Clim Dyn
DOI 10.1007/s00382-017-3652-7

More reliable coastal SST forecasts from the North American multimodel ensemble

G. Hervieux[1,2] · M. A. Alexander[2] · C. A. Stock[3] · M. G. Jacox[4,5] · K. Pegion[6] · E. Becker[7] · F. Castruccio[8] · D. Tommasi[9]

Received: 29 September 2016 / Accepted: 16 March 2017

Abstract The skill of monthly sea surface temperature (SST) anomaly predictions for large marine ecosystems (LMEs) in coastal regions of the United States and Canada is assessed using simulations from the climate models in the North American Multimodel Ensemble (NMME). The forecasts based on the full ensemble are generally more skillful than predictions from even the best single model. The improvement in skill is particularly noteworthy for probability forecasts that categorize SST anomalies into upper (warm) and lower (cold) terciles. The ensemble provides a better estimate of the full range of forecast values than any individual model, thereby correcting for the systematic over-confidence (under-dispersion) of predictions from an individual model. Probability forecasts, including tercile predictions from the NMME, are used frequently in seasonal forecasts for atmospheric variables and may have many uses in marine resource management.

Keywords Seasonal prediction · SST anomaly · Coastal ecosystems · Climate models · Multimodel ensemble forecast

1 Introduction

Numerical weather and climate forecasts have greatly improved over the last 30 years and now have the capability to provide useful seasonal forecasts (National Research Council 2010). Coupled global climate models (CGCMs), originally developed to study climate variability and change over centennial scales, are now being used to make forecasts on seasonal and even decadal timescales (e.g. Stockdale et al. 1998; Wang et al. 2009; Yang et al. 2012; Siedlecki et al. 2016). Due to the much lower frequency variability of the ocean compared to the atmosphere, seasonal forecast

This paper is a contribution to the special collection on the North American Multi-Model Ensemble (NMME) seasonal prediction experiment. The special collection focuses on documenting the use of the NMME system database for research ranging from predictability studies, to multi-model prediction evaluation and diagnostics, to emerging applications of climate predictability for subseasonal to seasonal predictions.This special issue is coordinated by Annarita Martiotti (NOAA), Heather Archambault (NOAA), Jin Huang (NOAA), Ben Kirtman (University of Miami) and Gabriele Villarini (University of Iowa).

Electronic supplementary material The online version of this article (doi:10.1007/s00382-017-3652-7) contains supplementary material, which is available to authorized users.

✉ G. Hervieux
gaelle.hervieux@noaa.gov

[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA

[2] NOAA Earth System Research Laboratory, Physical Sciences Division, 325 Broadway R/PSD1, Boulder, CO 80305-3328, USA

[3] NOAA Geophysical Fluid Dynamics Laboratory, Princeton University Forrestal Campus, 201 Forrestal Road, Princeton, NJ 08540-6649, USA

[4] Institute of Marine Sciences, University of California, Santa Cruz, CA 95064, USA

[5] NOAA Southwest Fisheries Science Center, Environmental Research Division, 99 Pacific St., Ste. 255A, Monterey, CA 93940, USA

[6] Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, 4400 University Drive, MS6C5, Fairfax, VA 22030, USA

[7] NOAA/Climate Prediction Center and INNOVIM LLC, 5830 University Research Court, College Park, MD 20740, USA

Published online: 30 March 2017

systems are expected to have more skill in predicting ocean variables (Goddard et al. 2001). Initially, CGCMs were primarily used to predict sea surface temperature (SST) anomalies in the tropical Pacific associated with El Niño and the Southern Oscillation (ENSO; Anderson et al. 2003; Latif et al. 1994; Kirtman and Zebiak 1997) but they have recently been shown to have skill in other regions (Becker et al. 2014; Kirtman et al. 2014; Stock et al. 2015). In addition to being a critical measure of the climate system, SST is one of the few ocean variables for which we have the extensive data coverage in space and time that is necessary for a broad evaluation of forecast skill. SST predictions can also be used readily in fisheries management applications due to the strong influence of temperature on the distribution and abundance of marine organisms. For example, SST forecasts are currently being used by managers to protect fish stocks and reduce the cost of fishing (Hobday et al. 2011; Eveson et al. 2015). Stock et al. (2015) highlighted the potential utility of CGCM-based forecast systems for SST prediction in coastal ecosystems and subsequent application to marine resources. Predictions of the marine environment may be even more critical in the future as the climate changes (Hobday and Hartog 2014).

Utility of predictions to the marine resource sector is not only dependent on achieving adequate forecast skill at the temporal and spatial scales of relevance to decision-makers, but also on reliably representing the uncertainty of the predictions. Prediction uncertainties arise from two types of error: uncertainties in model initialization and inadequacies in a model's formulation (resolution, parameterizations, etc.; e.g. Ji et al. 1998; Rosati et al. 1997). An ensemble of simulations with the same model but different initial conditions can be used to assess the former, while a multimodel ensemble strategy can be used for the latter (Jin et al. 2008; Palmer et al. 2004). The multimodel mean (MMM), including models with a wide range of skill, often outperforms all individual models or a subset of "better" models (Hagedorn et al. 2005; Weigel et al. 2008). Stock et al. (2015) assessed the SST forecast skill in coastal regions and mechanisms that underlie it using two state-of-the-art forecast systems. Here we evaluate SST predictions from the 14 individual models participating in Phase I of the North American Multimodel Ensemble (NMME; Kirtman et al. 2014) project, greatly expanding on the two CGCMs used by Stock et al. (2015). Using a suite of models to assess forecast skill is relatively new with the release of the North American Multimodel Ensemble data occurring within the last 2 years. While climate models have been used to make SST forecasts in the Nino region for more than a decade there are fewer papers that explore the prediction skill elsewhere over the global oceans, especially in coastal regions. Traditionally the coarse resolution of seasonal prediction models has been a barrier to their use for marine ecosystems applications. This work demonstrates not only that multimodel coastal SST predictions are skillful, but also that multimodel predictions greatly improve the probabilistic skills and provide a more reliable estimate of uncertainty.

In the following, we focus our analysis on whether using multiple models consistently improves skill relative to individual models across coastal regions, even when individual models within the ensemble may have notable deficiencies at the regional scale. Furthermore, we ask whether this improvement is reflected in both deterministic and probabilistic skill metrics.

2 Data and methods

2.1 Large marine ecosystems (LMEs)

Large marine ecosystems (LMEs, http://lme.edc.uri.edu/) are coherent ocean areas primarily located along continental margins where primary productivity is generally higher than in open ocean areas. LMEs have been defined based on ecological criteria, bathymetry, hydrography, productivity and trophic relationships (Sherman and Duda 1999). The LMEs are ~200,000 square kilometers or larger and produce ~80% of the catch of global marine fisheries (Sherman et al. 2009). We focus on the 11 LMEs located around the US and Canada including: East Bering Sea, Gulf of Alaska, California Current, Gulf of California, Gulf of Mexico, Southeast and Northeast US Continental Shelf, Scotian Shelf, Newfoundland-Labrador Shelf, Insular Pacific-Hawaiian and Aleutian Islands (Fig. 1).

2.2 North American multimodel ensemble (NMME)

The North American Multimodel Ensemble (NMME) is a collaborative effort based on seasonal forecast systems using coupled atmosphere–ocean–sea ice-land models from US and Canadian modeling centers (Kirtman et al. 2014). In addition to having different model formulations, the various forecast systems use different initialization methods. Our results are based on hindcasts (i.e., retrospective forecast experiments predicting what happened during the past) to validate the forecast system over a common 28-year period (1982–2009) for all the models from the NMME phase 1. Additional information about the forecast systems is provided in Table 1. While the 14 models have varying native resolution, all of the output has been interpolated to a 1° latitude by 1° longitude grid. The number of forecast

[8] NCAR/Climate and Global Dynamics, 1850 Table Mesa Drive, Boulder, CO 80305, USA

[9] Atmospheric and Oceanic Sciences Program, Princeton University, 300 Forrestal Road, Sayre Hall, Princeton, NJ 08544, USA

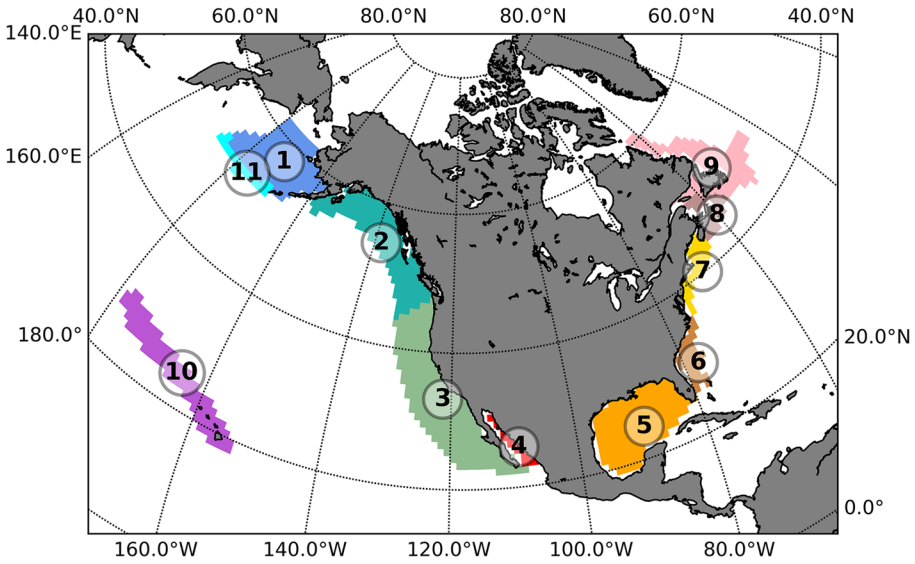

Fig. 1 Large marine ecosystems studied. *1* East Bering Sea (EBS; 1,193,601 km^2 or 179 grid pts), *2* Gulf of Alaska (GoA; 1,491,252 km^2 or 187 grid pts), *3* California Current (CC; 2,224,665 km^2 or 210 grid pts), *4* Gulf of California (GoC; 216,344 km^2 or 16 grid pts), *5* Gulf of Mexico (GoM; 1,530,387 km^2 or 135 grid pts), *6* Southeast US Continental Shelf (SEUS; 303,029 km^2 or 28 grid pts), *7* Northeast US Continental Shelf (NEUS; 308,554 km^2 or 30 grid pts), *8* Scotian Shelf (SS; 412,676 km^2 or 26 grid pts), *9* Labrador-Newfoundland (LN; 674,862 km^2 or 114 grid pts), *10* Insular Pacific Hawaiian (IPH; 975,493 km^2 or 89 grid pts), *11* Aleutian Islands (AI; 220,000 km^2 or 27 grid pts), which is numbered as 65 in the list of LME regions (http://www.lme.noaa.gov)

