

Article

Beyond Inflation: Backscatter Parameterizations to Address the Variability Deficit in Global Ocean Data Assimilation

Kate Boden ^{1,*}, Daniel E. Amrhein ^{2,3}, Jeffrey L. Anderson ³, Frederic S. Castruccio ²,
Mohamad El Gharamti ³ and Ian Grooms ¹

¹ Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA; ian.grooms@colorado.edu

² CGD Laboratory, NSF National Center for Atmospheric Research, Boulder, CO 80305, USA; damrhein@ucar.edu (D.E.A.); fredc@ucar.edu (F.S.C.)

³ Computational and Information Systems Laboratory, NSF National Center for Atmospheric Research, Boulder, CO 80305, USA; jla@ucar.edu (J.L.A.); gharamti@ucar.edu (M.E.G.)

* Correspondence: kate.boden@colorado.edu

Abstract

Global ocean models at non-eddying resolutions currently used for subseasonal to seasonal to decadal (S2S2D) prediction suffer from a severe deficit in internal variability. In ensemble data assimilation (DA), this can lead to under-dispersed ensembles that require inflation schemes. However, inflation corrections do not persist into the forecast phase, causing ensemble spread to collapse at longer lead times. This study evaluates an alternative approach: addressing the variability deficit directly within the model physics using a “Backscatter Package” (BackPack) consisting of the stochastic Stanley, stochastic GM+E, and Leith+E backscatter parameterizations. Implemented within a global MOM6/CESM framework at nominal 2/3° resolution using the DART ensemble DA package, the BackPack’s impacts are compared against cutting-edge adaptive inflation. The results demonstrate that the BackPack substantially increases internal variability and ensemble spread, successfully lowering the amount of required inflation. While reductions in ensemble-mean state errors are modest, the BackPack significantly improves ensemble calibration, assessed using a novel spread–error calibration ratio metric. Although this study only addresses the data assimilation phase, we expect that physics-based BackPack schemes may provide a physically sustainable pathway to maintain spread during the subsequent forecast phase.

Keywords: data assimilation; backscatter; inflation



Academic Editors: Jinyu Sheng and William Perrie

Received: 31 May 2026

Revised: 6 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The ability to predict environmental conditions from subseasonal to seasonal to decadal (S2S2D) timescales has emerged as a cornerstone of modern Earth system science, guiding strategic decisions regarding food security, water resource management, and disaster risk reduction. Improving skill across these timescales requires a global, coupled modeling approach that integrates complex interactions between the atmosphere, ocean, land surface, and cryosphere. The use of global models is essential because, on S2S2D horizons, regional variability is increasingly shaped by teleconnections, where remote climate drivers and large-scale oscillations exert a dominant influence on local conditions.

As in weather forecasting, the accuracy of forecasts on S2S2D timescales depends strongly on the accuracy of the initial conditions. As the forecast horizon lengthens from subseasonal to seasonal to decadal, the accuracy of the predictions depends more and more

strongly on the more slowly evolving components of the system, particularly the ocean. Initial conditions are commonly generated by data assimilation (DA), which encompasses a vast range of methods for combining forecast models with observations to estimate the state or parameters of a system. Modern coupled systems typically use some form of ensemble Kalman filter (EnKF) to generate state estimates, which are used to initialize forecasts or to produce coupled reanalyses [1–10].

The computational cost of global coupled ensemble simulations for data assimilation is such that many groups are currently unable to afford to use cutting-edge kilometer-scale models. Particularly for the ocean, which provides memory, and thus predictability on longer timescales, many prediction systems, including CESM [5], SPEAR [6], CAFE60v1 [7], and NorCPM [9,10], use models with non-eddy resolving resolutions on the order of $1/2$ – 1° . Ocean models at this resolution have unrealistically low levels of variability [11–15]. These unrealistically low levels of variability translate into low ensemble spread. In the context of ensemble prediction, this leads to poorly calibrated forecasts. In the context of data assimilation, this leads to poorly estimated initial conditions, because the degree to which EnKFs adjust the ensemble towards the observations depends on the relative magnitudes of the ensemble spread and the observation errors—when the ensemble spread is low, the EnKF tends to ignore the observations.

Because low ensemble spread is a common problem in EnKF data assimilation, methods to account for low spread are nearly as old as the EnKF itself [16]. These methods typically ‘inflate’ the ensemble either by multiplying ensemble perturbations (i.e., differences between an ensemble member and the ensemble mean) by inflation factors, or by adding inflation increments to the ensemble members. The prediction systems cited above all use some form of inflation to mitigate the under-dispersion of their ensembles. The old adage that ‘an ounce of prevention is worth a pound of cure’ suggests that it might be worthwhile to pursue strategies to fix the models, so that they exhibit realistic variability, rather than attempting to make up for this variability deficit after the fact.

The real ocean exhibits variability in many physical variables and on many time and length scales, much of which is associated with mesoscale eddies. Ocean mesoscale eddies have widths on the order of 100 km [17–19], which is too small to be resolved by the non-eddy resolving models commonly used in global coupled ensemble simulations. As a result, these models unavoidably lack a significant component of real-world variability [11–15]. Even on larger scales that the models are theoretically capable of resolving, they consistently underestimate variability. In the real ocean, some of this missing large-scale variability is driven by mesoscale eddies, whose effects are parameterized in low-resolution models. However, classical turbulent viscosity and mixing parameterizations typically remove energy from the resolved scales, failing to drive large-scale variability at all.

More recently, parameterizations have been developed for these models that are designed (among other things) to generate large-scale variability. While there are many parameterizations designed to increase the kinetic energy and variability of eddy-permitting ocean models, there are fewer targeting ocean models at non-eddy resolving resolutions. The stochastic parameterizations of Brankart [20] and Stanley et al. [21] address errors in the computation of density, and have been shown to increase variability in non-eddy resolving models [13,22,23]. The generic approach to stochastic parameterization by perturbing parameterization tendencies or parameters, developed by Buizza et al. [24] for weather prediction, has also been applied in non-eddy resolving ocean models to increase variability [22,25–28]. Finally, backscatter schemes, which inject energy into the resolved scales from the unresolved scales thereby enhancing variability, have recently been developed for targeting ocean models at non-eddy resolving resolutions [15,29,30].

In this paper we explore the potential for backscatter parameterizations to enhance ensemble spread and improve performance in global ocean data assimilation, leaving for future work an investigation of the effects of backscatter in the forecast phase of S2S2D prediction. A model without backscatter is used as a point of comparison, where the lack of ensemble spread is addressed using adaptive inflation. Comparisons between inflation and backscatter have appeared elsewhere in the literature. Whitaker and Hamill [31], for example, compared the effects of a stochastic kinetic energy backscatter (SKEB) parameterization [32] to their relaxation-to-prior spread (RTPS) and relaxation-to-prior perturbations (RTPP) inflation schemes in an idealized two-layer atmospheric model, and concluded

Tests replacing additive inflation with SKEB in the data assimilation show that it is surprisingly hard to improve upon additive inflation.

Grooms et al. [33] compared stochastic backscatter schemes to inflation in a two-layer QG model and found that the stochastic schemes were superior to tuned, non-adaptive covariance inflation. Ha et al. [34] compared the aforementioned SKEB stochastic backscatter parameterization to the adaptive inflation scheme developed by Anderson [35,36] using the WRF model for atmospheric mesoscale forecasts over the continental United States. They found that

Explicitly accounting for model uncertainty led to a better-tuned ensemble spread, a more skillful ensemble mean, and higher probabilistic scores, as well as significantly reducing the need for inflation. In particular, the stochastic backscatter scheme consistently outperformed . . . the control run with adaptive inflation over almost all levels of the atmosphere both deterministically and probabilistically.

Dufee et al. [37] compared a stochastic parameterization scheme to a tuned, non-adaptive inflation scheme in an idealized sQG model, and found that the stochastic scheme led to better accuracy and calibration than the inflation scheme. These four prior studies show that backscatter parameterizations can lead to improved ensemble data assimilation performance compared to inflation, but that whether backscatter is superior to inflation depends strongly on the backscatter scheme, the inflation scheme, and the model in question. In the specific configurations used here, we find that backscatter does not lead to significant improvements in the accuracy of state estimates as compared to adaptive inflation, but that it does lead to better uncertainty quantification via a better-calibrated ensemble.

Although the focus of this investigation is on model variability and ensemble spread, it bears noting that model bias is another major barrier to the skill of S2S2D forecasts [38,39]. Backscatter parameterizations can have a rectified effect on a model's mean state, thereby impacting (and hopefully reducing) model bias [15,23,40], while alternative methods also exist to account for biases both in the context of data assimilation and subsequent forecasting [41–45].

