## ARTICLE

Check for updates

# Ensemble forecasting greatly expands the prediction horizon for ocean mesoscale variability

Prasad G. Thoppil [1]✉, Sergey Frolov [2], Clark D. Rowley [1], Carolyn A. Reynolds [3], Gregg A. Jacobs [1], E. Joseph Metzger [1], Patrick J. Hogan[4], Neil Barton [3], Alan J. Wallcraft[5], Ole Martin Smedstad [6] & Jay F. Shriver [1]

Mesoscale eddies dominate energetics of the ocean, modify mass, heat and freshwater transport and primary production in the upper ocean. However, the forecast skill horizon for ocean mesoscales in current operational models is shorter than 10 days: eddy-resolving ocean models, with horizontal resolution finer than 10 km in mid-latitudes, represent mesoscale dynamics, but mesoscale initial conditions are hard to constrain with available observations. Here we analyze a suite of ocean model simulations at high (1/25°) and lower (1/12.5°) resolution and compare with an ensemble of lower-resolution simulations. We show that the ensemble forecast significantly extends the predictability of the ocean mesoscales to between 20 and 40 days. We find that the lack of predictive skill in data assimilative deterministic ocean models is due to high uncertainty in the initial location and forecast of mesoscale features. Ensemble simulations account for this uncertainty and filter-out unconstrained scales. We suggest that advancements in ensemble analysis and forecasting should complement the current focus on high-resolution modeling of the ocean.

[1] Naval Research Laboratory, Ocean Sciences Division, Stennis Space Center, MS, USA. [2] University of Colorado, Cooperative Institute for Research in Environmental Sciences at the NOAA Physical Sciences Laboratory, Boulder, CO, USA. [3] Naval Research Laboratory, Marine Meteorology Division, Monterey, CA, USA. [4] NOAA, National Centers for Environment Information, Stennis Space Center, MS, USA. [5] Florida State University, Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, USA. [6] Perspecta, Stennis Space Center, MS, USA. ✉email: prasad.thoppil@nrlssc.navy.mil

The ocean mesoscales, which include eddies, fronts, meanders, and rings of the boundary currents, have horizontal spatial scales of 10–300 km and timescales of weeks to months. They develop as a result of baroclinic flow instabilities[1] and direct wind forcing. The dominant scale of the baroclinic instability driving much of this variability is determined by the Rossby deformation radius, which varies between 250 km at the equator to about 10 km at higher latitudes[2]. The recent trend of increasing model resolution, fueled by increasing computational resources, has resulted in global ocean models that directly resolve mesoscale eddies in mi-latitudes. Conversely, the regional model can now resolve submesoscale features directly without a need for parameterization[3] of their impacts on larger scales. Such eddy-resolving models are essential for producing a realistic representation of energetics of the western boundary currents such as Gulf Stream[4] and Kuroshio, their interaction with the deep ocean[5], exchange of the heat and momentum fluxes with the atmosphere[6,7], and biological productivity of the ocean[8].

Eddy resolving, non-assimilative models are insufficient for accurate prediction of evolving mesoscale eddies and fronts due to mismatches in the initial locations of eddies and associated non-linear process, resulting in large forecast errors. This source of error could be diminished by accurate initialization of the ocean dynamic state through assimilation of ocean observations that are used to constrain the mesoscale features, thereby extending the forecast horizon[9,10]. Such initialization underpins forecasting for all components of the Earth system and is based on the optimal statistical estimate of the initial state given the prior guess from the numerical model and the sparse observations of the fluid. Unlike the atmosphere, which is transparent to many modalities of the remote sensing, the ocean is opaque to the electromagnetic sampling, hence, resulting in very sparse observations of the ocean interior (vertical profiles of temperature and salinity) with the majority of the observations confined to the surface (sea surface height, salinity, temperature, and ocean currents). With the assimilation of these observations[9–14], the current generation of the forecast models has forecast skill of 10–20 days[11] as measured by the anomaly correlation of the Sea Surface Height (SSH). Unlike the atmosphere where the limit of synoptic predictability has been a focus of sustained research[15–17], similar estimates for the ocean mesoscale are lacking. However, several recent studies suggest that using empirical methods and numerical model experiments, it is possible to establish predictive skill in excess of 30 days for isolated cases of large eddies in the northern South China Sea[18,19].

The current generation of ocean forecasting models is taking two complimentary approaches to improve the quality of the ocean forecasts: (1) improving realism of the ocean models, such as increasing model resolution beyond the mesoscale[20], inclusion of tides[21], and coupling between ocean and atmosphere[22,23]; and (2) directly accounting for uncertainty in the ocean initial conditions and forecast evolution by using ensembles of model forecasts[22]. Higher resolution models, while providing a more realistic representation of ocean features at scales smaller than the mesoscale, cannot be sufficiently constrained due to sparse observations available to provide necessary coverage. As a consequence, the source of error from the unconstrained scales could lead to larger forecast errors owing to growing dynamical instabilities[24]. We show that significant improvement in the ocean forecast can be attained by accounting for uncertainty in the initial conditions. Here, we use a suite of coupled lower resolution (1/12.5°), high-resolution (1/25°), and ensemble (1/12.5°) model forecasts (see Table 1 and Methods for specification of model formulations) to examine the role of each of these model enhancements in improving the ocean forecast. The forecast error calculated using ground truth observations across a variety of parameters among the models consistently indicate that the ensemble mean exhibits superior skill over the deterministic forecasts at all lead times. We provide a quantitative analysis that shows how the ensemble mean selectively filter out scales that contribute to the forecast error.

## Results

**Comparison of forecast skill**. We define the skill of the model as its ability to have lower root mean square error (RMSE) (compared to the ground truth observations) than a monthly mean climatology of the ocean observations. Figure 1 shows that the enhancements in ocean model realism (1/25°, blue) results in statistically indistinguishable RMSE compared to the lower resolution (1/12.5°, red) forecasts. It is possible that a selection of a different error metric might highlight the differences between the two deterministic models. In contrast, better representation of uncertainties through ensemble forecasting drastically increases the skill of the forecast (black) compared to deterministic (high and lower resolution) models (see summary in Table 2).

Specifically (Table 2), both deterministic models have similar skill for predicting upper ocean temperature (11 days), upper ocean salinity (5 days), Sea Surface Height Anomaly (SSHA, 8 days), Sea Surface Temperature (SST, 21 days), mixed layer depth (11–15 days), and surface heat flux (7 days). For SST being validated against the ERA5 SST reanalysis, the high-resolution model offers a slightly less skillful forecast than the lower resolution model (12 instead of 14 days). We speculate that this slight degradation of the forecast skill for the high-resolution model can be attributed to a combination of factors, stemming from the unconstrained features that have sharper gradients and, hence, lead to large errors when these features are displaced[24]. Though not statistically significant, the presence of small-scale features with sharper gradients in the high-resolution model is sensitive to the upper-ocean mixing and thus contributed to the slightly higher RMSE in the mixed layer depth.