Table 1 NMME models for phase 1

Model	Organization	Hindcast period	Ensemble size	Lead times (month)	References
Active					
NCEP-CFSv2	NCEP	1982–2010	24	0–8	Saha et al. (2014)
GFDL-CM2p1	GFDL	1982–2012	10	0–11	Delworth (2006)
GFDL-CM2p1-aer04	GFDL	1982–2015	10	0–11	Delworth (2006)
GFDL-CM2p5-FLOR-A06	GFDL	1980–2015	12	0–11	Vecchi et al. (2014)
GFDL-CM2p5-FLOR-B01	GFDL	1980–2015	12	0–11	Vecchi et al. (2014)
CMC1-CanCM3	Canadian Meteorological Center (CMC)	1981–2011	10	0–11	Merryfield et al. (2013)
CMC2-CanCM4	CMC	1981–2011	10	0–11	Merryfield et al. (2013)
COLA-RSMAS-CCSM4	NCAR	1982–2015	10	0–11	Infanti and Kirtman (2016)
NASA-GMAO-062012	NASA	1981–2015	10	0–8	Vernieres et al. (2012)
Retired					
NCEP-CFSv1	NCEP	1981–2009	15	0–8	Saha et al. (2006)
COLA-RSMAS-CCSM3	NCAR	1982–2015	6	0–11	Kirtman and Min (2009)
IRI-ECHAM4p5-AnomalyCoupled	IRI	1982–2012	12	0–7	DeWitt (2005)
IRI-ECHAM4p5-DirectCoupled	IRI	1982–2012	12	0–7	DeWitt (2005)
NASA-GMAO	NASA	1981–2009	8	0–8	Vernieres et al. (2012)

ensemble members for the individual models varies from 6 to 24, and the forecast length varies from 8 to 12 months (Table 1). Model drift is removed using a method similar to Stock et al. (2015), whereby a bias correction estimate, computed as the difference between the lead-dependent forecast and climatology of the forecast month, is subtracted from each forecast. However, here a cross-validation methodology is applied (e.g. Stockdale 1997) where

the value for the forecast being evaluated is excluded from the bias correction estimate. For those models that are not initialized on the first of the month (i.e. NCEP-CFS and NASA-GEOSS), we chose to use the ensemble members that started within 15–20 days prior to the first lead month.

2.3 Historical SST anomaly estimates

Following Stock et al. (2015), the observed SSTs are obtained from the NOAA Optimum Interpolation Sea Surface Temperature version 2 (OISSTv2; Reynolds et al. 2007), which has a nominal resolution of 0.25°. Stock et al. compared OISSTv2 SST anomalies with individual observations from the NOAA World Ocean Database (WOD13, Boyer et al. 2013) over 7 LMEs around the US. They found the two datasets to be consistent with each other with monthly correlations of the SST anomalies exceeding 0.75 for all but one LME. Individual in situ measurements can have RMS errors on the order of 1 °C and biases of several tenths of a degree (Reynolds et al. 2007). The OISSTv2 is a blended analysis that includes satellite measurements as well as observations from ships and buoys, where the data are weighted by their signal-to-noise ratio as well as the distance to the grid point. In addition, the uniform grid of the analysis enables a consistent comparison between model and observations across the disparate LMEs.

OISST data are averaged over individual LMEs to generate persistence forecasts and evaluate model skill. Following Stock et al. (2015) persistence forecasts are initiated using the SST anomaly in the month immediately preceding the first month's forecast. For comparison with the model predictions (which are mainly initialized on the first of the month), this is termed the "zero" month persistence forecast.

2.4 Skill metrics

We evaluate forecast skill using two common deterministic skill metrics, the anomaly correlation coefficient (ACC) and root mean square error (RMSE). The ensemble members from each model are averaged together and the ensemble mean is used to estimate the ACC and the RMSE. We also evaluate predictions using the Brier Score (BrS), a probabilistic forecast metric. The BrS is a measure of the mean-square error of probability forecasts for whether or not an event will occur, which can be relative to a probability category. In this study, the BrS is used to measure the forecast probability error in SST anomaly terciles, where the forecast probability ranges from 0 to 1, and the observed probability is either 0 or 1. For the BrS, we use the individual ensemble members for each model to estimate the forecast probability of an event occurring in a given tercile.

2.5 Anomaly correlation coefficient (ACC)

The ACC, a widely used measure for forecast verification (Jolliffe and Stephenson 2003), indicates the relative association between the predicted and observed anomalies but not the magnitude of their differences. If the predicted and observed anomalies are perfectly coincident then the ACC will have the maximum value of 1 and if they are 180° out-of-phase, the ACC will have a minimum value of −1.

The ACC as a function of initialization month (m) and lead time (t) can be written as

$$ACC(t,m) = \frac{\left(\sum_{\alpha=1}^{N}\left(F'_{\alpha}(t,m) \times O'_{\alpha}(t,m)\right)\right)}{\sqrt{\sum_{\alpha=1}^{N} F'_{\alpha}(t,m)^2 \sum_{\alpha=1}^{N} O'_{\alpha}(t,m)^2}}, \quad (1)$$

where F' is the forecast anomaly, O' is the verification field anomaly, and ACC is calculated over the period 1982–2009 ($N = 28$).

2.6 Root mean square error (RMSE)

The RMSE is also a common measure of forecast accuracy and indicates the magnitude of forecast error. The RMSE is 0 for perfect forecasts and increases with the amplitude of the difference between forecasts and observations.

The RMSE as a function of initialization month (m) and lead time (t) can be written as

$$RMSE(t,m) = \sqrt{\frac{1}{N}\sum_{\alpha=1}^{N}(F'_{\alpha}(t,m) - O'_{\alpha}(t,m))^2}, \quad (2)$$

where F' is the forecast anomaly, O' is the verification field anomaly, the RMSE is calculated over the period 1982–2009 ($N = 28$).

2.7 The Brier score (BrS)

Brier scores (BrS; Brier 1950; Wilks 1995) measure the average squared forecast probability error. The probability forecast is computed by the fraction of ensemble members exceeding a given threshold. Here the thresholds are the upper and lower terciles of the distribution, representing warmer and colder than average temperature, respectively; the BrS for the middle tercile is not presented as forecasts for "near-normal" conditions often exhibit little skill (e.g., see van den Dool and Toth 1991).

A perfect BrS is 0. As the difference between the forecast probability and the observed frequency increases, BrS increases to a maximum value of 1. The BrS weights larger errors more than smaller ones.

The BrS as a function of initialization month (m) and lead time (t) can be written as

$$\mathrm{BrS(t,m)} = \frac{1}{N}\sum_{\alpha=1}^{N}(f_{\alpha}(t,m) - o_{\alpha}(t,m))^2, \tag{3}$$

where f is the forecast anomaly probability of an event, o is the verification field anomaly probability of an event, the BrS is calculated over the period 1982–2009 ($N = 28$).

The BrS can be decomposed into three components (Murphy 1973) for K probability categories: Reliability (REL), Resolution (RES) and Uncertainty (UNC).

$$\mathrm{BrS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \tag{4}$$

$$\mathrm{REL} = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2, \tag{5}$$

$$\mathrm{RES} = \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2, \tag{6}$$

$$\mathrm{UNC} = \bar{o}(1 - \bar{o}), \tag{7}$$

where $\bar{o} = \sum_{t=1}^{N}\frac{o_t}{N}$ is the climatological base rate for the event to occur, n_k is the number of forecast with the same probability category, $\bar{o}_k$ is the relative observed frequency for a forecast probability f_k in the k probability class.

The reliability indicates how well the a priori predicted probability forecast of an event coincides with the posteriori observed frequency of the event. As defined here the reliability increases as REL decreases and approaches zero for good forecasts. The resolution indicates how well forecasts distinguish situations with distinctly different frequencies of occurrence; larger values indicate higher resolution. In the worst case, when the climatic probability is always forecast, the resolution is zero. In the best case, when the conditional probabilities are either zero or one, the resolution is equal to the uncertainty. The uncertainty measures the variability of the observations, and is independent of the forecast. It indicates the degree to which situations are easy or difficult to predict; its minimal value is zero when the event never or always occurs and its maximum value is reached when the event occurs 50% of the time. The uncertainty only depends on the observed frequency; since we are computing the BrS for terciles, the observed frequency is 1/3 and the uncertainty is $1/3 \times (1 - 1/3) = 0.22$.