The remainder of the paper is organized as follows. Section 2 presents the details of the ocean model configuration and of the assimilation system, i.e., the experimental setup. The results are presented in Section 3, and discussed in Section 4.

2. Materials and Methods

2.1. Model Configuration

We use the Modular Ocean Model version 6 (MOM6) ocean model [46] in the CESM2 framework [47] with active ocean and sea-ice components. Except as noted below, the configuration is the same as used by Grooms et al. [15].

The horizontal grid has a nominal $2/3^\circ$ spacing with meridional refinement to $1/4^\circ$ in an equatorial band. The vertical coordinate is z^* [48], with a vertical grid of 75 levels with

thicknesses from 1.8 m at the surface to 505 m at depth. The Redi [49] tracer diffusivity κ_{KHTR} and the GM [50,51] coefficient κ_{KHTH} are parameterized following [52]:

$$\kappa_{\text{KHTH}} = \kappa_{\text{KHTR}} = \alpha e p_1 \frac{1}{H} \int_{-H}^0 \frac{N(z)}{M^2(z)} dz \quad (1)$$

where e is the unresolved mesoscale eddy kinetic energy, $\alpha = 0.09$, H is the local ocean depth, and

$$N^2(z) = -\frac{g \partial_z \rho}{\rho}, \quad M^2(z) = -\frac{g |\nabla \rho|}{\rho}, \quad (2)$$

where ρ is the density and g is the acceleration due to gravity. The GM and Redi coefficients decay with depth following the profile of the first surface mode $p_1(z)$ as described by Grooms et al. [15]. The evolution of e is parameterized using the prognostic formulation of Jansen et al. [53]; their Equation (12). The lateral diffusivity of e is set to 75% of the rate for tracers κ_{KHTR} , following the findings of Grooms [54]. Mixed-layer restratification by submesoscale baroclinic instability is parameterized following [55] as described by Grooms et al. [15].

We use two configurations of the model: a Control case with no backscatter or stochastic parameterizations; and a case with three parameterizations that increase the variability—the Backscatter Package, or BackPack. The MOM6 parameter settings that differ between the two cases are summarized in Table 1, and discussed below. The first component of the BackPack is the Stanley parameterization of errors in the equation of state [21]. In both the Control and BackPack configurations, the Stanley coefficient c_{Stanley} (defined in Equation (25) of Stanley et al. [21]) is set to 0.5 (i.e., the STANLEY_COEFF option in the MOM6 implementation is set to 0.5). Based on offline analysis, Stanley et al. found $c_{\text{Stanley}} = 0.17$. Following Kenigson et al. [23] and Agarwal et al. [13], we find it beneficial to use a larger value in online simulations. The Control configuration uses the deterministic version of the Stanley parameterization (i.e., the STANLEY_A option in the MOM6 implementation is set to 0), while the BackPack configuration uses the stochastic version as described by Equation (49) of Stanley et al. [21] (i.e., the STANLEY_A option in the MOM6 implementation is set to 1). The Stanley parameterization's correction to density is only used in (i) the computation of the hydrostatic pressure (cf. [23]) and (ii) the Bodner parameterization of mixed-layer restratification by submesoscale baroclinic instability (cf. [13]). In the latter, the Stanley density correction is used in the computation of both the mixed-layer-averaged lateral buoyancy gradient $\nabla \bar{b}^z$ and the mixed-layer depth H that appear in Equation (27) of Bodner et al. [55]. The lower panel of Figure 1 shows the time-mean density correction at the surface arising from the Stanley parameterization from the BackPack reference simulation (described below), averaged over calendar year 2011. The Stanley parameterization only impacts density in regions where there is a large lateral temperature gradient, including off the southeast coast of Greenland, at the Brazil–Malvinas confluence, and in the Kuroshio extension and North Atlantic currents (cf. Figure 1 of Kenigson et al. [23]).

Table 1. A summary of MOM6 parameters that differ between the Control and BackPack model configurations.

Parameter Name	Control	BackPack
STANLEY_COEFF	0.5	0.5
STANLEY_A	0	1
SKEB_GM_COEF	0	0.75
LEITHY_CK	0	1

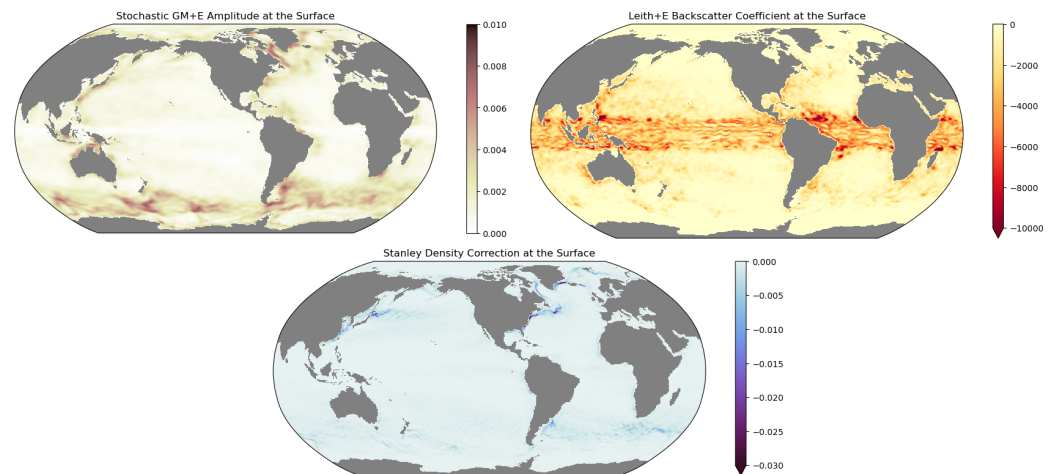


Figure 1. (Left) The stochastic GM+E backscatter amplitude at the surface [m/s]. (Right) The Leith+E harmonic viscosity coefficient at the surface [m²/s]. (bottom) The Stanley parameterization's density correction at the surface [kg/m³]. All are averaged over calendar year 2011 of the BackPack reference simulation.

The second component of the Backscatter Package is the stochastic GM+E parameterization [15]. This scheme introduces stochastic backscatter to the model's lateral momentum budget, where the backscatter rate is proportional to the rate at which the GM parameterization removes resolved potential energy. The configuration used here follows [15] with the following adjustments. First, the stochastic forcing is constructed using spherical harmonics with T192 resolution and a characteristic length scale of 300 km (i.e., $N = 192$ and $L_{\text{stoch}} = 300$ km in Equation (10) of Grooms et al. [15]). The spherical harmonic resolution is lower than in Grooms et al. [15], which slightly reduces the computational cost of the parameterization; the length scale of 300 km is between the 240 km and 360 km scales tested by Grooms et al. [15], striking a balance between having a stronger impact on variability in the northern hemisphere than the 240 km scale and yet staying small compared to the scale over which the backscatter amplitude varies. Second, the backscatter rate is set to 75% of the GM work rate (i.e., $c = 0.75$ in Equation (15) of [15], and $\text{SKEB_GM_COEF} = 0.75$ in the MOM6 implementation). Of the three values tested by Grooms et al. [15] (25%, 50%, and 75%), a backscatter rate of 75% produces the strongest response. The left panel of Figure 1 shows the stochastic GM+E backscatter amplitude at the surface from the BackPack reference simulation (described below), averaged over calendar year 2011.

The third component of the Backscatter Package is the Leith+E parameterization [56,57], which combines a biharmonic 2D Leith viscosity with a negative Laplacian backscatter. As described by Grooms et al. [56], the coefficients of these two terms are set by the following requirements:

1. The dissipation scale for the forward enstrophy cascade should be resolved on the model grid.
2. The rate of backscatter by the negative Laplacian term should be a prescribed fraction of the rate of energy dissipation by the biharmonic term.

Both the Control and BackPack configurations use this scheme, but in the Control simulation the backscatter coefficient is set to zero (i.e., $c_K = 0$ in Equation (24) of Grooms [56]; $\text{LEITH_CK} = 0.0$ in the MOM6 implementation), leaving only a 2D Leith biharmonic viscosity [58,59]. In the BackPack configuration the backscatter term is set to backscatter at the same rate that the biharmonic term is damping, resulting in no net energy dissipation (i.e., $c_K = 1$ in Equation (24) of Grooms [56]; $\text{LEITH_CK} = 1.0$ in the MOM6 implementation). The nondimensional Leith coefficient Y from [56] is set to $90^{1/6}$ (i.e., the LEITH_BI_CONST option in the MOM6 implementation is set to 90). This value is slightly

higher than the value of 75 used by Yassin et al. [57], which further enhances variability in the tropics.