In contrast, the ensemble mean is consistently more skillful than the deterministic forecasts. The forecast skill of the ensemble mean increases, relative to either deterministic forecast, by almost three-fold (Table 2) for all ocean variables, with the exception of the surface heat flux (which increases by 30%). For the ocean variables, the largest skill is for comparisons with in situ SST (greater than 60 days) and the smallest skill is for upper ocean salinity (24 days). The forecast skill for the upper ocean eddies, as measured by SSHA and the water column temperature, is around 30 days. We suggest that these discrepancies in skill reflect the availability of observations to constrain the initial conditions of the system: with most observations available for the SST (~50 M daily) and the least for in situ salinity (~0.5 K daily). There are about 300 K SSHA altimeter observations daily. The lowest skill of about 9 days and the smallest skill gain from the ensemble mean is for the surface heat flux. This lower skill is a reflection that heat fluxes are impacted by both ocean and atmospheric states. Specifically, the dominant role that cloud cover plays on the incoming short-wave solar radiation and outgoing long-wave radiation. In mid-latitudes, the prediction of cloud cover is limited by the atmospheric synoptic predictability of 5–10 days and lack of formal initialization of the cloud fields. Furthermore, the skill in the surface heat flux is consistent with the predictability of 5–10 days in surface winds and air-temperature[22], which in turn contribute through fluxes of latent and sensible heat to the lowest skill in surface heat flux. Although surface heat flux is a dominant forcing for the evolution of SST, the fluxes of momentum, heat and freshwater down to the mixed layer determine the SST. The extended SST forecast skill, which benefits from the longer ocean memory in response to the

**Table 1 List of model experiments used in this study.**

| Model experiment | HYCOM/CICE | NAVGEM | Ocean tides | Model set-up |
|---|---|---|---|---|
| DetHiRes | 1/25°, 41 layers | T0681 L60 (1/5°) | Yes | Deterministic |
| DetLoRes | 1/12.5°, 41 layers | T0359 L60 (1/3°) | No | Deterministic |
| EnsLoRes | 1/12.5°, 41 layers | T0359 L60 (1/3°) | No | Ensemble (16 members) |

*HYCOM* Hybrid Coordinate Ocean Model, *CICE* Community Ice CodE, *NAVGEM* NAVy Global Environmental Model, *DetHiRes* deterministic high-resolution model, *DetLoRes* deterministic lower resolution model, *EnsLoRes* ensemble lower resolution.
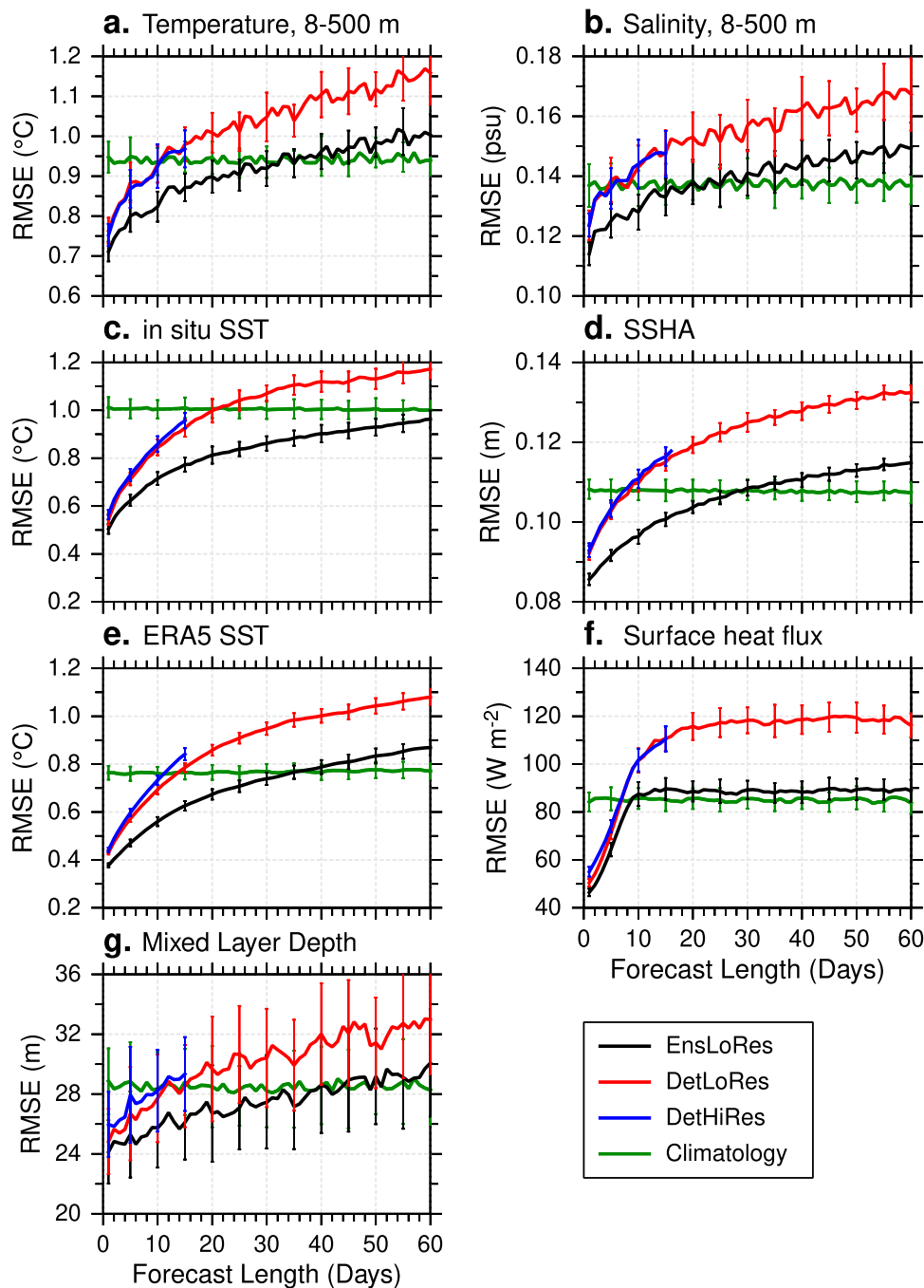


**Fig. 1 Comparison of forecast skills.** Global Root Mean Square Error (RMSE) of the ensemble mean (black), lower- (red) and high- (blue) resolution deterministic models, and monthly mean climatology (green) for (**a**) upper 500 m temperature (°C), (**b**) upper 500 m salinity (psu), (**c**) in situ SST (°C), (**d**) Sea Surface Height Anomaly (SSHA, m), (**e**) ECMWF ERA5 SST (°C), (**f**) ECMWF ERA5 surface heat flux (W m$^{-2}$), and (**g**) mixed layer depth (MLD, m), as a function of forecast length (days) from 1 to 60 days. The monthly mean climatologies include: (**a–c**, **g**) GDEM4[43], (**d**) AVISO SSHA for the period 1993–2018, (**e–f**) ECMWF ERA5 1989-2018. The verification is limited to region 50°S–50°N for the SSHA and ECMWF ERA5 and MLD fields. MLD is taken as the depth at which surface temperature decreased by 0.25 °C. Error bars are drawn at 95% confidence intervals (see methods for details).