3 Assessment of SST anomaly forecasts

Forecast skill of the NMME ensemble mean predictions for the 11 LMEs, as indicated by the ACC, is shown in Fig. 2. The multimodel mean (MMM) ACC values are shown as a matrix with initialization month on the x-axis and forecast lead on the y-axis. As found by Stock et al. (2015), the ACC varies widely by LME. ACC values are mostly positive and significantly above zero at the 95% level, except in the cases of the Southeast and Northeast US Continental Shelf and the Scotian Shelf. Skillful MMM forecasts that also exceed persistence occur in most regions at varying leads, including forecasts of more than 6 months. However, for very short lead times, the skill of persistence forecasts (see Figs. S1 and S2 in the supporting material) is often as good as or better than dynamical forecasts. This is not unexpected in most regions, as considerable skill derives from the large thermal inertia of the ocean surface mixed layer. ACCs are significantly above persistence for multiple initialization months and forecast leads in the Gulf of Alaska and the Insular Pacific-Hawaiian regions. While the ACCs are consistently high for all initialization months and leads in the Labrador-Newfoundland region, only a few exceed those based on persistence.

Many of the matrices in Fig. 2 exhibit higher (lower) ACC values along a diagonal, which occurs when there is more (less) skill for the month being predicted regardless of lead. For example, in the Gulf of Alaska region, the forecasts for February and March (e.g. 7–8 month forecasts initialized in July) have higher skill than forecasts for other months, resulting in higher ACC values from the upper left to lower right across the matrix. Enhanced ACCs for a predicted month tend to occur when the skill arises from reliable impacts that depend on the season, including: ENSO teleconnections (e.g. Barnston 1994; Jacox et al. 2017), the reemergence of winter temperature anomalies that persist below the mixed layer in summer (e.g. Alexander and Deser 1995) or Arctic sea ice extent conveying the imprint of fall SST anomalies over the winter season (Blanchard-Wrigglesworth et al. 2011). Diagonals of higher/lower skill are also apparent in Fig. 3, which shows the RMSE of the MMM as a function of initialization month and lead. For RMSE, however, it is harder to interpret the diagonals in terms of skill as the magnitude of the error may also reflect seasonal variability with higher RMSE during months of greater variance.

The ACCs averaged over all of the initialized months as a function of forecast lead-time for all 14 models, the MMM and persistence are shown in Fig. 4. The MMM generally gives the best forecast, even in cases where some models in the ensemble perform poorly. Exceptions are the Northeast US Continental Shelf and Scotian Shelf regions where the overall forecast skill is low and persistence is generally better than all of the model forecasts including the MMM. The forecast skill decreases with lead, although the rate of decline varies between regions; e.g., there is significant skill out to a year in the Labrador-Newfoundland

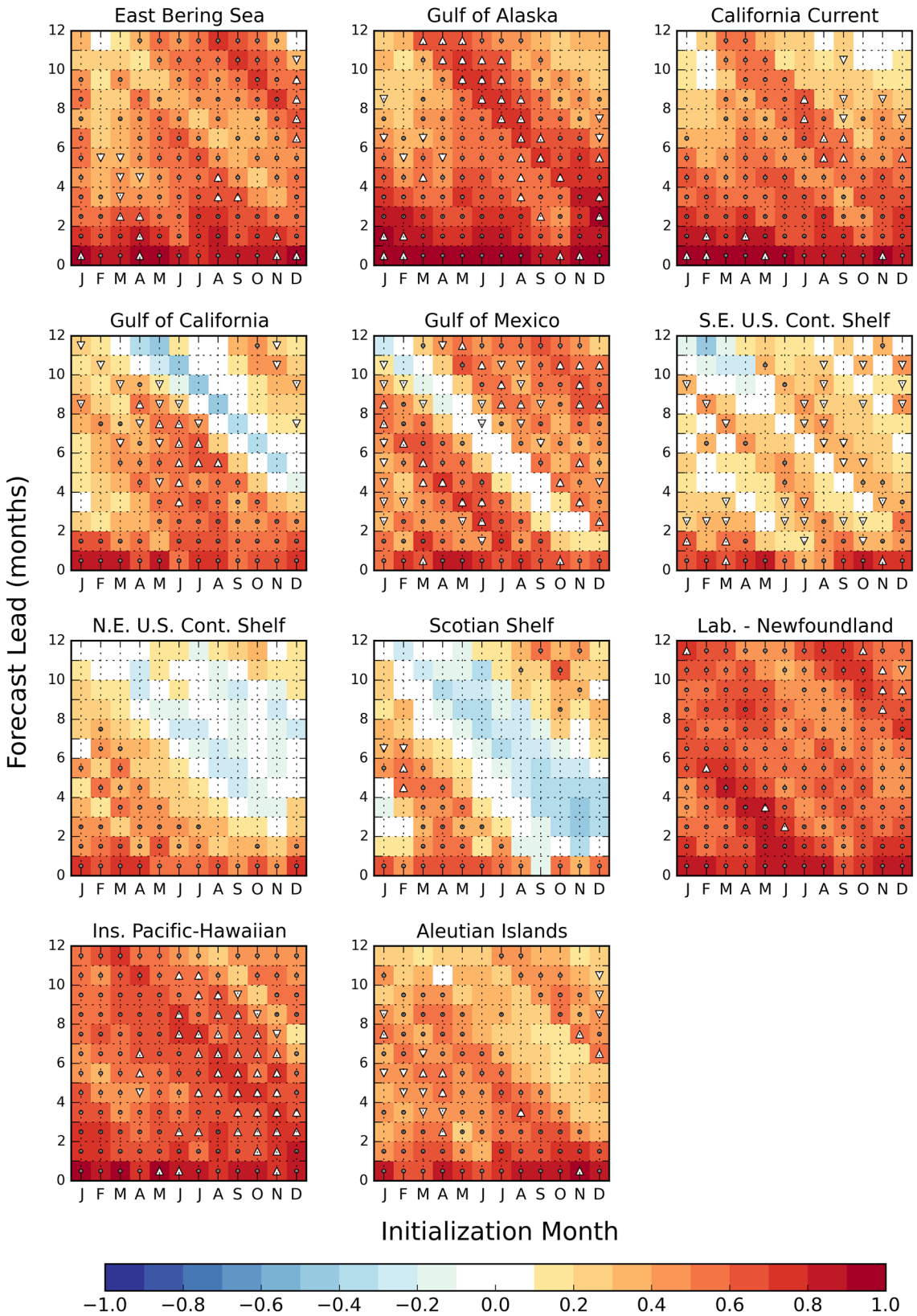

Fig. 2 Anomaly correlation coefficients (ACCs) between observations and the multimodel mean monthly forecasts as a function of the initialization month and lead time for the 11 LMEs. *Gray dots* indicate ACCs significantly above 0 at 95% level; *White upward triangles* indicate ACCs significantly above persistence at 90% level with ACC > 0.5; *White downward triangles* indicate ACCs significantly above persistence at 90% level with ACC < 0.5