As described by Grooms [56] and Yassin et al. [57], Leith+E backscatter is only active when the biharmonic coefficient predicted by the Leith closure is larger than some background level, which is $10^{12} \text{ m}^4 / \text{s}$ in these configurations. The biharmonic coefficient predicted by the Leith closure scales with the grid spacing to the sixth power. In the implementation here, the ‘grid spacing’ is taken to be the larger of the spacings in the zonal and meridional directions. Since the grid spacing decreases with latitude, the result is that the Leith+E backscatter is primarily active in the tropics. The right panel of Figure 1 shows the Leith+E backscatter coefficient at the surface from the Backpack reference simulation (described below), averaged over calendar year 2011.

All simulations presented in this paper are forced using the JRA55-do v1.5 data set [60]. The JRA55-do forcing data begins in calendar year 1958. We ran two reference simulations, one using the Control configuration and one using the Backpack configuration, initialized from rest and using temperature and salinity fields from WOA18 [61,62].

Figure 2 shows the differences in SST (left) and SSS (right) between the Control and Backpack simulations, averaged over calendar year 2011. The backscatter package clearly impacts not only the variability, but also the mean state. Unpacking the dynamical origins of these differences is beyond the scope of the current investigation (see [13,15,23,57] for more discussion).

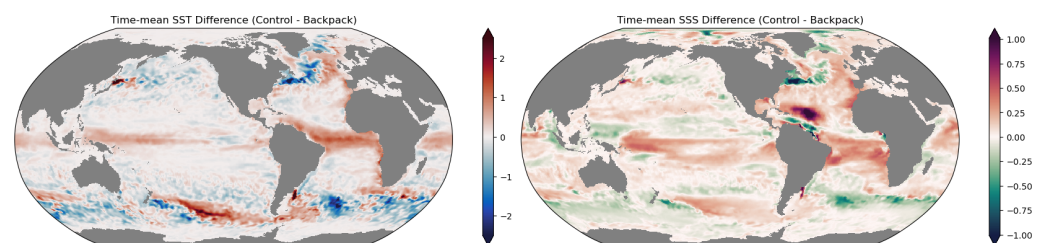


Figure 2. Difference between the Control and Backpack reference simulation sea surface temperatures [°C] (**left**) and sea surface salinity [psu] (**right**). Both are averaged over calendar year 2011.

Figure 3 shows the differences in the standard deviation of SST anomalies between the Control reference simulation and the OISST version 2.1 observational product [63] (top); between the Backpack and Control reference simulations (middle); and between the Backpack reference simulation and OISST (bottom). All results are for calendar year 2011. We compute SST anomalies at each spatial location by removing the projection of the SST time series onto the span of a constant (the mean) plus four Fourier harmonics of the annual period.

In order to ensure an apples-to-apples comparison, the daily OISST data set, which is available on a regular latitude/longitude grid at a resolution of $1/4^\circ$, is first low-pass filtered to remove length scales that are not resolvable on the grid used by the models. We use an anisotropic diffusion-based filter from [64] as implemented in the *gcm-filters* Python package ([65], [version 0.5.1]). We use the Gaussian filter shape with filter scale equal to twice the local model grid scale, anisotropically in each local grid direction. We compute SST anomalies for the filtered data set on the native grid, then compute standard deviations of the anomalies, then interpolate to the model grid for comparison.

The top panel of Figure 3 shows that the Control reference simulation lacks SST variability, even on length scales that it should in principle be able to resolve. (The excess variability in model SST compared to OISST around Antarctica may be related to the fact that SST is unobserved under sea ice, and OISST uses proxy values there.) The middle panel shows that the Backscatter Package significantly increases SST variability compared to the Control case, especially in the Tropical Atlantic and Pacific Oceans. The bottom

panel of Figure 3 shows that although the Backscatter Package improves the variability, the Backpack configuration still lacks variability.

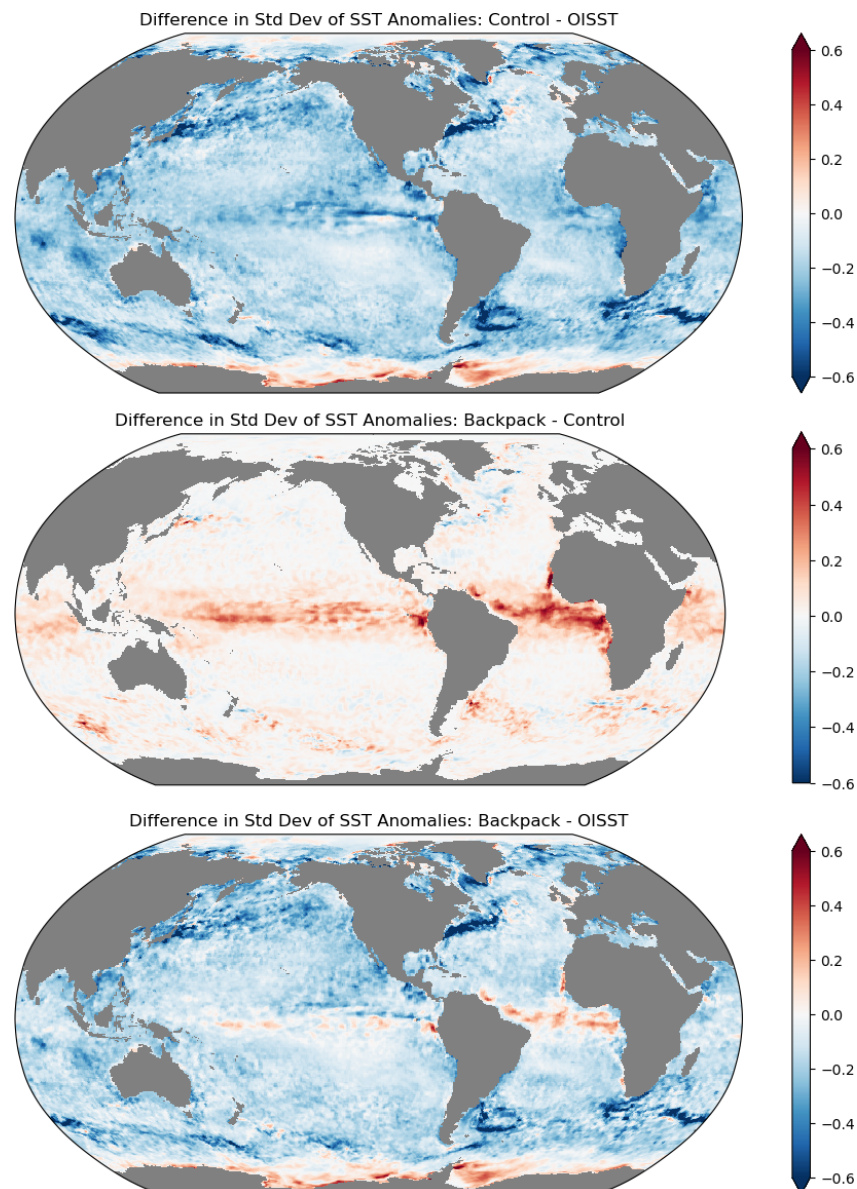


Figure 3. Difference between the standard deviation of SST anomalies in calendar year 2011 between various data sets [$^{\circ}\text{C}$]. (**Top**) Control minus OISST; (**middle**) Backpack minus Control; (**bottom**) Backpack minus OISST.

2.2. Experimental Design: Data Assimilation

The choice of experimental design in this paper was complicated by the realism of the model. Of the four previous comparisons of inflation to backscatter mentioned in the Introduction, three used idealized two-layer or sQG models [31,33,37]. These three investigations used a single reference case from a simulation at higher resolution than the ensembles. The remaining previous investigation [34] used a realistic model with real observations. The downside of using real observations is that there is no reference case available for verification, so [34] used a subset of withheld observations for verification.

The model configurations we adopt here are realistic, and the computational cost of using a high-resolution reference case, as adopted by Whitaker and Hamill [31] and Grooms et al. [33], is prohibitive. We could alternatively assimilate real observations, but such an approach is not without drawbacks. Using one set of observations for assimilation

and another set for verification requires a sufficiently large number of observations for both assimilation and verification. While there were sufficient observations in the atmosphere for this approach to be used by Ha et al. [34], it is not clear if there are enough subsurface observations in the ocean. In addition, the use of real observations requires a model of representation error that accounts for the errors in the assimilating model; performance strongly depends on the model of representation error, but there is no single best-practice approach to modeling representation error for non-eddy ocean models [66]. There are also ocean representation errors associated with uncertainty in the physical meaning of a model's temperature and salinity variables [67].