**Table 2 Comparison of forecast skills relative to climatology. Values come from Fig. 1. A Model forecast is skillful when the RMSE is smaller than that of climatology.**

| Model Experiment | RMSE skill compared to climatology (days) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Temperature (8–500 m) | Salinity (8–500 m) | in situ SST | ERA5 SST | SSHA | ERA5 Heat Flux | MLD |
| DetHiRes | 11 | 6 | 16[+] | 12 | 8 | 7 | 11 |
| DetLoRes | 11 | 5 | 21 | 14 | 9 | 7 | 15 |
| EnsLoRes | **33** | **24** | **60[+]** | **37** | **29** | **9** | **45** |

[+]Forecasts with no climatology crossing due to forecast length cutoff.
*SST* sea surface temperature, *SSHA* sea surface height anomaly, *ERA5* ECMWF reanalysis Version 5, *MLD* mixed layer depth.
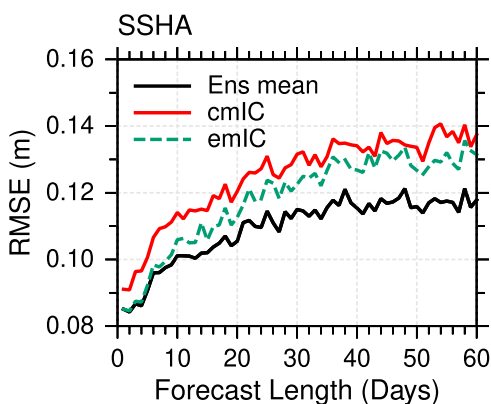


**Fig. 2 Illustration of non-linear filtering in a model experiment.**
Comparisons of global SSHA RMSE for a single forecast started on 5 July 2017 between (black) ensemble mean, (solid red) lower resolution deterministic initialized from the analysis (cmIC), and (green dashed) lower resolution deterministic initialized from the ensemble mean at the analysis time (lower resolution emIC). At short lead times the lower resolution emIC and ensemble mean RMSE match as a result of filtering out non-linear unconstraint scales. At long lead times growing dynamical instabilities in the lower resolution emIC lead to larger forecast errors similar to that in the lower resolution deterministic forecast. On the other hand, by filtering out unconstraint scales through averaging the ensemble mean maintains a consistently lower RMSE at all lead times. Although our sensitivity experiment seems to suggest that a forecast started from the ensemble mean is yielding skill that is better than the control member at shorter lead times, in practice, this is not necessarily a better approach. When we initialize from the ensemble mean, which is smoother than a single member, the features that exhibit sharp gradients are lost in the ensemble averaging. This will have a major impact on ocean mixing and stratification and it will take time for the model to recover.

atmospheric forcing, is comparable to the skill of the mixed layer depth forecasts. Furthermore, the ocean dynamic processes (e.g. upwelling) govern the SST in regimes where surface heat flux is not dominant and acts to counter the SST changes[25]. This disconnect between the surface heat flux and SST explains the superior skill in SST compared to surface heat flux. The comparison with ERA5 SST, which is relatively smoother than in situ SST, further confirms that the improvement is not only due to a smoother ensemble mean but also through filtering of the unconstrained scales in the eddy-resolving ocean.

**How do ensembles improve forecast skill?** Theory as well as the daily practice of weather forecasting show that an ensemble average of non-linear forecasts will have a comparable or lower RMSE than any individual ensemble forecast member[26,27]. The RMSE in the ensemble mean is improved relative to the control member through non-linear filtering of the less-predictable

smaller spatial scales. (For linear perturbation growth, the mean of an ensemble centered on a control member will remain equal to the control member). The impact of non-linear filtering through ensemble forecasting has been well-documented in the atmosphere[28,29] and thus ensemble weather forecasts are routinely used at major forecasting centers[26]. We illustrate the impact of non-linear filtering by starting a deterministic forecast from an ensemble mean analysis, referred to as the control forecast or emIC (Fig. 2). As expected, the RMSE score of the control forecast is similar to that of the ensemble mean over the first 5–7 days. However, at longer lead times, the dynamic instabilities in the control forecast grow, leading to deteriorating RMSE scores. Ultimately, the error of the control forecast converges to the error of the other deterministic forecasts (cmIC).

To distinguish between the effect of non-linear filtering in the ensemble mean and the fact that the ensemble mean is smoother and, hence, may verify better against observations in an RMSE sense, we use spectral analysis of the forecast error $R^2$ statistics. We define $R^2$ as the ratio of the model error variance to the variance of the altimeter SSHA (see methods section for details of the $R^2$ computation). The $R^2 > 1$ indicates that the model lacks skill in a specific spectral band because the variance of the model error exceeds the natural variability of the SSHA. The $R^2 < 1$ indicate a skillful forecast, that is, the model error is smaller than the natural variability. The $R^2 = 1$ indicates that the variance of the model error is equal to the variance of the signal in the spectral band, for example, when we use climatology as a forecast. The $R^2 = 2$ can occur when the model and the observations have equal variance but the signals are perfectly out of phase.

At 1-day forecast (Fig. 3a) both deterministic forecasts lack skill for scales smaller than about 300 km. In contrast, the ensemble mean has skill at all scales, with significant skill at scales larger than 150 km. The ensemble mean is also more skillful than a deterministic forecast at large scales (up to the maximum of 1000 km shown in Fig. 3). The lack of deterministic skill is consistent with the recent study that found that assimilation of ocean observations provides no constraint on ocean eddy kinetic energy at scales smaller than 161 km in the Kuroshio region[20]. Recall that the eddy kinetic energy is a derivative of the altimetry SSHA and, hence, we can expect that constrained kinetic energy spectra would have different scales than the SSHA.