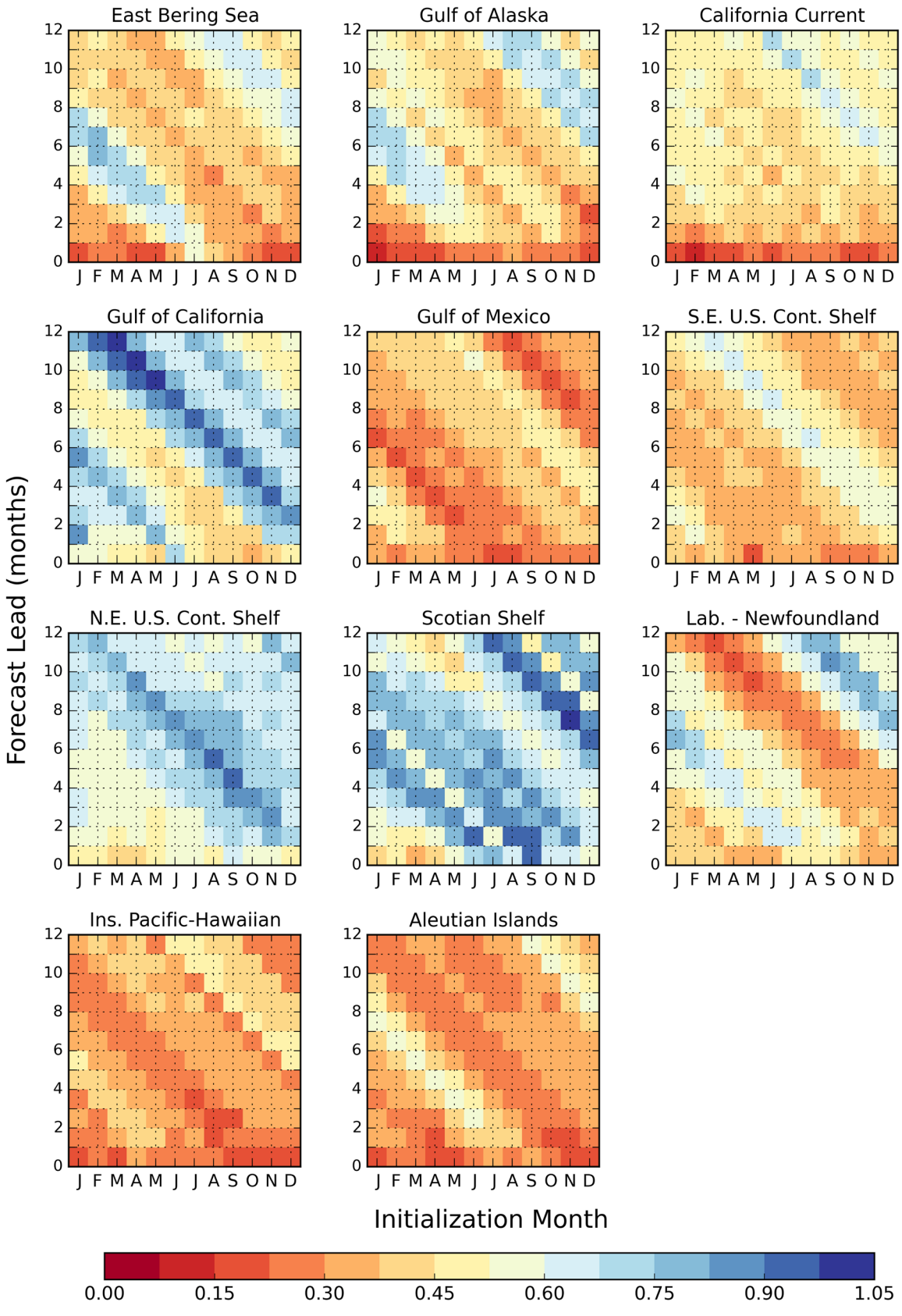

Fig. 3 Root mean square error (RMSE) of the multimodel mean as a function of forecast initialization month and lead time for each of the 11 LMEs

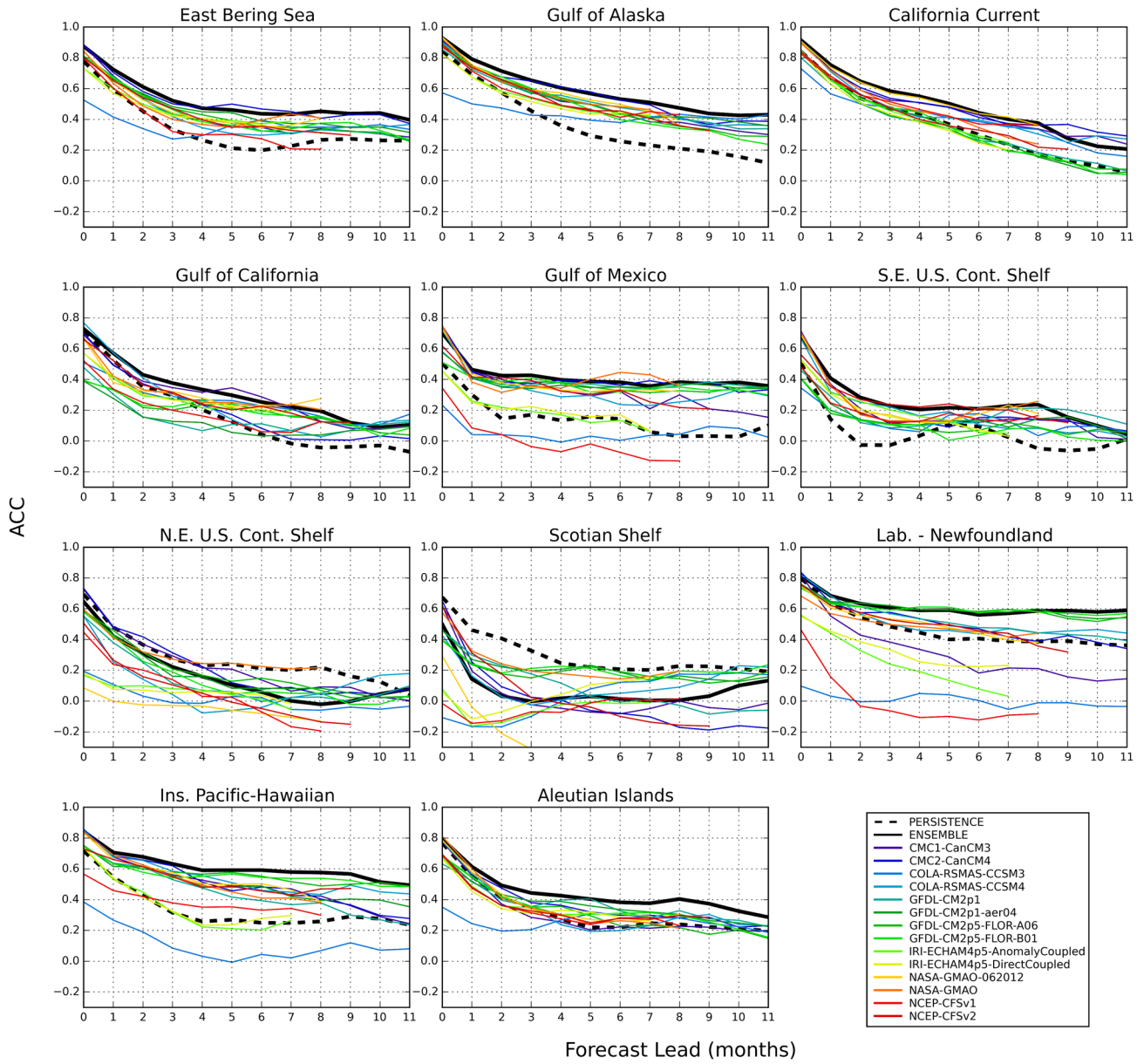

Fig. 4 Average of the ACCs over all 12 initialization months as a function of forecast lead time for each LME

region largely due to the strong persistence of SST anomalies in that LME.

The RMSE averaged over the initialization months as a function of forecast lead-time (Fig. 5) suggests that errors could be introduced when the model is initialized. For three models, COLA-RSMAS-CCSM3, IRI-ECHAM4p5-AnomalyCoupled and IRI-ECHAM4p5-DirectCoupled, the value of the RMSE is sometimes largest at the start of the forecast, which is opposite of what one would expect—that forecast errors would grow with time. The initialization errors are largest for higher latitude LMEs including the Aleutian Islands, East Bering Sea, Gulf of Alaska, Scotian Shelf and Labrador-Newfoundland. These three models all use the variational optimal interpolation scheme, an early ocean assimilation developed by Derber and Rosati (1989), to initialize the ocean component of the forecast system (Kirtman and Min 2009; DeWitt 2005). The sea-ice is not initialized from observations but is taken randomly from the free running model, meaning that the initial ice conditions can be far from the actual state. These ocean assimilation and prediction systems, along with the CFSv1 layed a foundation from which to build but have since been replaced with improved approaches.

To provide an overall evaluation of the individual models and the MMM forecast skill, we have computed the ACC, RMSE and BrS averaged over all initialization months and

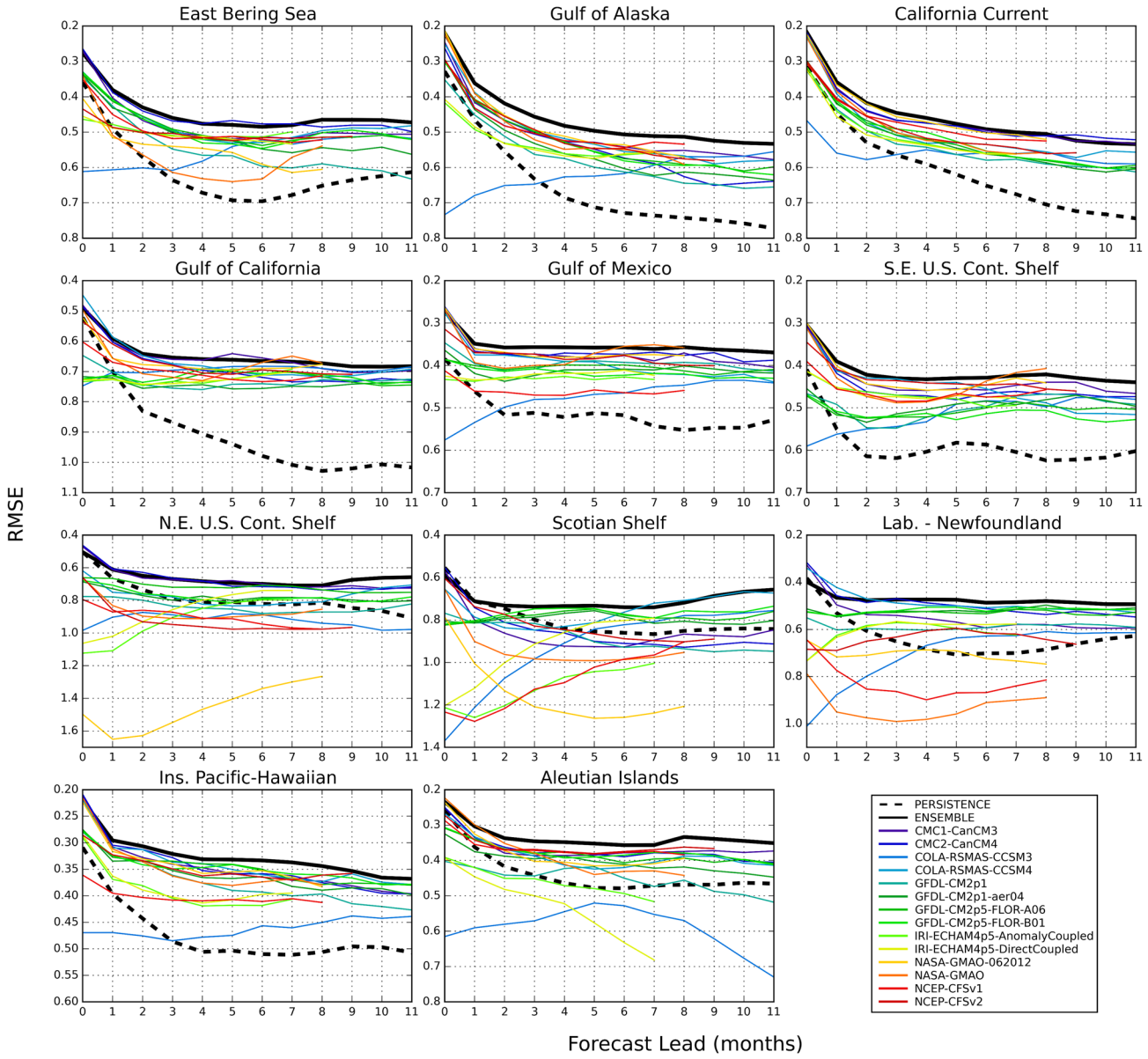

Fig. 5 Average of RMSEs over all initialized months as a function of forecast lead time for each LME. The abscises have been inverted to facilitate comparison with ACC plots