Faced with a choice between imperfect options, we opt to use a symmetric experimental design, where we generate synthetic observations from both the Control and BackPack reference cases, and assimilate these synthetic observations with both Control and BackPack ensembles. We therefore perform four OSSEs covering calendar year 2011, as summarized in Table 2. The OSSEs are denoted using two-letter sequences CC, CB, BC, BB, where the first letter denotes the model configuration used for the ensemble, and the second letter denotes the model configuration used to generate synthetic observations. All ensembles have 40 members, and the ensemble members are initialized on 1 January 2011, using the restart files from the corresponding reference case on 1 January of years 1971 through 2010. In the context of actual prediction systems, the performance is sensitive to model initialization, and an alternative initialization approach might be used [68].

Table 2. Ensemble and reference model configurations for each OSSE label.

OSSE Label	Ensemble Model Configuration	Reference Model Configuration
CC	Control	Control
CB	Control	BackPack
BC	BackPack	Control
BB	BackPack	BackPack

The synthetic observations are generated by evaluating temperature and salinity in the reference cases at the same spatial locations as in situ observations of temperature and salinity from the World Ocean Database (WOD). These synthetic observations are then perturbed by adding samples from a centered Gaussian distribution with variance 0.25, with units of degrees kelvin squared for temperature, and practical salinity units squared for salinity. The magnitude of the errors in the synthetic observations is larger than the instrument error of typical observing platforms because the observation error for models at this resolution is dominated by representation error [66]; the value chosen here is in the range of values used in other published studies [68–70]. To be clear, we do not use the actual observational values of temperature and salinity from the WOD; we only generate synthetic observations at the locations of observations in the WOD. These synthetic observations are then grouped into consecutive windows of length 5 days, with the first window centered at midnight on 5 January 2011. Each five-day window contains on the order of 250,000 observations. All observations in a window are assimilated simultaneously at the central time of the window. A window length of five days is in the range of other investigations using low-resolution ocean models (e.g., Ref. [7] uses a month-long window while Refs. [2,5] use one-day windows). Binning of observations into windows can lead to temporal representation errors, though this is mitigated here by the fact we perform OSSEs with models that do not include fast mesoscale and submesoscale features. Figure 4 illustrates the locations of temperature observations from six different observing platforms on 10 January 2011.

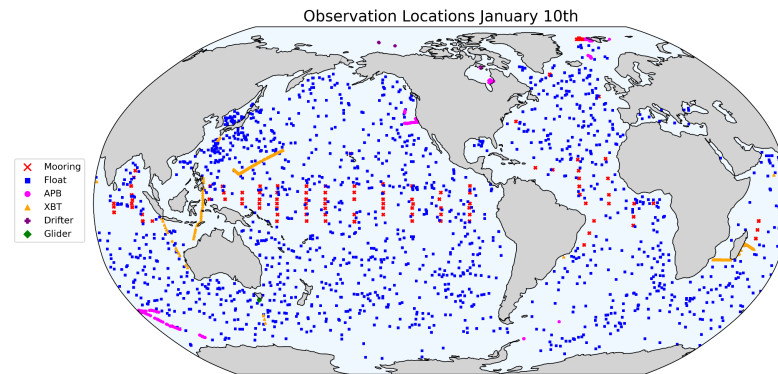


Figure 4. The locations of observations of temperature on 10 January 2011, from six different observing platforms.

The JRA55-do atmospheric forcing is not statistically stationary, so the ensembles on 1 January are biased compared to the reference cases. Figure 5 shows the SST bias between the ensemble means and the reference cases. There are broad-scale patterns of initial ensemble-mean SST bias shared across the four cases that result from their shared atmospheric forcing and shared method of constructing the initial ensemble. The cases with a BackPack reference case (CB and BB) display extra small-scale structure associated with the increased variability of the BackPack reference case. The CB and BC cases have the largest initial bias, since they combine biases from the initialization scheme with differences between the ensemble and reference model configurations (cf. Figure 2).

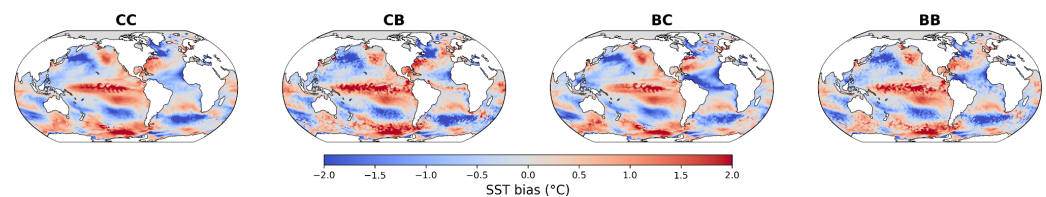


Figure 5. Difference between the initial ensemble-mean SST and the initial value of the reference case for the four OSSEs [°C]. Left to right: CC, CB, BC, BB.

The synthetic observations are assimilated using the Data Assimilation Research Testbed (DART) [71]. DART assimilates observations serially. Each observation is assimilated in a two-step process [72]. The first step applies an EAKF [73] in observation space, while the second step uses regression to convert the update in observation space to an update of the model state vector. Wang and Shen recently explored using deep neural networks in the second step [74]; here, we use the simple linear regression model from [72]. Observations of temperature and salinity are used to update the model's temperature, salinity, and lateral velocity fields only, for the sake of simplicity. Operational configurations typically update more variables, including variables related to sea ice and mixed-layer depth.

Several adaptive inflation algorithms are available in the literature [31,35,36,75–78]. All experiments here use the adaptive inflation algorithm of Gharamti [77], which adaptively sets multiplicative inflation factors for each physical variable at each location in space and time. Anderson developed an adaptive inflation scheme based on innovation statistics in [35], and updated it by making it Bayesian in [36], with a Gaussian prior on the inflation factors. Gharamti updated the method from [36] by using an inverse-gamma prior on the inflation factors, as described in [77], which ensures non-negative inflation factors (factors less than one—implying deflation—are allowed). From one assimilation cycle to the next the inflation factors are damped back towards a value of (1), using the default damping parameter of 0.9, as described by Gharamti [77]. We only apply inflation to

the forecast ensemble, though it can alternatively be applied to the analysis ensemble, or both, before and after assimilation [79]. The use of adaptive inflation in all simulations, rather than turning it off in the OSSEs (BC and BB) that use a Backpack ensemble, is not inconsistent: Inflation is needed to counteract the effect of finite ensemble size in the Backpack experiments. Since the inflation is adaptive, we track the inflation factors and expect the Backpack simulations to need less inflation than the Control simulations.

The state increments are horizontally localized using the Gaspari–Cohn localization function with a cutoff of 0.08 radians (≈ 510 km) for OSSEs with a Control ensemble (CC and CB), and 0.04 radians (≈ 255 km) for OSSEs with a Backpack ensemble (BB). (The ‘cutoff’ is half of the distance from the observation to the nearest location that is not affected by the observation.) No localization was used in the vertical direction. Guided by localization cutoffs reported in the literature [2,5,7,80], preliminary tests used the same localization cutoff (0.08) for Backpack and Control ensembles, but found that the Backpack ensembles were unstable with the longer cutoff. Further investigation revealed that the ensemble correlations are frequently shorter in the Backpack ensemble, so a large localization radius allows state-variable updates associated with spurious correlations, which can lead to instability.

This is illustrated in Figure 6, which shows ensemble correlations between SST centered at 140° W, 60° S (other centers display similar behavior) and other locations on 30 April 2011, for the CC (left) and BB (right) experiments. The black solid line shows the wider localization used in the CB and CC cases while the blue dashed line shows the narrower localization used in the BC and BB cases. Both ensembles display correlations at long ranges that are probably spurious, i.e., the result of sampling errors at a finite ensemble size. The range of realistic correlations in the BB OSSE appears to be smaller than in the CC OSSEs, which is consistent with the use of a narrower localization with the BB OSSE.

