	0.9	0.2	0.2	0.4	0.10(0.063)	0.63	0.25
MRI-ESM2-0	0.1	0.0	0.2	0.3	0.14(0.017)	0.68	0.26
NorESM2-LM	-4.7	1.0	0.1	0.7	0.09(<0.005)	0.61	-0.36
TaiESM1	-0.6	0.1	0.3	0.3	0.15(0.177)	0.69	0.24
UKESM1-0-LL	1.3	0.3	0.3	0.1	0.40(<0.005)	1.00	-0.04

S 39.The Bottom Temperature statistics against observation for top four models on the Atlantic Shelf for CSS, ESS, and WSS. Model Bias (Bs; unit: °C), standard deviation (STD), and trend (Tr; unit: °C/decade). The common time period is from 1970 to 2014.

	CSS			ESS			WSS		
	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend
Observation	-	0.9	0.15	-	0.8	-0.04	-	1.1	0.14
CNRM_ESM2_1	-0.1	1.1	0.02	1.3	1.1	-0.10	1.7	1.2	0.07
CNRM_CM6_1_HR	-1.7	1.4	0.56	0.1	0.9	0.31	-0.5	1.3	0.52
MRI_ESM2_0	-1.0	0.6	0.28	2.2	0.8	0.45	0.1	0.5	0.27
MPI_ESM1_2_LR	0.6	0.8	0.21	-0.4	0.7	0.21	0.3	1.1	0.26

S 40.The Bottom Temperature statistics against observation for top four models on the Atlantic Shelf for CNS and SNS at Spring and Fall. Model Bias (Bs; unit: °C), standard deviation (STD), and trend (Tr; unit: °C/decade) for the period of 1980-2014.

	CNS_Spr.			CNS_Fal.			SNS_Spr.			SNS_Fal.		
	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend
Observation	-	0.5	0.24	-	0.4	0.15	-	0.7	0.22	-	0.6	0.06
CNRM_ESM2_1	0.3	0.6	0.43	0.5	0.6	0.47	1.8	0.8	0.11	2.3	0.6	0.05
CNRM_CM6_1_HR	0.3	0.4	0.19	1.0	0.2	0.13	0.9	0.5	0.16	2.0	0.4	0.17
MRI_ESM2_0	1.6	0.3	0.07	1.4	0.3	0.06	2.6	0.6	0.43	2.6	0.6	0.36
MPI_ESM1_2_LR	1.4	0.5	0.30	1.4	0.5	0.21	0.8	0.4	0.21	0.6	0.5	0.17

S 41.The Bottom Temperature statistics against observation for top four models on the Atlantic Shelf for NNS and SLS in Fall. Model Bias (Bs; unit: °C), standard deviation (STD), and trend (Tr; unit: °C/decade) for the period of 1980-2014.

	NNS_Fal.				SLS_Fal.	
	Bias	STD	Trend	Bias	STD	Trend
Observation	-	0.5	0.31	-	0.6	0.40
CNRM_ESM2_1	-0.1	0.6	0.50	0.2	0.6	0.50
CNRM_CM6_1_HR	-0.3	0.3	0.16	-0.5	0.3	0.22
MRI_ESM2_0	1.0	0.3	0.09	1.1	0.3	0.07
MPI_ESM1_2_LR	1.1	0.4	0.23	0.5	0.4	0.19

S 42.The Bottom Temperature statistics against observation for top four models on the Pacific Shelf for EBS in Summer for the period of 1982-2014, GAK yearly for the period of 1987-2014, NV Spring and Fall for the period of 1998-2014. Model Bias (Bs; unit: °C), standard deviation (STD), and trend (Tr; unit: °C/decade).