leads (Fig. 6). There is general agreement between these three metrics, which indicate that: (1) in most cases the MMM forecast has better deterministic skill than any of the individual models, (2) in all cases the MMM forecast has much better probabilistic skill than any of the individual models (3) in most cases individual models are more skillful than persistence, (4) different models have fairly similar behavior in each LME.

The best individual models often vary by system. Four models, CFS-v1, CCSM3, and the IRI-Anomaly and IRI Direct-Coupled, have lower skill within the Hawaiian region and LMEs situated in the Atlantic Ocean as indicated by the ACC values (Fig. 6, top left). Nevertheless, the two IRI models that have less skill in the Insular Pacific-Hawaiian LME have better skill in the Gulf of Alaska and California Current relative to other models. Variations in skill may offer process level insights into the relative strengths and weaknesses of the individual models.

The mean RMSE ranges from approximately 0.3° to 1.0 °C for the different LMEs and forecast systems when averaged over both lead and initialization month (Fig. 6, top right). In general, the ACC and RMSE provide similar assessments of the relative forecast skill; e.g., that forecasts are poor in the northeast US and Scotian Shelf and

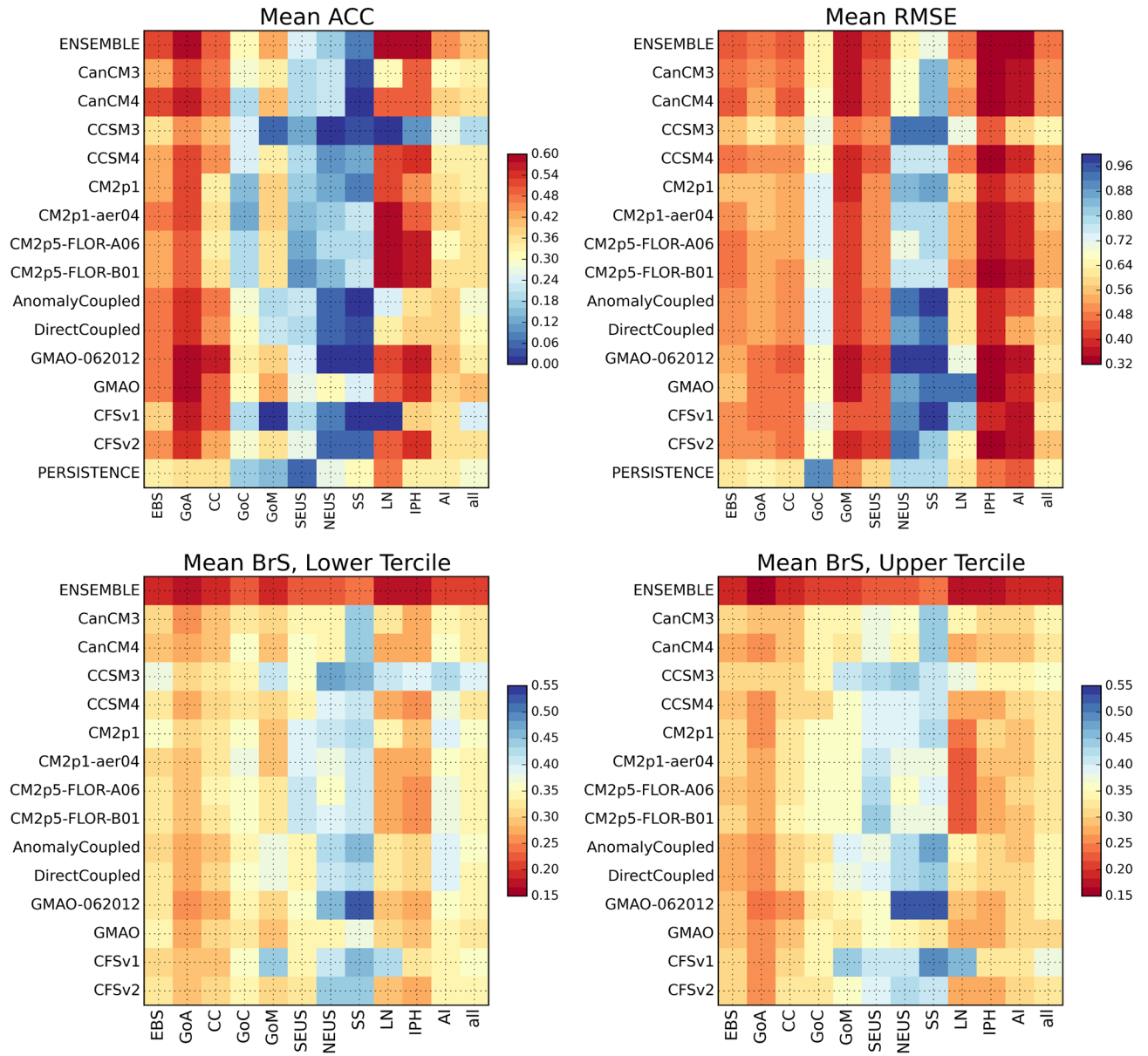

Fig. 6 Skill metrics averaged over all initialization months and lead times for each LME as well as all LMEs combined (all) (*x*-axis) and for each model as well as persistence and the ensemble mean (*y*-axis). (*top left*) ACC, (*top right*) RMSE, (*bottom*) BrS for (*left*) lower or cold tercile and (*right*) upper or warm tercile. The *color scale* for all metrics is arrayed so that higher skill is shown in *red* and lower skill in *blue*

relatively skillful for the Hawaiian and Aleutian Island LMEs. The relatively low RMS errors in the Gulf of Mexico, the Southeast U.S. Continental Shelf, the Insular Pacific-Hawaiian, and the Aleutian Islands are partly due to the limited SST variability in those four regions (Table 2), which results in smaller departures between the predicted and observed SST. Forecasts for low-amplitude anomalies can have both small ACC and RMSE values (Koh et al. 2012), as is the case for forecasts of December-January SSTs in the Gulf of Mexico region (see Figs. 2, 3). Furthermore, the MMM forecasts generally under-predict the observed variability (compare Table 3 with Table 2), which results in an underestimate of the RMSE (e.g. Taylor 2001; Koh et al. 2012).

The mean Brier Scores, presented for the probability forecasts for SST anomalies in the upper and lower terciles, are shown in the bottom of Fig. 6. The BrS estimated from all models in the NMME is substantially better (lower values) than the BrS of any individual model. The probabilistic skill depends on the spread among

Table 2 The monthly SST standard deviation averaged over all months or 3-month seasons for the 1982–2009 period of OISSTv2 spatial average for each LME

LME	EBS	GoA	CC	GoC	GoM	SE US	NE US	SS	LN	IPH	AI
Annual	0.53	0.59	0.54	0.71	0.39	0.43	0.65	0.66	0.54	0.41	0.37
D-J-F	0.43	0.51	0.51	0.92	0.39	0.53	0.70	0.52	0.35	0.44	0.30
M-A-M	0.44	0.56	0.60	0.68	0.51	0.48	0.60	0.53	0.40	0.45	0.30
J-J-A	0.75	0.74	0.54	0.72	0.31	0.33	0.63	0.81	0.86	0.44	0.48
S-O-N	0.50	0.55	0.52	0.52	0.36	0.39	0.68	0.73	0.64	0.31	0.41

Table 3 The monthly SST standard deviation averaged over all months or 3-month seasons for the 1982–2009 period of of MMM spatial average for each LME at 6-month lead

LME	EBS	GoA	CC	GoC	GoM	SE US	NE US	SS	LN	IPH	AI
Annual	0.28	0.38	0.26	0.20	0.17	0.16	0.31	0.31	0.30	0.18	0.25
D-J-F	0.29	0.41	0.23	0.16	0.23	0.13	0.26	0.26	0.34	0.16	0.28
M-A-M	0.25	0.28	0.22	0.18	0.12	0.12	0.33	0.30	0.28	0.13	0.22
J-J-A	0.28	0.34	0.29	0.27	0.16	0.14	0.34	0.42	0.29	0.20	0.23
S-O-N	0.30	0.49	0.30	0.19	0.17	0.25	0.31	0.29	0.29	0.23	0.27

ensemble members (Sooraj et al. 2012). The multimodel ensemble better represents the actual range of predicted values as the total number of members is higher but also because different models provide a wider diversity of outcomes that is more consistent with observed ranges than any individual model.