Figure 4 further illustrates the spectral analysis results presented in Fig. 3 using a single snapshot of the SSH from 1-day forecast in the Kuroshio extension region of the North West Pacific. The forecasts agree on the representation of large-scale warm and cold core eddies (as indicated by highs and lows in SSH) and fronts in the Kuroshio extension region. Conversely, the models disagree at the exact location of smaller eddies (e.g. eddy around 162°E and 30°N highlighted by the black dashed line box). It is these unconstrained features at the analysis time that are leading to large forecast errors owing to growing dynamical instabilities. The error growth in the high-resolution model is even more rapid as indicated by slight degradation of skill in SST
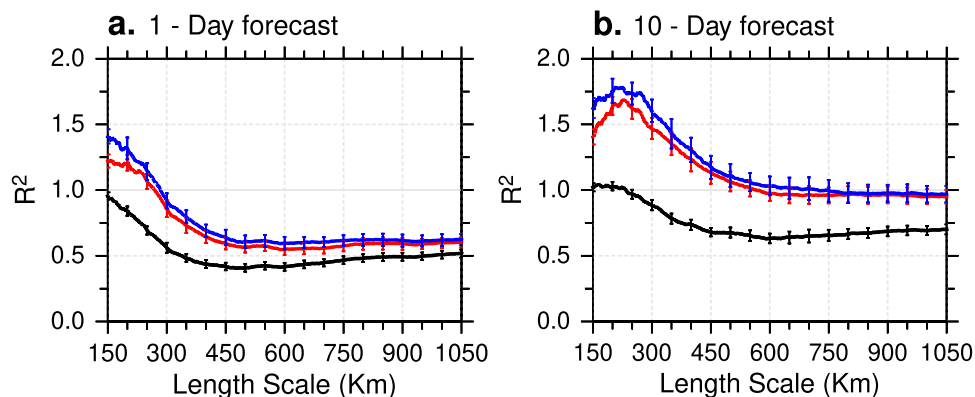
**Fig. 3 Improved forecast skill in the ensemble mean as revealed by power spectral analysis of along-track SSHA.** Scales constrained in the (**a**) 1 day forecast and (**b**) 10 day forecast in the high-resolution model (blue), lower resolution (red), and ensemble mean (black). $R^2$, which is the ratio of the model error variance to the variance of the altimeter SSHA, values above 1 mean that variance of error in this frequency band is greater than variance of the altimeter signal and, hence, these scales are not constrained. $R^2$ values below 1 indicate that variance of errors are lower than the variance of the signal and, hence, analysis and forecast have skill in this frequency band. The $R^2$ values are spectrally smoothed and averaged over multiple forecasts (see details in the methods section). We excluded scales lower than 150 km as the power spectra density flattens off at this scale. The analysis is carried out for the globe (50°S to 50°N). The error bars of the $R^2$ are computed at 95% confidence level (see methods for details).
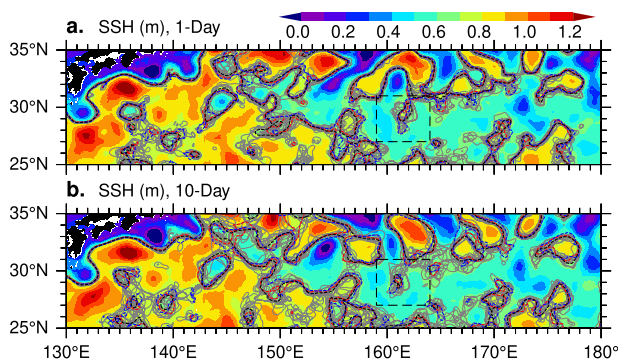


**Fig. 4 Models depiction of mesoscale features.** Snapshot of model Sea Surface Height (SSH, m) forecast initialized from the analysis on 1 March 2017 for the Kuroshio Extension region (southeast of Japan) for the (**a**) 1-day (2 March 2017), and (**b**) 10- day (11 March 2017) forecasts. Shaded regions indicate ensemble mean SSH. A contour of the 0.7 m SSH is shown for ensemble mean (black), high- (blue) and lower- (red) resolution models to demarcate the mesoscale features. Gray contours indicate spaghetti isolines of 16 ensemble members 0.7 m SSH. Regions of high (low) SSH indicate warm (cold) core mesoscale eddies. The spread of the ensemble members indicates forecast uncertainty in the location of mesoscale features, which is causing larger forecast errors in the deterministic forecasts. The models agree on the large-scale mesoscale features and fronts. However, the exact location of the small-scale features differs among the models (highlighted by dashed black box). It is the fast-growing instabilities on these scales that are responsible for forecast errors.

(Fig. 1e) as the error from smaller scales feed onto the underlying mesoscales. This is consistent with other reported findings that increasing model resolution towards the submesoscale leads to a degradation in the forecasting of mesoscale features in a regional modeling study[24].

Figures 3b, 4b show that our findings for the 1-day forecast are also valid at a longer forecast time (we used 10 day forecast here as an example). As expected, the $R^2$ for the deterministic forecast increases across all scales, indicating degraded skill of the forecast compared to the 1-day forecast (Figs. 3a, 4a). Furthermore, in agreement with Fig. 1, the $R^2$ for deterministic forecasts exceeds 1

at all scales indicating lack of skill (recall that from Fig. 1d the skill of the deterministic forecast for SSHA is less than 9 days). By contrast, the $R^2$ ratio of the ensemble mean is below 1 at a wide range of scales, particularly scales larger than 200 km, consistent with Fig. 1d. The fact that these independent analyses produced similar results provides confidence in our results.

Qualitative comparisons with SSHA observations shown in Fig. 5 illustrate the extended predictability of the mesoscale features in the ensemble mean. At 1-day forecast (Fig. 5a), the characteristics (shape and position) of mesoscale features (cold and warm core eddies) are in good agreement with the along-track SSHA observations owing to the assimilation of observations at the analysis time. The evolution of the mesoscale circulation becomes less skillful at 10-day forecasts (Fig. 5b) due to growing dynamical instabilities. This can be seen around 155°E (see dashed box) where an elongated cyclonic circulation with a warm core eddy to the north in the 1-day forecast becomes baroclinically unstable, generating two cold core eddies by eddy separation. However, the observations have maintained its shape and position closer to March 2, 2017. The eddy separation may also be the result of interaction with the warm core eddy to the north. In the 30-day forecast (Fig. 5c), the mesoscale features in the ensemble mean do not match well with the along-track SSHA. While there is reasonable agreement for the presence of mesoscale features in the observations, it is evident that their positions are displaced. This is the case for a cold core eddy at 132°E (see dashed box) where it has displaced slightly to the northeast compared to the observations.