	EBS_Sum.			GAK1_Year			NV_Spr.			NV_Fal.		
	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend
Observation	-	0.8	-0.18	-	0.3	-0.02	-	0.7	-0.47	-	1.2	1.07
ACCESS_CM2	1.4	0.7	-0.26	-0.8	0.8	0.08	-0.2	0.2	0.17	-0.7	0.2	0.08
AWI_CM_1_1_MR	-0.5	0.7	0.04	-0.1	0.5	0.29	-0.9	0.4	0.16	-1.2	0.2	0.17
CNRM_CM6_1_HR	0.8	0.8	0.01	1.2	0.4	0.03	1.6	0.5	-0.41	1.9	0.4	-0.48
CESM2	1.4	0.5	-0.03	0.6	0.4	-0.12	0.2	0.5	-0.48	0.9	0.4	-0.25

S 43.The Bottom Temperature statistics against observation for top four models on the Pacific Shelf for SV spring (1985-2014) and Fall (1986-2014) and BCS Spring (1998-2014). Model Bias (Bs; unit: °C), standard deviation (STD), and trend (Tr; unit: °C/decade)

	SV_Spr.			SV_Fal.			BCS_Spr.		
	Bias	STD	Trend	Bias	STD	Trend	Bias	STD	Trend
Observation	-	0.6	-0.17	-	0.6	-0.25	-	0.3	-0.19
ACCESS_CM2	1.9	0.3	0.05	1.9	0.2	0.00	1.3	0.3	0.16
AWI_CM_1_1_MR	1.3	0.3	0.06	0.9	0.1	-0.05	0.4	0.3	0.09
CNRM_CM6_1_HR	4.3	0.3	0.10	3.9	0.2	0.02	3.7	0.2	-0.22
CESM2	1.9	0.4	-0.04	1.3	0.2	0.06	1.0	0.3	-0.27

S 44.Regionally averaged depth for each model and STD of the mean depths in meter for all models.

Models/Region	North Atlantic											Arctic				North Pacific		
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	BB	CAA	SBS	SC	BS	AS	BCS
ACCESS-CM2	93	100	128	90	99	85	150	224	202	196	103	794	255	186	106	72	134	151
ACCESS-ESM1-5	93	100	128	90	99	85	150	224	202	196	103	794	255	186	106	72	134	151
AWI-CM-1-1-MR	116	109	135	98	179	95	174	289	230	206	109	856	151	255	116	74	113	113
CAMS-CSM1-0	109	125	103	144	107	144	143	261	259	241	102	863	162	236	114	82	179	146
CanESM5	116	100	136	111	168	109	148	272	239	198	114	866	230	211	97	74	111	140
CESM2	199	190	205	201	165	264	198	301	346	307	108	822	244	297	122	116	400	273
CESM2-WACCM	199	190	205	201	165	264	198	301	346	307	108	822	244	297	122	116	400	273
CMCC-CM2-SR5	111	95	125	103	160	103	139	258	229	190	108	835	220	200	93	70	104	134
CNRM-CM6-1	112	97	129	107	163	105	141	259	232	192	110	844	223	205	94	72	107	137
CNRM-CM6-1-HR	112	101	139	98	168	103	157	268	214	195	107	816	202	181	91	75	152	150
CNRM-ESM2-1	112	97	129	107	163	105	141	259	232	192	110	844	223	205	94	72	107	137
EC-Earth3	118	100	141	111	180	111	149	279	245	198	114	870	239	212	98	75	113	142
GISS-E2-1-G	141	183	147	200	188	218	143	316	310	277	110	848	273	274	103	107	260	205
IPSL-CM6A-LR	118	100	141	111	180	111	149	279	245	198	114	870	233	212	98	75	113	142
MIROC6	291	339	325	316	149	284	253	458	519	404	212	975	339	417	225	158	961	1245
MIROC-ES2L	291	339	325	316	149	284	253	458	519	404	212	975	339	417	225	158	961	1245
MPI-ESM1-2-HR	120	115	151	124	177	120	169	270	234	222	106	860	201	181	93	83	170	217
MPI-ESM1-2-LR	129	75	132	120	188	106	196	314	285	240	111	794	236	247	115	99	208	302
MRI-ESM2-0	220	292	219	175	174	212	200	316	313	308	111	835	218	293	109	101	323	372
NorESM2-LM	152	207	144	127	196	150	155	307	273	199	113	870	238	212	94	83	201	242
TaiESM1	199	190	205	201	165	264	198	301	346	307	108	822	244	297	122	116	400	273
UKESM1-0-LL	116	100	135	111	169	109	147	268	239	198	113	866	230	211	98	74	110	140
STD	58	80	60	66	27	73	34	59	88	67	30	47	43	67	37	248	248	317