The Brier score depends on the reliability (agreement between the forecast and observed probability of an event occurring), the resolution (ability to separate situations into different categories), and the uncertainty (a measure of the observed variability). As illustrated by reliability diagrams for forecasts of the probability of an SST anomaly in the upper or lower tercile at 0 and 4 month leads, the individual models are under-dispersed ("over confident"), especially when compared with the multimodel ensemble, as the former under-forecast rare events and over-forecast events with high frequency of occurrence (Fig. 7). In contrast, the MME forecast is very similar to the actual probability of an SST anomaly being in the upper or lower tercile over nearly the full range of probabilities. For example, when the multimodel ensemble predicts an 80% probability of the 4-month forecast being in the upper tercile the actual probability of occurrence was ~83% but when the individual models forecast a probability of 80% the observed SST anomalies occurred in the upper tercile only about 40–60% of the time (Fig. 7, 3rd panel).

For a reliable prediction system, which has on average the correct amount of spread given its skill, the RMSE should be equal to the spread (Buizza 1997). We estimate the spread using the standard deviation across the ensemble members of the individual ensemble members or for all in the NMME and plot SST spread normalized by the RMSE as a function of forecast lead (Fig. 8). The spread of all of the individual NMME models is smaller than the RMSE (Fig. 8), which is typical of under-dispersive or overconfident prediction systems. By this measure the multimodel ensemble is somewhat over-dispersive.

Skill matrices for the reliability and resolution components of the BrS for all LMEs and prediction systems including the multimodel ensemble are shown in Fig. 9. The values are based on all forecasts irrespective of the initialization month and lead. The results indicate that the higher BrS skill of the multimodel ensemble forecasts come from improved resolution as well as better reliability.

4 Summary and discussion

We have analyzed the forecast skill of the North American Multimodel Ensemble using three different metrics: anomaly correlation, root mean square error and the Brier score. The results indicate that current global climate forecast systems with relatively coarse oceanic and atmospheric resolution have skill in forecasting SST anomalies in many coastal LME-scale regions, confirming the findings of Stock et al. (2015). Forecast skill is highly dependent on the month being predicted, with certain months producing higher or lower seasonal predictability regardless of the initialization time and duration of the forecast. The forecast skill varied widely by region, with relatively high skill in the Pacific, especially in the Bering Sea and Gulf of Alaska, and in the vicinity of Newfoundland, but limited skill in regions along the US east coast.

Several factors influence regional forecast skill, including the ability of the models to simulate large-scale climate phenomena, such as ENSO and its teleconnections (e.g., Jacox et al. 2017), ocean–ice interactions and gyre circulations. Changes in monthly forecast skill can result from the capacity of models to simulate physical processes that evolve over the seasonal cycle. In the vicinity of Hawaii,

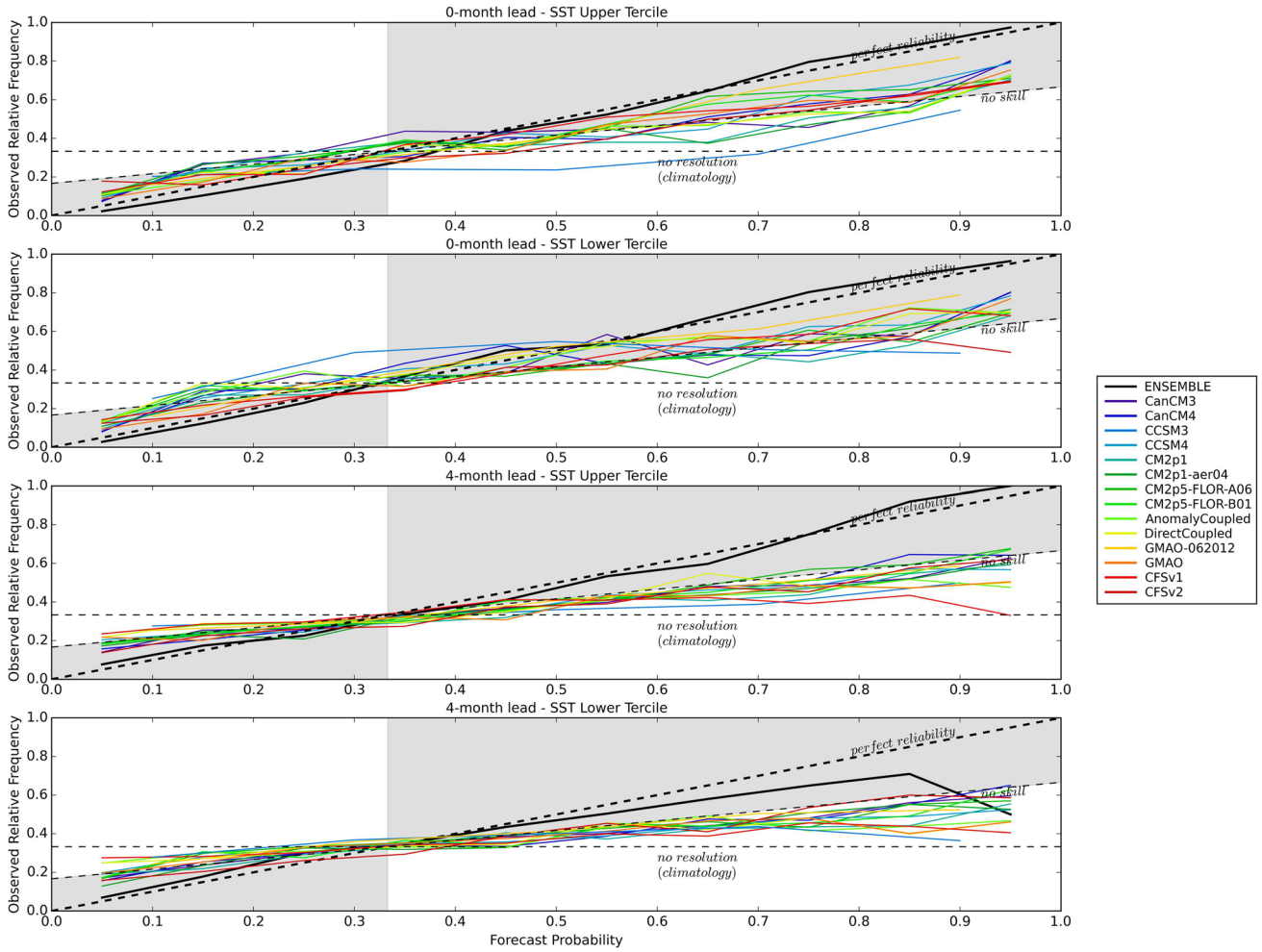

Fig. 7 Reliability diagrams for SST anomaly forecasts for the upper and lower terciles at 0-month lead (*top*, the first month forecast) and 4-month lead (*bottom*) from the individual NMME models and the ensemble mean. The values are obtained using all LMEs. The reliability diagram groups the forecasts into bins according to the issued probability (*horizontal axis*). The frequency with which the event was observed to occur for this sub-group of forecasts is then plotted against the *vertical axis*. For perfect reliability the forecast probability and the frequency of occurrence should be equal (*thick dashed line* along the diagonal). The *horizontal dash line* is the climatological forecast probability, by definition of a value of 1/3. A forecast climatology does not discriminate at all between events and non-events, and thus has no resolution. The *grey shaded area* defines the skill region (positive Brier skill score). An overconfident forecast system is where the forecast probability is greater than the observed frequency

skill has been linked to accurate simulation of the seasonal latitudinal migration of SST fronts across the North Pacific, while along the west coast of North America, skill has been linked to accurate simulation of the emergence of the signatures of Pacific basin scale variability above less predictable local variability (Stock et al. 2015; Jacox et al. 2017). In midlatitude areas with pronounced seasonal cycles in mixed layer depth, SST anomalies can recur from one winter to the next fall/winter due to the "reemergence mechanism" (e.g. Alexander and Deser 1995), where SST anomalies created by surface flux anomalies during winter persist below the seasonal thermocline in summer and are re-entrained into the surface mixed layer in the following fall and winter. In regions with seasonal ice cover, fall SST anomalies influence the winter sea ice thickness and sea-ice can serve as a reservoir for transmitting fall SST anomalies to the following summer (Blanchard-Wrigglesworth et al. 2011). In addition, sea ice melt during the spring season influences ocean temperature anomalies that persist during summer and impact sea ice and SST in the following ice growth season (Blanchard-Wrigglesworth et al. 2011). These two sea-ice mediated mechanisms may enhance SST forecast skill in the East Bering Sea and Newfoundland Labrador shelf from fall to the following spring and from spring to the following winter, resulting in higher ACC values for forecasts ending in July and December,

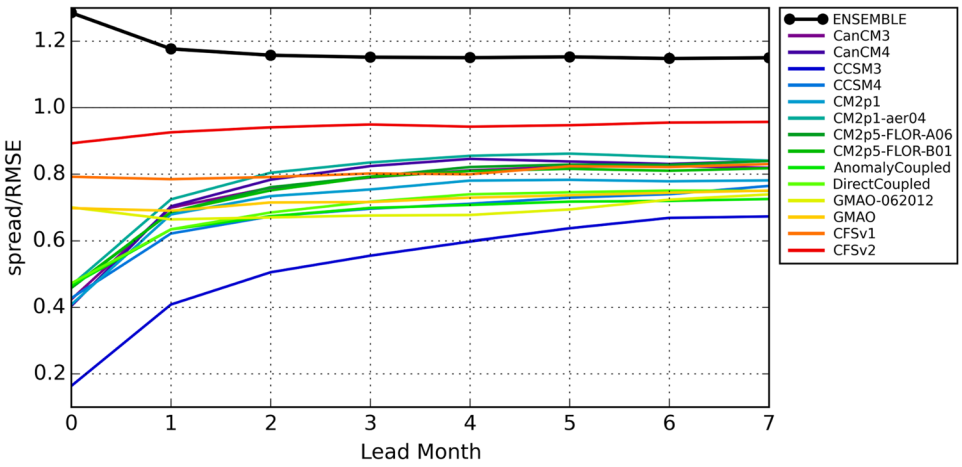

Fig. 8 SST spread normalized by the RMSE as a function of forecast lead averaged over all the LMEs. The spread is the standard deviation of all the members from an individual model or the standard deviation of all of the forecasts in the multimodel ensemble. For a reliable prediction system, the RMSE should be equal to the spread; in under-dispersive or overconfident prediction systems the spread/RMSE <1

although these processes also contribute to skill in persistence forecasts.