## Discussion

The growth in computational power for ocean modeling has enabled increased horizontal resolution beyond mesoscales. We find that current deterministic data assimilation only constrains SSHA scales that are larger than 300 km, which far exceeds scales that can be numerically resolved by a modern ocean model. For example, the high-resolution (1/25°) model used in this study can numerically resolve and propagate features as small as five-to-ten grid cells or about 20–40 km. As a result, both deterministic models constrained similar scales despite having differences in the model resolutions, thereby producing comparable skill. Hence, we attribute the lack of skill improvements in the high-resolution model to the lack of high-density observations that are
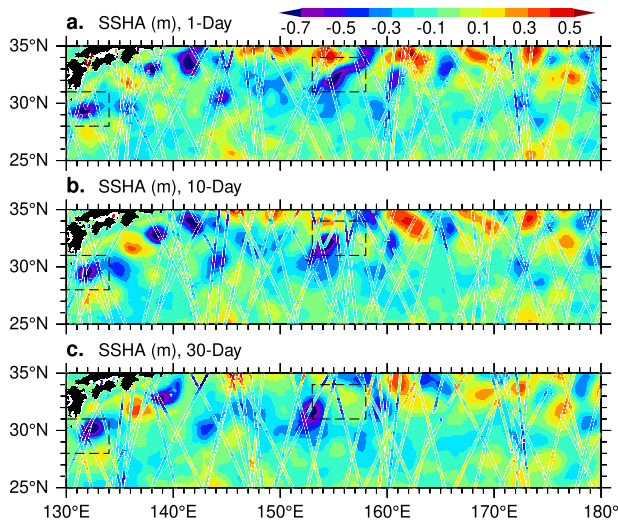
**Fig. 5 Qualitative comparisons with SSHA observations.** Model ensemble mean sea surface height anomaly (SSHA, m) of the Kuroshio Extension region (southeast of Japan) on the forecast days (**a**) 1 (2 March 2017), (**b**) 10 (11 March 2017) and (**c**) 30 (31 March 2017) for the forecasts started from the 1 March 2017 analysis. The shaded regions indicate the ensemble mean SSHA. The along-track altimetry SSHA observations are shown as filled squares with white edges using the same color bar. Because a single day along-track SSHA observations have a limited spatial coverage on depicting mesoscale eddies, we combined three days along-track observations (e.g. along-track observations from 9 to 11 March 2017 are used for forecast verification on 11 March 2017) with the assumption that eddy movement over three days is relatively small. Regions of high (low) SSHA indicate warm (cold) core mesoscale eddies. The location of two cold core eddies (cyclonic) is highlighted by dashed black boxes.

needed to constrain the model. Future high-density observations mission such as Surface Water Ocean Topography (SWOT)[30] could benefit high-resolution models in constraining scales smaller than that in the current ocean observations system.

The results of this study suggest that the continuous increase in resolution and realism of the ocean models has limited potential to further improving the forecast skill of eddy-resolving models without reducing the uncertainty of the model initial conditions. One way to reduce these uncertainties is to increase the density of the ocean sampling network (such as the upcoming SWOT[30] and the planned Winds and Currents Mission (WaCM)[31] satellite missions). The scales constrained by the data assimilation are specific to the density of observations, assuming model resolution can resolve those scales in the observations. Recent studies[20,32] suggest that the addition of the SWOT observations will translate to modest improvement in constrained scales in the ocean forecast (reduction from 161 km to 139 km for ocean eddy kinetic energy). It is likely that the proposed WaCM mission to observe surface currents will be able to further constrain global ocean mesoscale features. Conversely, it is possible to improve the realism of the resolved scales in the high-resolution models by improved ocean model physics and parameterization of sub-grid scale processes.

As an alternative, but also in addition to the increase in ocean satellite observing missions, this article demonstrates that ensemble forecasting methods can greatly reduce error in both the ocean analysis and forecast. Specifically, at the 1-day forecast, the ensemble mean analysis presented here had skill for the entire range of examined scales (from 150 to 1000 km), while the deterministic analysis had no skill below 300 km. The same

benefit of non-linear filtering extends to the forecast, which translates to a three-fold gain in the skillful forecast range over climatology (ensemble mean is more skillful than climatology up to 29 days for SSHA and over 60 days as verified against in situ SST).

Given the results presented in this paper, what is the benefit of further investments into higher resolution forecast models? We hypothesize that the ensemble of such high-resolution models (once it becomes practicable with an increase in computational resources) is likely to be more skillful than the ensemble of lower resolution models that we investigated in this paper. Such increased skill is likely to result from (a) a more realistic representation of the natural variability of the ocean, and (b) better representation of the energy cascade in the high-resolution model.

Ensemble forecast systems similar to the one presented here can be developed readily and at a modest computational cost (the cost of the 16-member lower resolution ensemble is roughly twice the cost of the high-resolution ocean model used in this study). The addition of the ensemble methodology to the array of emerging ocean forecasting applications can have positive benefits for multiple endeavors, including improvements in ocean reanalyses and improved forecasting of ocean ecosystems and harmful algal blooms and storm surge, and other application such as safety of navigation in the Arctic, search and rescue at sea, and extended range weather forecasting.

## Methods

**Models**. A set of coupled global models are used. A summary of the model configurations is presented in Table 1. The model components and their configurations are based on existing Navy forecast systems. The ocean model is the HYbrid Coordinate Ocean Model (HYCOM)[11,33] with sea ice—Community Ice CodE (CICE)[34]; and the atmospheric model is the NAVy Global Environmental Model (NAVGEM)[35]. The models are fully coupled using National Unified Operational Prediction Capability (NUOPC)[36] coupler. The high-resolution model is configured at 1/25° (~4.4 km in the tropics) and lower resolution is 1/12.5° (~9 km in the tropics). All models have 41 vertical layers. Vertical mixing is parameterized using the KPP scheme. These experiments utilize the same settings apart from differences in resolution except that high-resolution deterministic model includes tidal forcing. Because the models are fully coupled, the forcing can be slightly different between the models due to different horizontal resolutions (see Table 1) which is not likely to impact the forecast skills. The family of coupled models, referred as Navy-ESPC[22], were undergoing transition from research to operations at the U.S. Fleet Numerical Meteorological and Oceanographic Center (FNMOC) at the time these experiments were conducted.