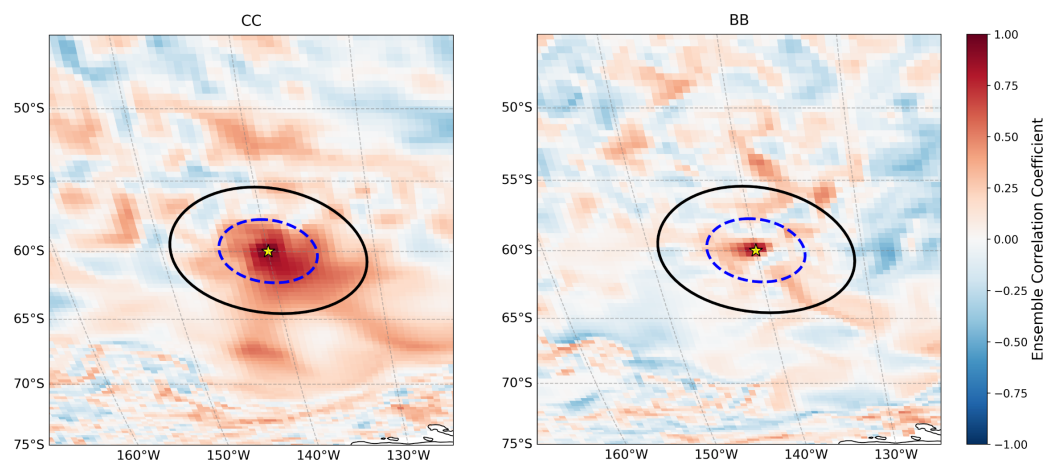


Figure 6. Ensemble correlations between SST at the location labeled with a star (140° W, 60° S) and other locations on 30 April 2011, for the CC experiment (left) and the BB experiment (right). The black solid line shows the wider localization used in the CB and CC cases while the blue dashed line shows the narrower localization used in the BC and BB cases.

Note that no formal tuning was performed to arrive at these filter parameters. The initial value 0.08 was chosen to be in the range of values previously reported in the literature [2,5,7,80], with the lower value of 0.04 consistent with values used in models with higher resolution [81]. Practitioners might use one of the adaptive localization schemes available in the literature to avoid having to manually tune a localization radius [82–84].

A final detail of the assimilation setup involves the treatment of sub-freezing seawater temperatures. A free run of the ocean model typically does not produce temperatures below the local freezing point; instead, temperatures below freezing result in the formation of sea

ice. Data assimilation increments can produce sub-freezing temperatures though, which can lead to instability in a model run. To combat this, we adjusted the temperatures between the end of an assimilation and the beginning of the next forecast to remove sub-freezing temperatures. Specifically, whenever the temperature T falls locally below $-0.054S$, where S is the local salinity, we replace the sub-freezing value of T by $-0.054S + 0.01$.

3. Results

3.1. Backpack and Spread

To demonstrate the effect of Backpack on the ensemble spread, we ran free (non-assimilating) year-long 40-member ensemble simulations for both the Backpack and Control model configurations. Figure 7 shows the temporal evolution of the spatially averaged ensemble spreads in SSS (left) and SST (right) for both configurations. Ensemble spread is defined here as the square root of the spatial mean of the ensemble variance. Early on, both configurations exhibit comparable spread, which gradually decays since the ensemble members see the same atmospheric forcing. Over time, however, the Backpack ensemble maintains greater spread in both variables compared to the Control. This demonstrates the capacity of Backpack to sustain ensemble variance in a free run, which is consistent with the higher temporal variability observed in its single-member free run relative to the Control configuration (Figure 3).

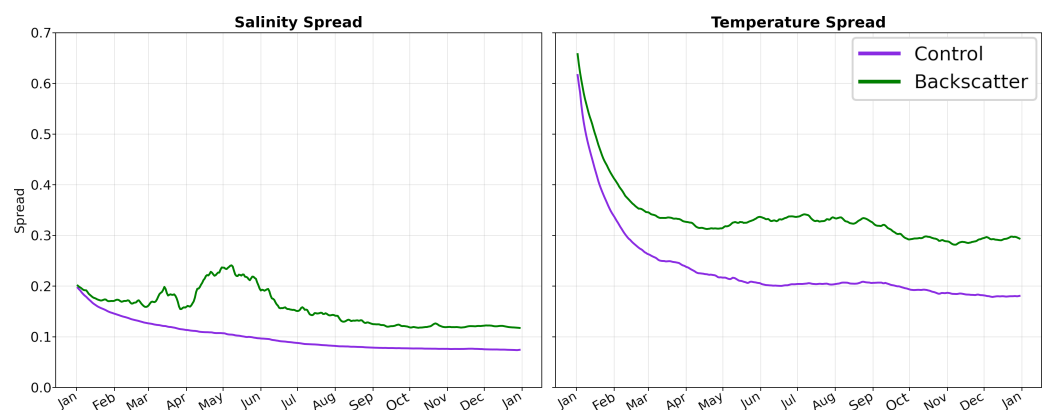


Figure 7. Spread of ensemble SSS [psu] (left) and SST [°C] (right) for two ensemble simulations with no DA using the Control and Backpack model configurations. The value plotted here is the square root of the spatial mean of the ensemble variance of SSS or SST.

3.2. Inflation

Figure 8 shows the adaptive inflation factors for SST averaged in space as a function of time for the four OSSEs. The inflation factors for SSS are similar. All experiments begin with an biased ensemble (cf. Figure 5). The BB and BC cases require less inflation to overcome this initial bias, in contrast to the CC and CB cases which both start with higher inflation values. Once the initial bias has been largely erased, the CC and BB experiments have no further differences between their reference and ensemble, and inflation primarily acts to account for sampling errors.

In contrast, the CB and BC cases have a mismatch between ensemble and reference cases throughout their runs. The CB case requires considerably more inflation than all other cases throughout the run to deal with this ensemble–reference mismatch, whereas the BC case, which has the same ensemble–reference mismatch, requires about the same amount of inflation as the CC and BB cases. The CC case eventually requires less inflation than the BB and BC cases, but the CC case is not a good proxy for real operational data assimilation because the smooth Control model lacks realistic variability, artificially making the data assimilation problem easier.

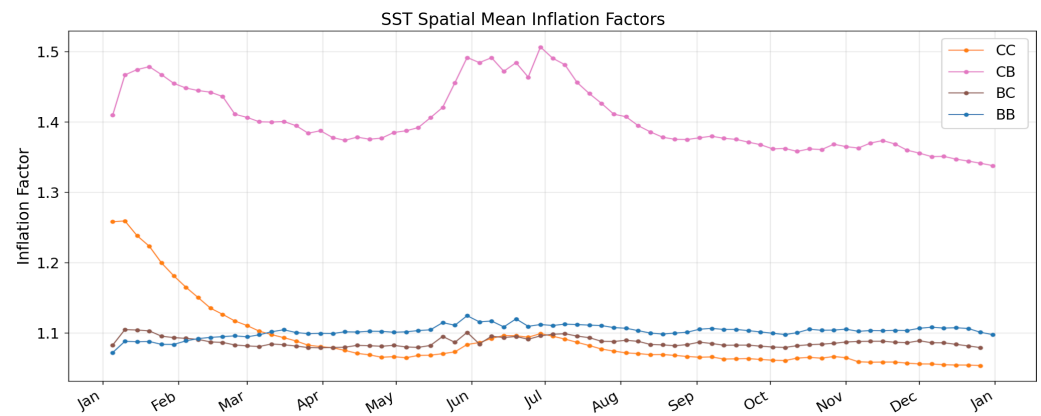


Figure 8. Spatially averaged adaptive inflation factors for SST as a function of time for the four OSSEs.

Figure 9 shows maps of the adaptive inflation factors for SST averaged in time from April through December for the four OSSEs. The salinity inflation factors are extremely similar. In all cases, inflation is high in the Tropical Pacific, reflecting the adaptive inflation's efforts to maintain spread, countering the spread-reducing effect of the dense observing system associated with the TAO/TRITON array. Inflation factors are also elevated in all cases in regions of enhanced variability, like the Southern Ocean and western boundary current extensions.

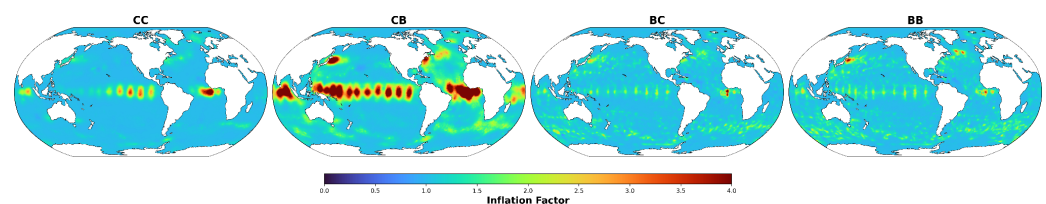


Figure 9. Time-averaged inflation factors for SST for all four OSSEs. The averaging window is April through December to avoid the effects of the initial ensemble bias.