Errors may also be due in part to the coarse model resolution, particularly in the Northeast and Southeast US LMEs, entirely situated over the poorly resolved continental shelf. The Gulf Stream tends to extend too far north before separating from the coast in coarse resolution ocean models, potentially leading to inaccurate temperature anomaly forecasts along the eastern seaboard. Fine-scale resolution also appears to be necessary to represent flow along the continental shelf and its penetration into the Gulf of Maine and other areas with complex bathymetry (Saba et al. 2016). Fine resolution may also improve forecasts in other small regions such as the Scotian Shelf or the Gulf of California, whose SST dynamics are partially driven by small-scale processes not accurately represented in current climate prediction systems, and in regions with coastal upwelling, including the California Current System (Jacox et al. 2017).

Three older models (COLA-RSMAS-CCSM3, IRI-ECHAM4p5-AnomalyCoupled and IRI-ECHAM4p5-DirectCoupled) had poor forecast accuracy (large RMSE) during the first few forecast months due to initialization errors in the LMEs at higher latitudes. Removing those three models in the computation of the MMM led to improvement in the initial forecast accuracy, with a reduction of the MMM RMSE by up to 17%, but degraded the forecast skill at longer leads. This illustrates that the method used for the model initialization, as well as the forecast model itself, can impact seasonal forecast skill.

The ACC and RMSE indicated that on average the multimodel mean has higher deterministic skill than any single model, although an individual NMME model could provide a slightly better forecast than the MMM for a given LME, lead and month, especially in regions where forecasting was particularly challenging. Averaging over the ensemble can improve forecasts due to error cancellation, i.e., the individual model biases vary with location, time of initialization, forecast lead, etc., and when they are averaged together the error is reduced (e.g. Hagedorn et al. 2005). While including all of the individual models (even ones with more limited skill) contributed to the overall forecast skill of the NMME, more sophisticated methods such as Bayesian Model Averaging (BMA; Raftery et al. 2005), where individual models in a multimodel average are first weighted based on their skill, could improve multimodel forecasts. In practice, however, it has proven difficult to exceed the skill of the ensemble mean where all ensemble members are weighted equally (Tippett and Barnston 2008; DelSole et al. 2013).

The improvement in skill using the multimodel ensemble was especially pronounced for probability forecasts as indicated by the Brier score. Using a suite of models to assess forecast skill is relatively new with the release of the North American Multimodel Ensemble data occurring within the last 2 years (Kirtman et al. 2014). Given the chaotic nature of the climate system, very small differences in the initial condition can lead to a wide range of credible forecasts. The multimodel ensemble better represents this potential distribution of forecast values than any individual model. While increasing the number of simulations from a single model can partly ameliorate this problem, using simulations from multiple models often provides a better estimate of the actual range of predicted values (e.g. Hagedorn et al. 2005; Becker and van den Dool 2016) and thus better probability forecasts. Using a synthetic forecast generator, Weigel et al. (2008) found that multimodel ensembles outperform a 'best-model' approach if the single-model ensembles are under-dispersive, as occurred for the individual models within the NMME.

The improvement in skill for the tercile metric using the NMME may translate to improvements to the application of seasonal forecasting for marine resource management. All organisms maximize their performance (e.g., growth or

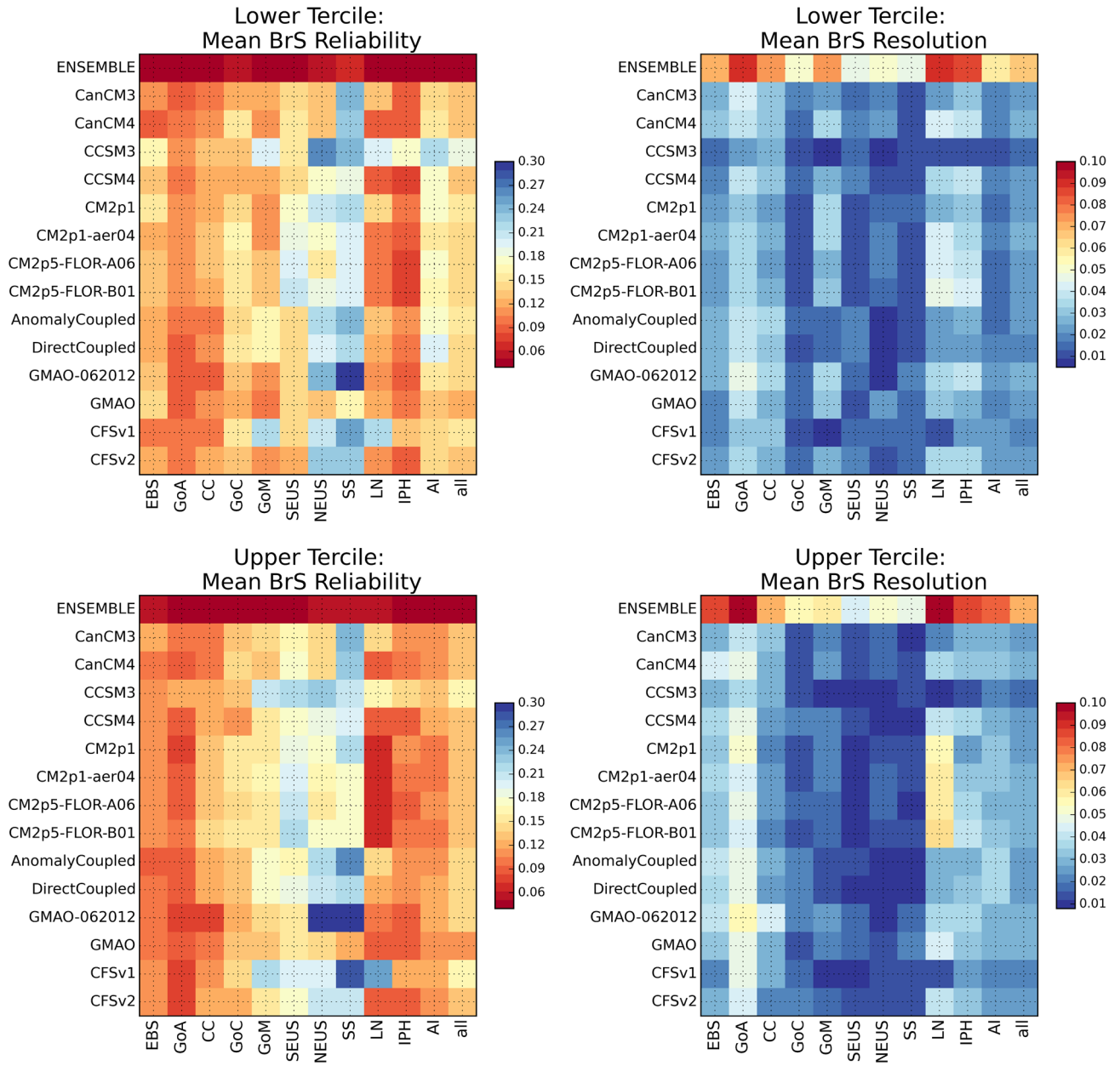

Fig. 9 The reliability (REL, *left*) and resolution (RES, *right*) components of the Brier score for SST anomaly forecasts for the lower or cold tercile (*top*) and upper or warm tercile (*bottom*) averaged over all initialization months and lead times. Results are presented for each LME and all the LMEs combined (all) (*x-axis*) and each model and the multimodel ensemble (*y-axis*). The *color scale* for all metrics is arrayed so that higher skill is shown in *red* and lower skill in *blue*

reproductive success) at an optimal range of environmental conditions. Productivity of many fish species sharply declines outside of an optimal temperature range (Pörtner and Farrell 2008) and specific life stages, such as spawners and juveniles, generally display a narrower optimal thermal window, making them particularly vulnerable to extreme temperature fluctuations (Pörtner and Peck 2010). Hence, the probability of occurrence of warm (upper tercile) or cold (lower tercile) SST anomalies during specific seasonal windows may be very useful to marine resource managers (e.g. Spillman et al. 2015). To base decisions on a comprehensive assessment of risk, managers may be advised to use multimodel rather than single-model ensemble forecasts, as the former produce a better probabilistic forecast and a more reliable estimate of uncertainty.