**Data assimilation**. Atmospheric observational data are assimilated using a 6-hour update cycle in the NRL Atmospheric Variational Data Assimilation System—Accelerated Representer (NAVDAS-AR)[37]. A three-Dimensional Variational Data Assimilation (3DVAR) method is used for the assimilation of ocean and sea ice observations using a 24-hour update cycle and the Navy Coupled Ocean Data Assimilation software (NCODA)[38,39] to constrain the mesoscale features to the available observations. A weakly coupled data assimilation system is used for the Navy-ESPC model[22], which is defined by separate DA systems for the ocean (NCODA) and atmosphere (NAVDAS-AR) and a fully coupled forecast model for the first guess (Navy-ESPC). A 3-h incremental analysis update[40] is used to synchronize the initial conditions in the ocean and the atmosphere and to minimize the impact of the dynamical imbalances in the analysis increments. The sea ice analysis is directly inserted at the start of each 12Z forecast cycle.

**Ensemble generation**. Ensemble experiments with the Navy-ESPC are configured as a 16-member ensemble. Ensemble generation used the perturbed observation approach, in which observations are perturbed separately in NAVDAS-AR and NCODA scaled by the observational error. In the present configuration, 15 members add perturbations to the observations and one "unperturbed" member does not. Note, that this unperturbed member is still part of a random draw from the unknown true distribution of errors. Each ensemble member maintained an independent forecast-assimilation cycle in which perturbations are introduced through the observations assimilated by adding random perturbations to the observations. The perturbations are scaled with the assumed observation error of each observation (i.e., random draws from the normal distribution with zero mean and the standard deviation of the observation error). The introduction of the random perturbations

produces differences in the analyses, in turn, causing differences in the forecasts. In NAVDAS, observations are perturbed in the NAVDAS-AR 4DVAR solver after they are thinned and quality controlled. Statistics of the perturbations are consistent with the observation error statistics used by the 4DVAR system. In NCODA, satellite observations of SSHA, SST, and sea ice concentration and in situ surface observations are perturbed before thinning, synthetic profiles are generated using perturbed SSHA and SST predictors, and in situ profile observations are perturbed individually independent of the profile thinning. Due to our implementation constraints we do not enforce the requirement that the ensemble perturbations average-out to zero at initial time.

As this is a fully coupled ocean-atmosphere-sea ice system, perturbations introduced in the atmospheric variables (e.g. winds), in turn, feedback into the ocean thereby serving as a source of uncertainty to the ocean forecasts. The model uncertainties, those caused by approximations in model formulations and physical parameterizations, are not accounted for in the present configuration. While considerable progress has been made to account for structural model uncertainties in atmospheric forecast models[41], progress for ocean models is still in the early stages[42]. This is due to the lack of a clear understanding of sources of error and how to represent them in ocean forecasting models. One aspect of this research that separates itself from earlier studies is that we have presented results from a global, fully coupled ocean-atmosphere-ice global ensemble prediction system and compares the ensemble skill relative to a high- and lower- resolution deterministic systems, out to 15–60 days.

**Analysis-forecast cycle.** The analysis for the coupled ensemble started on 15 December 2016. All 16 members started on this date using the same restart files and ran until 24 January 2018. After a 45-day spin-up period in order to give the parallel update cycles time to diverge, 60-day forecasts were performed once per week on Wednesdays at 12Z consisting a total of 52 forecasts spanning 1 February 2017 and 24 January 2018. A similar set-up is used for high-resolution coupled run, except that the length of forecast is limited to 16 days. This set-up enables comparisons of the skill between the ensemble and deterministic forecasts.

**Observations.** To evaluate the skill of the forecasts, a variety of ocean observations were used. We used observations during forecast verification time, hence, the specific observations used for verification have not been yet assimilated by our system and can be considered independent as is a common practice for the verification of atmospheric and oceanic forecasts. However, because of the paucity of observational data, this is the same type (Argo profiles, SST observations, etc.) of the observation as we use in assimilation. This is, possibly, with the exception of the ERA5 SST and heat flux, which come from an external reanalysis product.

During the model forecast runs, daily-mean temperature and salinity fields were calculated in HYCOM. For each file, a corresponding 24-hour window of verifying profile or surface observations was stored in a matchup file together with the model prediction of the temperature and salinity at the observation location, and on the standard 40 depth levels. The profiles or surface observations are retrieved from the NCODA database include in situ temperature and salinity from Argo and Conductivity Temperature Depth (CTD) among other sources. Profiles are quality controlled for outliers against climatology and profiles with larger than three standard deviations are excluded from the analysis. The primary reference comparison for the ocean forecast is climatology, because there is presently no extended-range (more than 7 days) global ocean forecast system. The climatology used for ocean temperature and salinity is the Generalized Digital Environmental Model[43] (GDEM4).

For the SST forecast skill evaluation, two sets of SST are used. The in situ based SST observations made primarily by drifting buoys and ship of opportunity among other sources are quality controlled in the NCODA database. Matchup files consisting of model values at observation location and the observations of SST ($\sim$30 K day$^{-1}$) are used to calculate the forecast error. The observations which are mostly confined along the ship tracks have non-uniform spatial coverage. The reference climatology is taken from the GDEM4[43]. In situ SST complimented by more spatially uniform SST from the ERA5 reanalysis. The ERA5 reanalysis, which is on spatial resolution of 0.25° × 0.25° grid, is sub-sampled by 1° × 1° grid so that they can be treated as representative observations rather than correlated values from adjacent grid points. Unlike in situ SST, the analysis region is near-global and limited between 50°S and 50°N. For climatology, the monthly mean ERA5 reanalysis for the period 1979–2018 is used. The surface heat flux represents the balance between the radiative (solar and long-wave radiation) and turbulent heat fluxes (sensible and latent heat fluxes) and primarily governs the evolution of SST through one-dimensional mixed layer thermodynamic processes among other ocean dynamic processes. The forecast skill of the surface heat flux is evaluated against ERA5 reanalysis and climatology for the period 1979–2018.

The Sea Surface Height Anomaly (SSHA) measured by remotely sensed satellite altimetry provides a quantitative measure of mesoscale eddies in the ocean. The along-track SSHA observations are taken from the NCODA ocean observation system. The SSHA is calculated using the reference mean SSH from observations averaged over the period 1993–2008. For comparing with altimetry SSHA, the 20-year (1993–2012) mean SSH from an ocean reanalysis[44] is subtracted from the model SSH to obtain model SSHA. Model SSHA is extracted along the satellite track using a linear interpolation. The skill of model forecasts is compared relative

to a climatology generated using a 26-year (1993–2018) monthly mean SSHA mapped on a spatial resolution of 0.25° × 0.25° from AVISO (Archiving, Validation and Interpretation of Satellite Oceanographic data).

**Error quantification.** The performance of forecast models is quantified in terms of RMSE using all available independent observations. Although RMSE is dominated by observations with large errors, the averaging across 52 forecasts minimize this and accurately represents the performance of the forecasting systems. The sampling between high- and lower resolution models is treated the same as most observations are coarser than model grid points. The statistics presented are global error and should be comparable between the models.