The CB case has much higher inflation than the other cases throughout the tropics and in western boundary currents, presumably because of the difference between the ensemble and reference model configurations, which is large in these regions (Figure 2). It is interesting to note that the BC case, which has the same difference between ensemble and reference model configurations, does not require extra inflation compared to the BB case in these regions, because the BackPack model configuration leads to sufficient ensemble spread.

To underscore that these results are not confined to the surface, in Figure 10 we plot the horizontal-mean inflation factors for temperature as a function of depth and time for the four OSSEs. The results are qualitatively consistent with the results for SST in Figure 8: the CB case requires by far the most inflation. Both at and below the surface, the BC case requires less inflation than the BB case, presumably because the BC case is assimilating observations from a smooth model, which are easier to match than observations from a more variable model. Differences in inflation factors between the cases with Control and BackPack ensembles may also be related to the different localization radii used in these configurations.

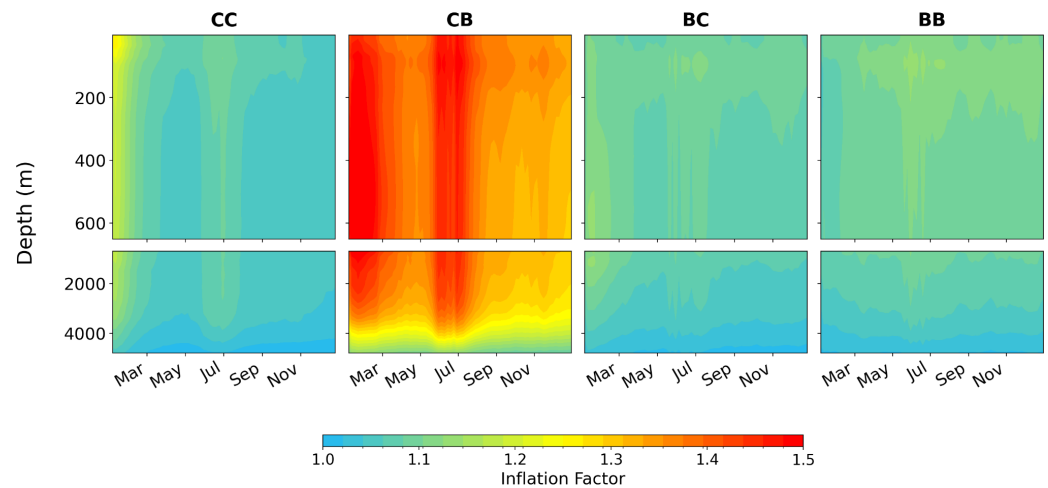


Figure 10. Horizontally averaged inflation factors for temperature as a function of depth and time for the four OSSEs.

3.3. Accuracy

To evaluate the accuracy of the states produced by the assimilation, this section shows the RMSE for the four OSSEs. We begin with Figure 11, which shows RMSEs for SST and SSS, computed as the square root of the spatial mean of the squared error between the reference case and the ensemble mean. The CC and BB cases have initially biased ensembles (cf. Figure 5) but no systematic model bias. Figure 11 shows that this initial bias is largely erased in the first month in the CB, BC, and BB cases, and in the first two to three months in the CC case. After this initial adjustment period, we find that although the BB case has the largest SST errors, only slightly larger than the CB and BC cases, the CB case has worse SSS errors than all the other cases. For both SST and SSS, the CC case, which is a perfect model with limited variability, has the lowest RMSE.

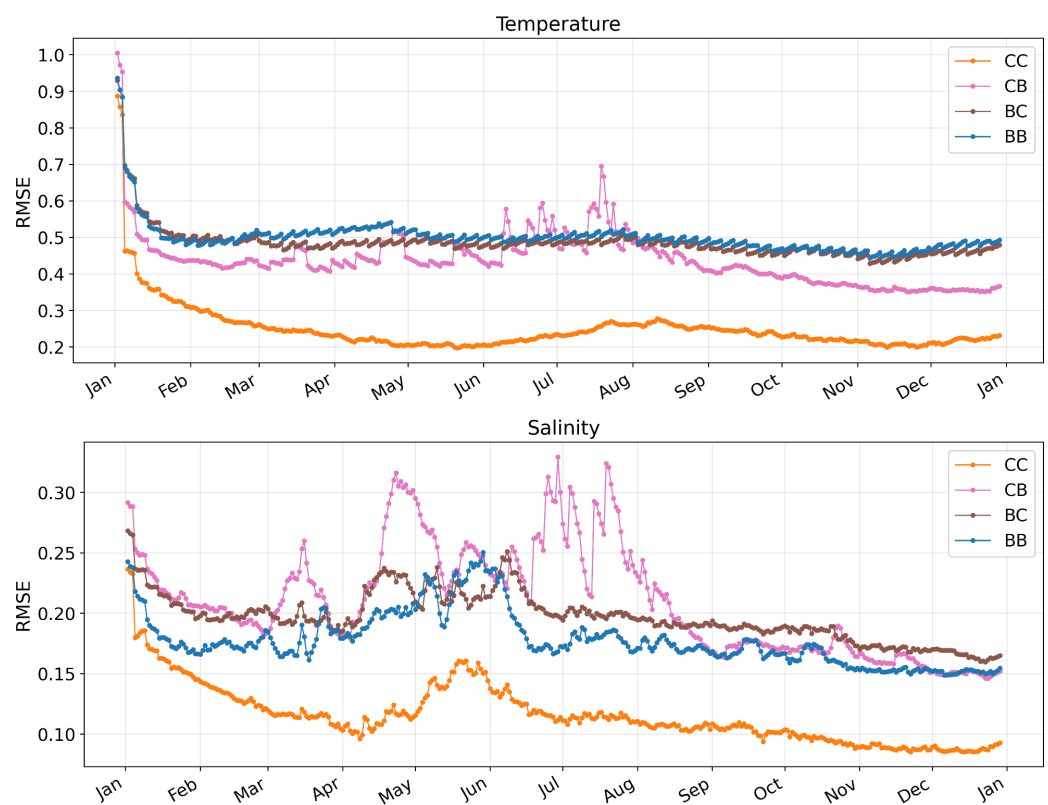


Figure 11. Horizontally averaged RMSE for SST [°C] (**top**) and SSS [psu] (**bottom**) from the four OSSEs.

The spatial structure of the RMSE at the surface is shown in Figure 12, averaged from April through December to avoid the effects of the initial ensemble bias. The BB and BC cases exhibit larger SST errors in the Southern Ocean than the other two cases, reflecting the increased spatial variability of the BackPack model in that region, generated primarily by the stochastic GM+E component of the backscatter package. This increased RMSE may reflect a ‘double-penalty’ effect, where a model that has a small-scale feature in the wrong place exhibits higher error than a model that completely lacks the small-scale feature [85–88]. This also suggests that the in situ observing system is insufficient to completely resolve the BackPack model’s variability in the Southern Ocean. The double-penalty effect may also explain why the SST RMSE is lower for the BC case than the BB case in Figure 11: the BB case is penalized for having a small-scale feature in the wrong place and for lacking a small-scale feature in the right place, while the BC case is only penalized for having a small-scale feature where there is not one in the reference case. Increased temperature and salinity RMSEs in the Tropical and North Atlantic Ocean in the CB case probably reflect the systematic difference between the Control and BackPack models in that region shown in Figure 2.

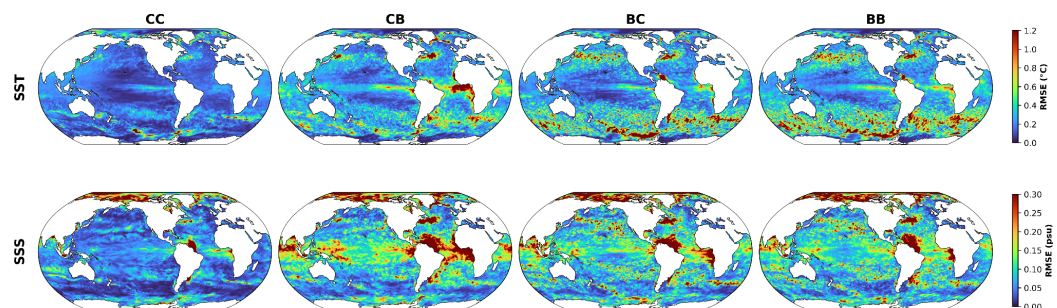


Figure 12. Time-averaged RMSE for SST [°C] (top) and SSS [psu] (bottom) from the four OSSEs. The averaging window is April through December to avoid the effects of the initial ensemble bias.