S 45. Model biases for SST in each ocean along the percent of cold bias occurrence in each ocean. In the North Atlantic, 37.19% of model instances have a cold bias and 62.18% of instances have warm biases. In the Arctic, 57.95% and 42.05% of model instances have cold and warm biases, respectively. In the North Pacific, 59.09% and 40.91% of model instances have cold and warm biases, respectively.

Models/Region	North Atlantic											Arctic				North Pacific		
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	BB	CAA	SBS	SC	BS	AS	BCS
ACCESS-CM2	1.45	2.08	2.01	0.32	-0.42	0.22	-0.70	-0.89	-0.97	-0.42	-0.21	-0.70	-0.02	-0.06	-0.40	-1.32	-1.03	-0.01
ACCESS-ESM1-5	4.38	5.52	5.53	3.52	1.58	2.59	0.80	-0.19	-0.50	-0.50	0.12	-0.40	0.15	0.13	0.12	-0.94	-0.33	0.62
AWI-CM-1-1-MR	-0.44	-0.85	0.02	-0.34	0.53	-1.31	-0.97	1.17	0.61	0.67	1.21	0.42	0.67	0.13	0.07	-1.45	-3.39	-1.29
CAMS-CSM1-0	1.86	2.73	1.02	0.13	-0.21	-1.48	-2.24	-1.64	-1.64	-1.05	-0.28	-0.56	0.07	-0.61	-0.42	-2.31	-4.34	-2.92
CanESM5	2.38	3.82	1.32	0.64	1.09	-0.81	-2.53	-2.10	-2.03	-1.31	0.23	-1.22	-0.04	-0.23	-0.33	-1.24	-3.93	-1.34
CESM2	5.83	7.07	6.06	3.31	0.28	2.11	1.06	0.73	0.30	-0.03	-0.61	-0.58	-0.08	-0.24	0.10	-0.10	-0.58	0.46
CESM2-WACCM	5.89	7.15	6.20	3.45	0.20	2.19	1.11	0.79	0.39	0.01	-0.65	-0.62	-0.12	-0.28	-0.24	-0.55	-0.83	0.23
CMCC-CM2-SR5	3.74	4.66	2.18	0.96	1.49	0.65	-1.12	-0.30	-0.31	0.38	1.09	0.60	0.80	1.06	1.10	0.69	-0.68	0.88
CNRM-CM6-1	1.68	1.36	0.18	0.21	1.07	0.12	-0.29	0.74	0.89	1.30	1.11	0.62	0.48	0.82	0.65	0.79	0.40	2.04
CNRM-CM6-1-HR	-0.75	-0.97	-0.94	-1.09	0.51	-0.97	-0.69	0.30	0.46	0.80	0.53	0.89	0.22	-0.08	0.17	0.74	1.19	2.28
CNRM-ESM2-1	2.69	2.50	1.07	1.05	1.85	0.94	0.47	1.53	1.63	1.84	1.55	0.86	0.58	1.38	0.98	1.70	1.16	2.53
EC-Earth3	2.80	3.42	0.40	-0.20	-0.67	-0.84	-1.85	-1.88	-1.78	-0.93	-0.09	-0.85	0.09	-0.17	0.28	1.10	0.20	1.15
GISS-E2-1-G	4.84	5.15	4.24	3.41	0.92	2.04	1.82	0.96	0.63	-0.36	-1.71	-1.24	-0.17	-0.57	-0.63	-3.19	-5.68	-4.35
IPSL-CM6A-LR	2.21	2.50	0.10	-0.05	0.98	0.13	-0.37	0.35	0.41	1.16	1.44	-0.40	0.35	0.38	0.57	0.97	-0.73	0.29
MIROC6	2.53	2.64	1.91	1.12	2.52	-0.10	-0.32	0.24	0.14	0.31	0.52	-0.78	-0.10	-0.27	-0.33	0.12	0.23	1.81
MIROC-ES2L	1.26	1.88	1.31	1.28	2.39	1.18	0.51	0.92	0.84	1.11	0.93	0.07	-0.07	0.03	-0.54	-1.44	-0.22	1.45
MPI-ESM1-2-HR	0.09	-0.66	-0.77	-0.25	-0.05	-1.15	-1.02	0.54	0.27	0.63	0.61	-0.23	0.17	-0.49	-0.40	-1.40	-2.27	-1.14
MPI-ESM1-2-LR	-0.91	-1.99	-1.98	-2.05	-0.87	-1.45	-1.22	0.01	-0.19	0.33	0.62	0.19	-0.04	-0.56	-0.66	-2.16	-4.05	-2.31
MRI-ESM2-0	0.97	1.10	0.29	-0.74	-0.24	-0.67	-0.51	-0.32	-0.32	0.38	0.12	-0.25	0.26	0.01	0.30	0.65	0.69	1.33
NorESM2-LM	3.48	3.46	1.46	0.01	-0.60	-0.62	-0.90	-0.69	-0.80	-0.43	-0.62	-0.61	-0.10	-0.22	0.01	-0.41	-1.60	-0.67
TaiESM1	6.49	7.42	6.33	3.23	0.78	2.49	1.37	1.14	0.71	0.01	-0.57	-0.87	-0.16	-0.48	-0.36	-0.97	-1.37	-0.13
UKESM1-0-LL	0.64	0.93	-1.63	-1.77	-0.81	-1.85	-2.29	-1.63	-1.45	-0.70	-0.45	-0.70	-0.09	-0.62	-0.41	-1.28	-1.52	0.26
Number of cold biases	3	4	4	8	8	11	15	9	10	9	9	15	11	14	11	14	16	9
Total number of Cold biases	90											51				39		
Total instances	22*11=242											22*4=88				22*3=66		
Percent of cold biases	%37.19											%57.95				%59.09		