While RMSE penalizes the sharp features with location error in the high-resolution model, the so-called the double-penalty effect[45,46], the spectral analysis, which independently confirms the error analysis, is not prone to the double-penalty effect for the scales considered (>150 km). We also note that the skill improvements in the ensemble mean are significantly greater than the deterministic models; hence, it is highly unlikely that the double-penalty effect, which is more prevalent at shorter lead times, could affect the major conclusions of this study.

**Spectral analysis of error.** The spectral analysis of error presented in Fig. 3 has been computed using the following steps: (1) Matchups between altimeter SSHA and deterministic model runs and ensemble mean are computed (matchups file consists of observed SSHA, time, geographical coordinates, and model SSHA at the observation locations) (2) Latitude-longitude position of the matchups are converted into the along-track coordinates (3) Matlab© implementation of the Lomb-Scargle method is used to compute power spectral density (PSD) of the raw along-track altimeter signal (#2) and the model-altimeter matchups (#1) (4) PSD are smoothed using a moving average filter in the frequency space that used the sliding window of the nearest ±1000 frequencies, which, in the range of interest between 50 and 1000 km, averaged frequencies within ±2% of the central frequency. For example, the resulting smoothed estimate for the frequency centered on 1/300 km$^{-1}$ is averaged between 1/294 km$^{-1}$ and 1/306 km$^{-1}$ (5) Smoothed PSD are interpolated on a common set of 5000 frequencies, which are logarithmically spaced between 20 and 2000 km (6) The $R^2$ statistics are computed by taking a ratio of the smoothed PSD of the errors to the altimeter signal.

We examined the slope of the power spectral density of the SSHA signal (Supplementary Fig. 1), which revealed that the slope flattens for wavelengths below 150 km, indicating that below 150 km SSHA observations are contaminated by an increasing amount of observational noise. As a result, we excluded scales smaller than 150 km from the forecast error $R^2$ with no apparent implications on the major conclusions drawn from this study.

**Error bars.** Error bars of the estimated RMSE (Fig. 1) and $R^2$ statistics (Fig. 3) are computed at 95% confidence level as Error = $1.96\times \text{std}(q)/\sqrt{N/4}$, where $q$ is a quantity of interest (the RMSE or the $R^2$) and $N/4$ is the number of degrees of freedom in the sample. Specifically, $N$ is the number of forecasts (52) and 4 is an empirical factor that accounted for possible correlation between weekly forecasts.

## Data availability
Data displayed in the figures are publicly available from https://github.com/frolovsa/papers-naturecee2021. The observations being used for forecast evaluation are publicly available for download from these websites https://ioos.noaa.gov/data/access-ioos-data/; https://www7320.nrlssc.navy.mil/altimetry/. The monthly mean SSHA climatology used in this study were processed by SSALTO/DUCAS and distributed by AVISO+ with support from CNES and available for download from website https://www.aviso.altimetry.fr. The ECMWF ERA5 reanalysis products being used here are obtained from https://cds.climate.copernicus.eu/. The GDEM climatology is the product of U. S. DoD and available upon submitting product request form to the Naval Meteorology and Oceanography Command Secretary of the OAML. The Navy-ESPC analysis and forecast model output data are stored at the Navy DoD Supercomputing Resource Center (DSRC). Access to the Navy DSRC may be obtained through a request to the DoD HPCMP (https://www.hpc.mil/). Once an account has been established, contact the corresponding author for information to access the archived data.

## Code availability
The ocean model is available at https://www.hycom.org/. The data assimilation and observation processing scripts are the intellectual property of the Department of Navy. The Navy-ESPC code is only available with pre-approval from sponsors. Collaborators outside of NRL with model access have agreements and grants with sponsors (i.e., Office of Naval Research (ONR)). If interested in working with the Navy-ESPC code, please contact the corresponding author and code 32 of ONR (https://www.onr.navy.mil/Science-Technology/Departments/Code-32).