To illustrate the subsurface behavior, in Figure 13 we show the horizontally averaged RMSE for temperature (top row) and salinity (bottom row) for the four OSSEs. Below the surface, at around 100–200 m depth, the temperature RMSE for the BB case is generally lower than for the CB and BC cases, which reverses the behavior at the surface shown in Figure 11. The CB case evidently has an initial bias in both temperature and salinity at around 1 km depth that takes some time to resolve. This bias is associated with the fact that the CB case has a mismatched ensemble and reference model, and not with the construction of the initial Control ensemble, since this feature at 1 km depth is not seen in the CC case. The BC case resolves this bias more effectively than the CB case. As at the surface (Figure 11), the salinity RMSEs for the BB case are generally lower than for the CB and BC cases.

3.4. Spread–Error Calibration

To assess the quality of the uncertainty quantification provided by the BackPack and Control ensembles we consider spread–error calibration. Spread–error calibration is generally only a necessary but not a sufficient condition for perfect ensemble calibration, which occurs when the true state is statistically indistinguishable from the ensemble members [89]. A more complete probabilistic verification of ensemble calibration might include evaluation of the continuous ranked probability score (CRPS [90]), rank histograms [91], and reliability diagrams [92].

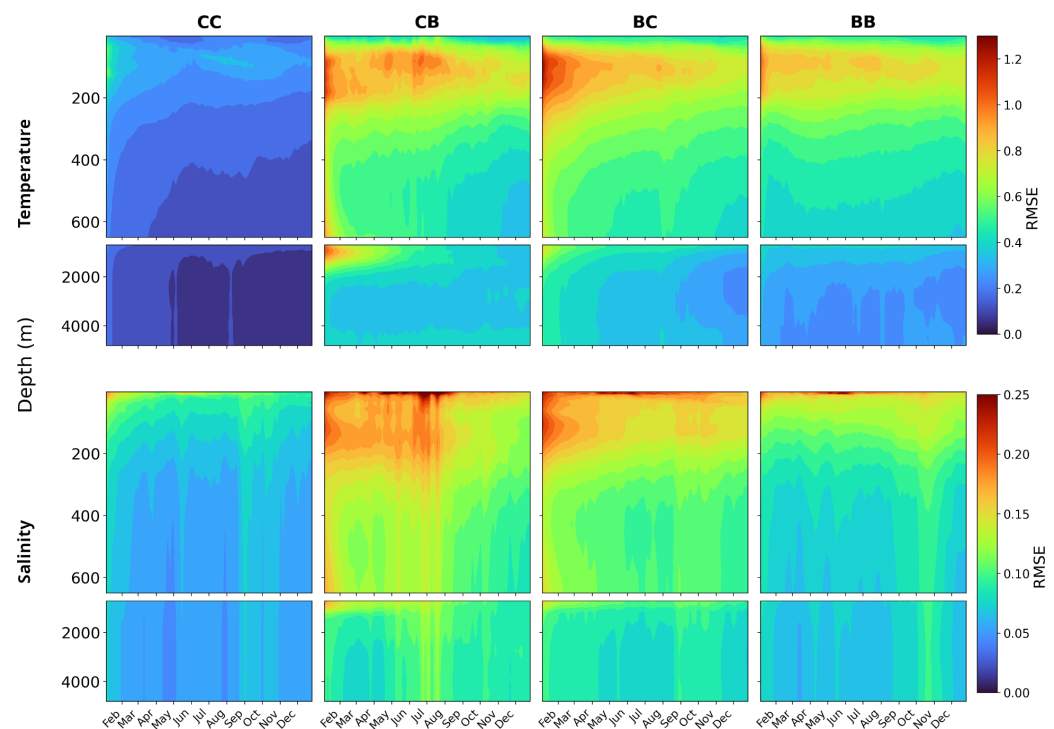


Figure 13. Horizontally averaged RMSE for temperature [°C] (top) and salinity [psu] (bottom) from the four OSSEs.

Let $x(s, t)$ denote a quantity of interest, e.g., temperature or salinity, at some location (s, t) in space and time. The true value of $x(s, t)$ is obtained from the reference case, and is denoted $x_0(s, t)$. The ensemble produces a set of values $\{x_n(s, t)\}_{n=1}^N$, where in all our experiments $N = 40$. The ensemble mean $\bar{x}(s, t)$ and standard deviation (spread) $s(s, t)$ are

$$\bar{x}(s, t) = \frac{1}{N} \sum_{n=1}^N x_n(s, t), \quad s(s, t) = \left[\frac{1}{N-1} \sum_{n=1}^N (x_n(s, t) - \bar{x}(s, t))^2 \right]^{1/2}. \quad (3)$$

The traditional approach to assessing spread–error calibration computes the ratio of average spread to average error, defined as

$$\left[\frac{\langle s(s, t)^2 \rangle}{\langle (x_0(s, t) - \bar{x}(s, t))^2 \rangle} \right]^{1/2} \quad (4)$$

where $\langle \cdot \rangle$ denotes an average in space or in time. Sometimes correction factors are included in the ratio to account for finite sample-size effects [93]. Values less than one are commonly interpreted as implying an under-dispersed ensemble, while values greater than one imply an over-dispersed ensemble.

The theory behind the traditional approach is based on modeling the ensemble and the truth as realizations of random variables [93]. For an unbiased ensemble of infinite size, the expected value of the variance equals the expected value of the squared error in the ensemble mean. But the variance of both quantities could be large, so that the ensemble variance and squared error in the mean computed at a single location or time might differ drastically from each other, even at the limit of infinite ensemble size. The spatial or temporal average $\langle \cdot \rangle$ that appears in the traditional ratio is intended to decrease this variance, but it does not necessarily achieve this reliably. Averages, whether in space or in time, are sensitive to outliers, so that both $\langle s(s, t)^2 \rangle$ and $\langle (x_0(s, t) - \bar{x}(s, t))^2 \rangle$ could be dominated by small regions with large errors or large spread.

To address this limitation of the traditional approach, we develop here a novel measure of spread–error calibration. We denote the local error–spread ratio as

$$r(s, t) = \frac{x_0(s, t) - \bar{x}(s, t)}{s(s, t)}, \quad (5)$$

and define the spread–error calibration ratio (SECR) to be

$$C_R = \left[\frac{\gamma}{\langle r^2(s, t) \rangle} \right]^{1/2} \quad (6)$$

where

$$\gamma = \frac{N-1}{N-3} \left(1 + \frac{1}{N} \right). \quad (7)$$

The factor of $\sqrt{\gamma}$, which is approximately 1.04 at $N = 40$, accounts for finite sample-size effects, while inverting after averaging produces an easily interpretable result: C_R values equal to one suggest a well-calibrated ensemble, while values above and below one indicate over-dispersed and under-dispersed ensembles, respectively. The finite-sample-size correction is determined by the following theorem.

Theorem 1. Let $\{X_n\}_{n=1}^N$ be independent random variables with normal distribution $\mathcal{N}(\mu_e, \sigma_e^2)$, and let X_0 be independently distributed as $\mathcal{N}(\mu_t, \sigma_t^2)$. Define

$$\bar{X} = \frac{1}{N} \sum_{n=1}^N X_n, \quad V = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X})^2, \quad R^2 = \frac{(X_0 - \bar{X})^2}{V}. \quad (8)$$

For $N > 3$, the expected value of R^2 is

$$\frac{N-1}{N-3} \left(\frac{\sigma_t^2 + (\mu_t - \mu_e)^2}{\sigma_e^2} + \frac{1}{N} \right). \quad (9)$$

The proof can be found in Appendix A. Insight can be gained by examining special cases.

- In the case of a perfectly calibrated ensemble, which has population mean μ_e equal to the true mean μ_t and population variance σ_e^2 equal to the true variance σ_t^2 , the expected value of R^2 equals $\frac{N-1}{N-3} \left(1 + \frac{1}{N} \right)$, which justifies the sample correction factor γ in the definition of C_R .
- In the case of an unbiased ensemble, the expected value of R^2/γ reduces to

$$\frac{1 + N \frac{\sigma_t^2}{\sigma_e^2}}{1 + N}. \quad (10)$$

When the ensemble is under-dispersed ($\sigma_e^2 < \sigma_t^2$), the expected value of R^2/γ is greater than one, and vice versa, which naturally results in a C_R value less than one, matching our intuitive definition of under-dispersion.