Acknowledgements We thank the NOAA Climate Program Office (CPO) for providing funding for this research. DT was funded by

a Special Early-Stage Exploration and Development grant from NOAA's office of oceanic and atmospheric research (OAR) with additional support from NOAA's National Marine Fisheries Service.

References

Alexander MA, Deser C (1995) A mechanism for the recurrence of midlatitude SST anomalies during winter. J Phys Oceanogr 25:122–137

Anderson DLT et al (2003) Comparison of the ECMWF seasonal forecast system 1 and 2, including the relative performance for the 1997/8 El Nino. Tech. Memo. 404. ECMWF, Reading, pp 93

Barnston AG (1994) Linear statistical short-term climate predictive skill in the Northern Hemisphere. J Clim 7:1513–1564. doi:10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2

Barth JA, Menge BA, Lubchenco J, Chan F, Bane JM, Kirincich AR, McManus MA, Nielsen KJ, Pierce SD, Washburn L (2007) Delayed upwelling alters nearshore coastal ocean ecosystems in the northern California current. Proc Natl Acad Sci 104:3719–3724

Becker E, van den Dool HM (2016) Probabilistic seasonal forecasts in the North American multi model ensemble: a baseline skill assessment. J Clim 29:3015–3026. doi:10.1175/JCLI-D-14-00862.1

Becker E, van den Dool HM, Zhang Q (2014) Predictability and forecast skill in NMME. J Clim 27:5891–5906. doi:10.1175/JCLI-D-13-00597.1

Blanchard-Wrigglesworth E, Armour KC, Bitz CM, DeWeaver E (2011) Persistence and Inherent Predictability of Arctic Sea Ice in a GCM Ensemble and Observations. J Clim 24(1):231–250

Boyer TP et al (2013) World ocean database 2013. NOAA Atlas NESDIS 72. Levitus S, Ed Mishonov A (eds) Silver Spring, pp 209. doi:10.7289/V5NZ85MT

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Buizza R (1997) Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. Mon Weather Rev 125:99–119. doi:10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2

DelSole T, Yang X, Tippett MK (2013) Is unequal weighting significantly better than equal weighting for multi-model forecasting? Q J R Meteorol Soc 139:176–183. doi:10.1002/qj.1961

Delworth TL et al (2006) GFDL's CM2 global coupled climate models. Part I: formulation and simulation characteristics. J Clim 19:644–667

Derber J, Rosati A (1989) A global oceanic data assimilation system. J Phys Oceanogr 19:1333–1347

DeWitt DG (2005) Retrospective forecasts of interannual sea surface temperature anomalies from 1982 to present using a directly coupled atmosphere–ocean general circulation model. Mon Weather Rev 133:2972–2995

Eveson JP et al (2015) Seasonal forecasting of tuna habitat in the Great Australian Bight. Fish Res 170:39–49

Goddard L, Mason SJ, Zebiak SE, Ropelewski CF, Basher R, Cane MA (2001) Current approaches to seasonal-to-interannual climate predictions. Int J Climatol 21:1111–1152

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57:219–233. doi:10.1111/j.1600-0870.2005.00103.x.

Hobday AJ, Hartog JR (2014) Derived ocean features for dynamic ocean management. Oceanography 27(4):134–145. doi:10.5670/oceanog.2014.92

Hobday AJ et al (2011) Ecological risk assessment for the effects of fishing. Fish Res 108:372–384. doi:10.1016/j.fishres.2011.01.013

Infanti JM, Kirtman BP (2016) Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. J Geophys Res: Atmos 121(21):12690–12701

Jacox MG, Alexander MA, Hervieux G, Stock CA (2017) On the skill of seasonal sea surface temperature forecasts in the California current system and its connection to ENSO variability. Clim Dyn. doi:10.1007/s00382-017-3608-y

Ji M, Behringer DW, Leetmaa A (1998) An improved coupled model for ENSO prediction and implications for ocean initialization. Part II: the coupled model. Mon Weather Rev 126:1022–1034

Jin EK et al. (2008) Current status of ENSO prediction skill in coupled ocean–atmosphere model. Clim Dyn 31(6):647–664. doi:10.1007/s00382-008-0397-3.

Jolliffe IT, Stephenson DB (2003) Forecast verification: a practitioner's guide in atmospheric science. Wiley, Chichester, p 240

Kirtman BP et al (2014) The North American multimodel ensemble phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteorol Soc 95:585–601. doi:10.1175/BAMS-D-12-00050.1

Kirtman BP, Min D (2009) Multimodel ensemble ENSO prediction with CCSM and CFS. Mon Weather Rev 137:2908–2930

Kirtman BP, Zebiak SE (1997) ENSO simulation and prediction with a hybrid coupled model. Mon Weather Rev 125:2620–2641. doi:10.1175/1520-0493(1997)125\2620:ESAPWA[2.0.CO;2

Koh T-Y, Wang S, Bhatt BC (2012) A diagnostic suite to assess NWP performance. J Geophys Res 117:D13109. doi:10.1029/2011JD017103

Latif M, Barnett TP, Cane MA, Flugel M, Graham NE, von Storch H, Xu JS, Zebiak SE (1994) A review of ENSO prediction studies. Clim Dyn 9:167–179. doi:10.1007/BF00208250.

Merryfield WJ et al (2013) The Canadian seasonal to interannual prediction system. Part I: models and initialization. Mon Weather Rev 141:2910–2945

Murphy AH (1973) A New Vector Partition of the Probability Score. J Appl Meteorol 12(4):595–600

National Research Council (2010) Assessment of intraseasonal to interannual climate prediction and predictability. The National Academies Press, Washington, DC. doi:10.17226/12878

Palmer TN et al (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bull Am Meteorol Soc 85:853–872. doi:10.1175/BAMS-85-6-853

Pörtner HO, Farrell AP (2008) Physiology and climate change. Science 322:690–692

Pörtner HO, Peck MA (2010) Climate change effects on fishes and fisheries: towards a cause-and-effect understanding. J Fish Biol 77:1745–1779

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133:1155–1174

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20:5473–5496

Rosati A, Miyakoda K, Gudgel R (1997) The impact of ocean initial conditions on ENSO forecasting with a coupled model. Mon Weather Rev 125:754–772

Saba VS, Griffes SM, Anderson WG, Winton M, Alexander MA, Delworth TL, Hare JA, Harrison MJ, Rosati A, Vechhi GA, Zhang R (2016) Enhanced warming of the Northwest Atlantic Ocean under climate change. J Geophys Res Oceans 121:118–132. doi:10.1002/2015JC011346

Saha S et al (2006) The NCEP climate forecast system. J Clim 19:3483–3517

Saha S et al (2014) The NCEP climate forecast system version 2. J Clim 27:2185–2208

Sherman K, Duda AM (1999) An ecosystem approach to global assessment and management of coastal waters. Mar Ecol Prog Ser 190:271–287

Sherman K, Belkin IM, Friedland KD, O'Reilly J, Hyde K (2009) Accelerated Warming and Emergent Trends in Fisheries Biomass Yields of the World's Large Marine Ecosystems. AMBIO: J Hum Environ 38(4):215–224

Siedlecki SA, Kaplan IC, Hermann A, Nguyen T, Bond NA, Williams G, Newton J, Peterson WT, Alin S, Feely RA (2016) Experiments with seasonal forecasts of ocean conditions for the Northern region of the California current upwelling system. Nat Sci Rep 6. doi:10.1038/srep27203

Sooraj KP, Annamalai H, Kumar A, Wang H (2012) A comprehensive assessment of CFS seasonal forecast over the tropics. Weather Forecast 27: 3–27. doi:10.1175/WAF-D-11-00014.1.

Spillman CM, Hartog JR, Hobday AJ, Hudson D (2015) Predicting environmental drivers for prawn aquaculture production to aid improved farm management. Aquaculture 447:56–65

Stock CA, Pegion K, Vecchi GA, Alexander MA, Tommasi D, Bond NA, Fratantoni PS, Gudgel RG, Kristiansen T, O'Brien TD, Xue Y, Yang X (2015) Seasonal sea surface temperature anomaly prediction for coastal ecosystems. Prog Oceanogr 137:219–236. doi:10.1016/j.pocean.2015.06.007

Stockdale TN (1997) Coupled ocean–atmosphere forecasts in the presence of climate drift. Mon Weather Rev 125:809–818

Stockdale TN, Anderson DLT, Alves JOS, Balmaseda MA (1998) Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. Nature 392:370–373

Taylor KA (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106(D7):7183–7192

Tippett MK, Barnston AG (2008) Skill of multimodel ENSO probability forecasts. Mon Weather Rev 136:3933–3946. doi:10.1175/2008MWR2431.1

van den Dool HM, Toth Z (1991) Why do forecasts for "near normal" often fail? Weather Forecast 6:76–85. doi:10.1175/15200434(1991)006<0076:WDFFNO>2.0.CO;2

Vecchi GA et al (2014) On the seasonal forecasting of regional tropical cyclone activity. J Clim 27:7994–8016

Vernieres G, Keppenne C, Rienecker MM, Jacob J, Kovach R (2012) The GEOS-ODAS, description and evaluation. NASA technical report series on global modeling and data assimilation, NASA/TM–2012–104606, vol 30

Wang B et al (2009) Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). Clim Dyn 33:93–117. doi:10.1007/s00382-008-0460-0

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260. doi:10.1002/qj.210

Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press, Dublin, p 467

Yang X et al (2012) A predictable AMO-like pattern in the GFDL fully coupled ensemble initialization and decadal forecasting system. J Clim 26:650–661