## References

1.  Chelton, D. B., Schlax, M. G. & Samelson, R. M. Global observations of nonlinear mesoscale eddies. *Prog. Oceanogr.* **91**, 167–216 (2011).
2.  Chelton, D. B., deSzoeke, R. A., Schlax, M. G., El Naggar, K. & Siwertz, N. Geographical variability of the first baroclinic rossby radius of deformation. *J. Phys. Oceanogr.* **28**, 433–460 (1998).
3.  Hurlburt, H. E. The potential for ocean prediction and the role of altimeter data. *Mar. Geod.* **8**, 17–66 (1984).
4.  Chassignet, E. P. & Xu, X. Impact of horizontal resolution (1/12° to 1/50°) on Gulf Stream separation, penetration, and variability. *J. Phys. Oceanogr.* **47**, 1999–2021 (2017).
5.  Thoppil, P. G., Richman, J. G. & Hogan, P. J. Energetics of a global ocean circulation model compared to observations. *Geophys. Res. Lett.* **38**, L15607 (2011).
6.  Kirtman, B. P. et al. Impact of ocean model resolution on CCSM climate simulations. *Clim. Dyn.* **39**, 1303–1328 (2012).
7.  Small, R. J. et al. Air-sea interaction over ocean fronts and eddies. *Dyn. Atmos. Ocean.* **45**, 274–319 (2008).
8.  Chelton, D. B., Gaube, P., Schlax, M. G., Early, J. J. & Samelson, R. M. The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science* **334**, 328–332 (2011).
9.  Metzger, E. J. et al. *Validation Test Report for the Global Ocean Forecast System V3.0-1/12 deg HYCOM-NCODA: Phase II.* https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/ADA518693.xhtml (2010).
10. Smith, S. et al. *Validation Test Report for the Navy Coupled Ocean Data Assimilation 3D Variational Analysis (NCODA-VAR) System, Version 3.43.* NRL Report NRL/MR/7320-11-9363 (Naval Research Laboratory, 2011).
11. Metzger, E. J. et al. US Navy operational global ocean and Arctic ice prediction systems. *Oceanography* **27**, 32–43 (2014).
12. Hurlburt, H. E. et al. in *Ocean Modeling in an Eddying Regime, Geophysical Monograph Series* Vol. 177 (eds Hecht, M. W. & Hasumi, H.) 353–382 (Wiley, 2008).
13. Drévillon, M. et al. The GODAE/Mercator-Ocean global ocean forecasting system: results, applications and prospects. *J. Oper. Oceanogr.* **1**, 51–57 (2008).
14. Martin, M. J., Hines, A. & Bell, M. J. Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. *Q. J. R. Meteorol. Soc.* **133**, 981–995 (2007).
15. Lorenz, E. N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963).
16. Lorenz, E. N. The predictability of a flow which possesses many scales of motion. *Tellus* **21**, 289–307 (1969).
17. Zhang, F. et al. What is the predictability limit of midlatitude weather? *J. Atmos. Sci.* **76**, 1077–1091 (2019).
18. Li, J., Wang, G., Xue, H. & Wang, H. A simple predictive model for the eddy propagation trajectory in the northern South China Sea. *Ocean Sci* **15**, 401–412 (2019).
19. Xu, D., Zhuang, W. & Yan, Y. Could the two anticyclonic eddies during winter 2003/2004 be reproduced and predicted in the northern South China Sea? *Ocean Sci.* **15**, 97–111 (2019).
20. D'Addezio, J. M. et al. Quantifying wavelengths constrained by simulated SWOT observations in a submesoscale resolving ocean analysis/forecasting system. *Ocean Model.* **135**, 40–55 (2019).
21. Arbic, B. K., Wallcraft, A. J. & Metzger, E. J. Concurrent simulation of the eddying general circulation and tides in a global ocean model. *Ocean Model* **32**, 175–187 (2010).
22. Barton, N. et al. The Navy's Earth System Prediction Capability: A new global coupled atmosphere-ocean-sea ice prediction system designed for daily to subseasonal forecasting. *Earth Space Sci.* **8**, e2020EA001199 (2021).
23. Grist, J. P. et al. Increasing Atlantic Ocean heat transport in the latest generation coupled ocean-atmosphere models: the role of air-sea interaction. *J. Geophys. Res. Ocean.* **123**, 8624–8637 (2018).
24. Sandery, P. A. & Sakov, P. Ocean forecasting of mesoscale features can deteriorate by increasing model resolution towards the submesoscale. *Nat. Commun.* **8**, 1566 (2017).
25. Yu, L., Jin, X. & Weller, R. A. Role of net surface heat flux in seasonal variations of sea surface temperature in the tropical Atlantic Ocean. *J. Clim.* **19**, 6153–6169 (2006).
26. Palmer, T. The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Q. J. R. Meteorol. Soc.* **145**, 12–24 (2019).
27. Palmer, T. N. Stochastic weather and climate models. *Nat. Rev. Phys.* **1**, 463–471 (2019).
28. Palmer, T. N. Medium and extended range predictability and stability of the Pacific/North American mode. *Q. J. R. Meteorol. Soc.* **114**, 691–713 (1988).
29. Surcel, M., Zawadzki, I. & Yau, M. K. On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Weather Rev.* **142**, 1093–1105 (2014).
30. Fu, L.-L. & Ubelmann, C. On the transition from profile altimeter to swath altimeter for observing global ocean surface topography. *J. Atmos. Ocean. Technol.* **31**, 560–568 (2014).
31. Chelton, D. B. et al. Prospects for future satellite estimation of small-scale variability of ocean surface velocity and vorticity. *Prog. Oceanogr.* **173**, 256–350 (2019).
32. Carrier, M. J., Ngodock, H. E., Smith, S., Souopgui, I. & Bartels, B. Examining the potential impact of SWOT Observations in an Ocean Analysis–Forecasting System. *Mon. Weather Rev.* **144**, 3767–3782 (2016).
33. Bleck, R. An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. *Ocean Model* **4**, 55–88 (2002).
34. Hunke, E. C. & Lipscomb, W. H. *CICE: The Los Alamos Sea Ice Model Documentation and Software User's Manual Version 4.0.* Technical Report LA-CC-06-012, T-3 Fluid Dynamics Group (Los Alamos National Laboratory, 2015).
35. Hogan, T. F. et al. The Navy Global Environmental Model. *Oceanography* **27**, 116–125 (2014).
36. Theurich, G. et al. The earth system prediction suite: toward a coordinated U. S. modeling capability. *Bull. Am. Meteorol. Soc.* **97**, 1229–1247 (2016).
37. Rosmond, T. & Xu, L. Development of NAVDAS-AR: non-linear formulation and outer loop tests. *Tellus A* **56**, 45–58 (2006).
38. Cummings, J. A. & Smedstad, O. M. in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* Vol II (eds Park, S. & Xu, L.) 303–343 (Springer, 2013). https://doi.org/10.1007/978-3-642-35088-7_13.
39. Cummings, J. A. Operational multivariate ocean data assimilation. *Q. J. R. Meteorol. Soc.* **131**, 3583–3604 (2005).
40. Bloom, S. C., Takacs, L. L., Da Silva, A. M. & Ledvina, D. Data assimilation using incremental analysis updates. *Mon. Weather Rev.* **124**, 1256–1271 (1996).
41. Leutbecher, M. et al. Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Q. J. R. Meteorol. Soc.* **143**, 2315–2339 (2017).
42. Andrejczuk, M. et al. Oceanic stochastic parameterizations in a seasonal forecast system. *Mon. Weather Rev.* **144**, 1867–1875 (2016).
43. Carnes, M. R., Helber, R. W., Barron, C. N. & Dastugue, J. M. *Validation Test Report for GDEM4.* NRL Tech Report NRL/MR/7330--10-9271 (Naval Research Laboratory, 2010).
44. Thoppil, P. G., Metzger, E. J., Hurlburt, H. E., Smedstad, O. M. & Ichikawa, H. The current system east of the Ryukyu Islands as revealed by a global ocean reanalysis. *Prog. Oceanogr.* **141**, 239–258 (2016).
45. Rossa, A., Nurmi, P. & Ebert, E. in *Precipitation: Advances in Measurement, Estimation and Prediction* (ed Michaelides, S.) (Springer, 2008). https://doi.org/10.1007/978-3-540-77655-0_16.
46. Crocker, R., Maksymczuk, J., Mittermaier, M., Tonani, M. & Pequignet, C. An approach to the verification of high-resolution ocean models using spatial methods. *Ocean Sci.* **16**, 831–845 (2020).

## Author contributions

P.G.T. conceived the study, performed data processing and forecast error analysis, and wrote the first draft of the manuscript. S.F. computed spectral analysis, interpreted the results, and contributed substantially to the writing of the manuscript. C.A.R. designed sensitivity experiment and P.G.T. performed it. G.A.J. helped clarifying spectral analysis results and mixed layer evaluation. A.J.W. contributed to HYCOM developments. C.D.R. and N.B. aided in Navy-ESPC ocean and atmospheric ensembles developments. P.J.H., S. F., and N.B. performed ensemble simulations and O.M.S., E.J.M., and J.F.S. performed deterministic model simulations. All authors revised, read, and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43247-021-00151-5.

**Correspondence** and requests for materials should be addressed to P.G.T.

**Peer review information** Primary handling editor: Heike Langenberg.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.