S 46.Model biases for BT in each ocean along the percent of cold bias occurrence in each ocean. In the North Atlantic, 7.02% of model instances model have a cold bias and 92.98% of instances have a warm bias. In the Arctic, 48.86% and 41.14% of model instances have cold and warm biases, respectively. In the North Pacific, 38.63% and 61.37% of model instances have cold and warm biases, respectively.

Models/Region	North Atlantic											Arctic				North Pacific			
	GoM	WSS	CSS	ESS	GSL	SNS	CNS	NNS	SLS	NLS	HB	BB	CAA	SBS	SC	BS	EBS	AS	BCS
ACCESS-CM2	3.34	4.05	2.52	4.26	1.81	4.86	2.18	0.46	0.32	0.96	0.77	-1.21	-1.04	1.00	-0.82	-0.12	0.39	0.09	1.19
ACCESS-ESM1-5	6.56	7.86	6.08	7.97	4.35	7.54	2.69	0.34	0.23	0.77	1.06	2.58	-0.53	0.76	-0.29	0.93	1.42	1.23	1.80
AWI-CM-1-1-MR	0.13	3.14	2.53	2.87	1.45	1.27	3.28	2.46	2.80	4.67	1.57	3.47	-0.26	0.19	-0.13	-0.30	-0.48	-1.35	-0.39
CAMS-CSM1-0	2.04	5.36	2.23	4.86	2.40	3.70	1.09	1.59	0.48	0.94	-0.39	3.59	-0.14	0.38	-0.37	-1.44	-1.34	-1.90	-1.00
CanESM5	3.47	7.19	6.65	6.31	4.74	4.16	0.45	0.34	0.24	0.11	0.73	-0.7	-0.98	0.80	-0.59	-0.15	-0.02	-2.45	0.77
CESM2	2.10	4.68	3.18	5.78	3.15	4.59	3.71	2.83	3.58	4.09	0.08	3.68	0.19	0.31	0.10	0.68	0.90	-0.19	0.44
CESM2-WACCM	1.97	4.62	3.18	5.79	3.05	4.70	3.82	2.81	3.55	4.01	0.11	3.35	0.04	0.18	-0.24	0.33	0.59	-0.30	0.25
CMCC-CM2-SR5	3.59	7.41	5.35	4.33	3.88	4.14	0.86	0.76	0.93	0.40	0.42	-0.48	-0.50	0.09	0.62	2.06	2.25	0.32	0.97
CNRM-CM6-1	-0.01	1.50	-0.85	0.38	0.51	1.40	0.12	-0.47	0.03	0.45	0.36	-1.02	-0.44	0.34	-0.12	-0.10	-0.05	-0.36	2.14
CNRM-CM6-1-HR	-0.61	0.23	-0.62	0.36	0.02	0.99	0.48	-0.53	-0.40	-0.16	-0.15	-0.4	-0.55	0.46	-0.26	0.32	0.45	1.20	2.94
CNRM-ESM2-1	0.93	2.70	0.32	1.04	0.81	1.59	0.50	0.00	0.45	0.84	0.23	-1.12	-0.56	0.48	-0.29	0.07	0.23	0.12	2.23
EC-Earth3	3.00	7.56	7.36	8.33	4.24	3.83	0.23	0.23	0.36	0.21	1.31	-0.79	0.17	0.06	0.30	2.03	2.34	1.18	1.72
GISS-E2-1-G	3.62	4.19	2.97	4.94	3.40	4.44	2.66	0.72	0.99	0.47	-0.48	-1.82	-0.47	0.10	-0.65	-1.95	-2.00	-3.71	-2.74
IPSL-CM6A-LR	3.68	6.69	5.02	5.56	2.09	2.91	0.29	0.40	0.70	0.81	1.32	-0.36	0.40	0.30	0.01	0.27	0.28	-0.07	2.00
MIROC6	4.00	5.15	3.92	6.75	7.24	5.77	2.70	2.06	3.22	3.52	0.55	4.42	0.14	0.45	-0.45	0.53	1.13	0.14	-2.97
MIROC-ES2L	2.36	4.03	2.90	5.73	6.29	6.14	3.85	3.24	4.06	4.36	0.23	5.19	-0.25	0.27	-0.74	-0.33	0.25	-0.11	-2.98
MPI-ESM1-2-HR	2.44	4.41	4.02	6.07	3.43	1.90	1.35	1.00	1.17	1.64	1.07	0.94	-0.13	0.07	-0.25	0.17	0.24	0.12	1.40
MPI-ESM1-2-LR	2.44	1.50	1.63	0.00	0.92	0.44	1.25	0.98	0.84	0.86	0.86	2.5	0.20	0.29	-0.38	-0.02	0.29	-0.80	0.86
MRI-ESM2-0	-1.13	0.77	0.31	3.20	3.16	2.51	1.27	0.98	1.33	1.99	0.51	2.51	0.39	0.83	0.32	0.97	1.18	0.58	0.07
NorESM2-LM	-1.42	-0.83	-2.22	0.91	1.48	2.64	4.24	2.87	3.46	4.59	0.56	-0.07	-0.05	0.05	-0.28	0.10	0.40	-3.21	-4.66
TaiESM1	2.00	4.28	2.93	4.71	2.12	3.16	3.27	2.21	3.00	3.38	-0.22	2.39	-0.19	0.00	-0.36	-0.42	-0.19	-1.39	-0.64
UKESM1-0-LL	2.01	5.28	3.41	2.80	1.91	1.40	0.48	0.34	0.44	0.01	0.40	-1.13	-0.64	0.59	-0.56	0.14	0.36	0.26	1.26
Number of cold biases	4	1	3	0	0	0	0	3	1	1	4	11	15	0	17	9	6	12	7
Total number of Cold biases	17											43				34			
Total instances	22*11=242											22*4=88				22*4=88			
Percent of cold biases	%7.02											%48.86				%38.63			