- All else being equal, the expected value of R^2 is larger for a biased ensemble than for an unbiased ensemble. The effect of a biased ensemble is therefore qualitatively similar to the effect of having an under-dispersed ensemble, as it produces C_R values less than one.
- If the ensemble is perfectly calibrated but non-Gaussian, then the expected value of R^2/γ may deviate from unity.

To further motivate the SECR, consider that we can compute $r^2(s, t)/\gamma$ at a location in space and time using the ensemble and true state, but we can only regard this as a realization of a random variable $R^2(s, t)/\gamma$ whose expected value would equal unity if the ensemble were Gaussian and perfectly calibrated. The variance of this random variable could be large, in which case the computed value $r^2(s, t)/\gamma$ could be far from one even for a perfectly calibrated Gaussian ensemble. Computed values $r^2(s, t)$ at single points of space and time are therefore not reliable indicators of ensemble calibration.

We resort to spatial or temporal averaging to reduce the variance in individual computed values of $R^2(s, t)$. The intuition that averaging reduces variance can be justified by the law of large numbers. There are many laws of large numbers that differ in their assumptions and in the precise nature of their conclusions; they all conclude that the arithmetic mean of a set of random variables converges to the arithmetic mean of the expected values of those variables [94]. On this basis, we expect that for a perfectly calibrated Gaussian ensemble the variance of the spatial or temporal mean $\langle R^2(s, t) \rangle$ should decrease to zero as the domain of averaging increases to infinity, unless some improbable and pernicious correlation structure prevents the law of large numbers from applying. Practically speaking, we presume that for a sufficiently large averaging domain, computed values $\langle r^2(s, t) \rangle/\gamma$ far from unity reliably indicate poorly calibrated ensembles.

To connect with the intuition that values less than unity correspond to under-dispersion, the SECR considers the inverse value $\gamma/\langle r^2(s, t) \rangle$, and to connect with traditional metrics of spread vs. error (rather than variance vs. squared error), the SECR takes the square root. From a theoretical standpoint, the computed value C_R should be regarded as a realization of the random variable $\sqrt{\gamma/\langle R^2(s, t) \rangle}$. Jensen's inequality indicates that C_R is a realization of a random variable whose expected value is larger than one, though the Jensen gap is small when the variance of $\langle R^2(s, t) \rangle/\gamma$ is small:

$$1 = \left[\frac{\gamma}{\mathbb{E}[\langle R^2(s, t) \rangle]} \right]^{1/2} \leq \mathbb{E} \left[\left(\frac{\gamma}{\langle R^2(s, t) \rangle} \right)^{1/2} \right]. \quad (11)$$

Practically speaking, this means that the computed C_R for a perfectly calibrated ensemble might exhibit a slight positive bias (yielding values slightly greater than 1), but that this bias becomes negligible as the averaging domain grows and the variance of $\langle R^2(s, t) \rangle$ shrinks.

An advantage of the SECR is that by taking the ratio before averaging the result it is not overly sensitive to small regions with large errors. An advantage of the traditional approach over the SECR is that it does not involve any assumption of Gaussianity. The newly proposed SECR is derived under an assumption of Gaussianity, and should therefore not be used for variables like sea-ice concentration or precipitation, whose distributions are often non-Gaussian. We apply it here to ocean temperature and salinity, whose distributions are consistent with Gaussianity. Both approaches only provide a necessary, but not sufficient, condition for the ensemble to be well-calibrated [89].

Figure 14 plots the spatial-mean spread–error calibration ratio C_R for SST (top) and SSS for the four OSSEs. In all four experiments the ensembles are under-dispersed, though the spread–error calibration is better for salinity than for temperature. The BB and BC cases are both better calibrated than the CB and CC cases, which are comparable to each other.

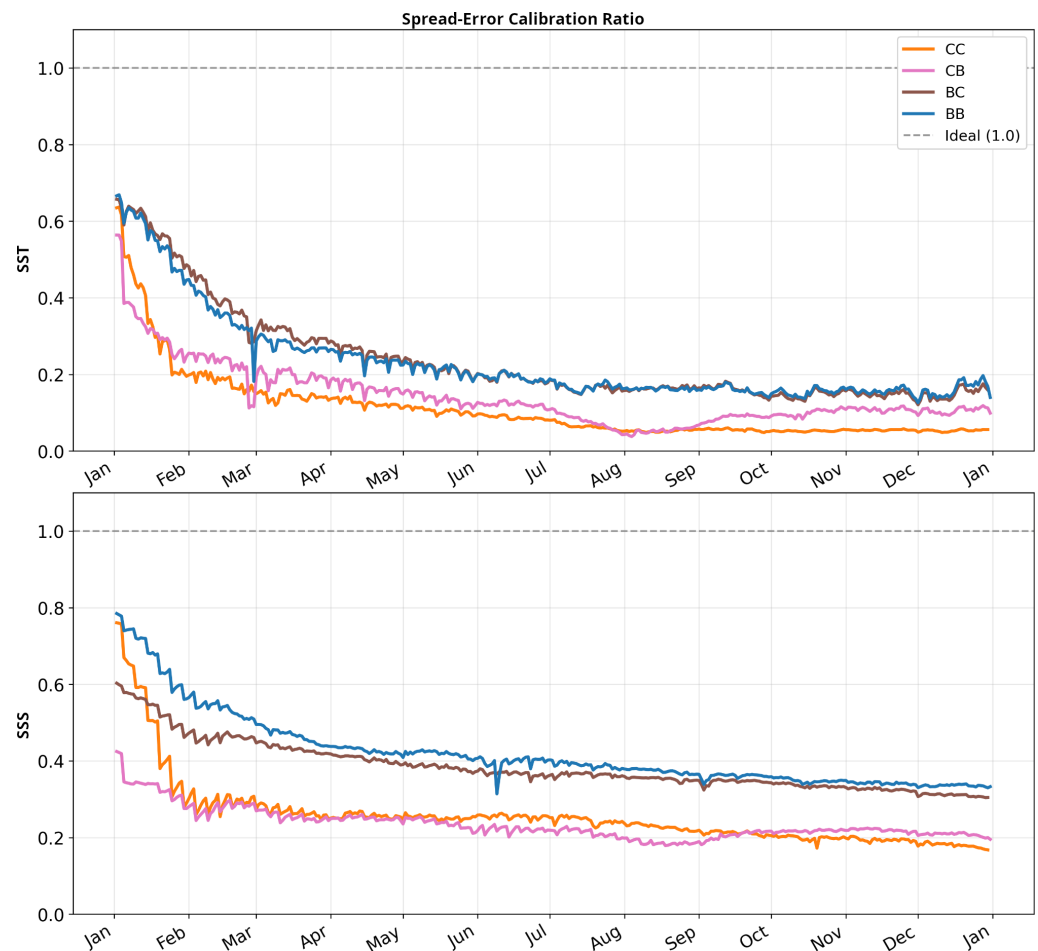


Figure 14. Horizontally averaged spread–error calibration ratios C_R for SST (top) and SSS (bottom) from the four OSSEs. Values above 1 are over-dispersed, while values below 1 are under-dispersed.

Figure 15 shows maps of the time-mean spread–error calibration ratio C_R for SST (top) and SSS (bottom) for the four OSSEs. The averaging window is April through December to avoid the effects of the initial ensemble bias. Whereas the spatial-mean spread–error calibration ratios show that the ensembles are significantly under-dispersed on average globally, the spatial maps show that the ensembles, though still not over-dispersed, are well calibrated in some regions. The BB experiment is clearly the best calibrated in terms of both SST and SSS, with the Backscatter Package leading to improved spread–error calibration primarily in the tropics and Southern Ocean, and also to a lesser extent in the North Atlantic and North Pacific. Even the BB ensemble remains significantly under-dispersed in the subtropical South Pacific and subpolar North Pacific; this is consistent with Figure 1, which shows that neither the stochastic GM+E nor the Leith+E backscatter schemes are active in those areas.

Since the SECR is a novel measure of spread–error calibration, we provide for comparison the traditional ratio Equation (4) in Figure 16, where the averaging window is April through December to avoid the effects of the initial ensemble bias. The results are broadly similar to the SECR results shown in Figure 15, except that the ensembles all appear less under-dispersed using the traditional metric. This suggests that the new approach provides a more stringent test of spread–error calibration than the traditional approach. Differences in inflation factors between the cases with Control and Backpack ensembles may also be related to the different localization radii used in these configurations.

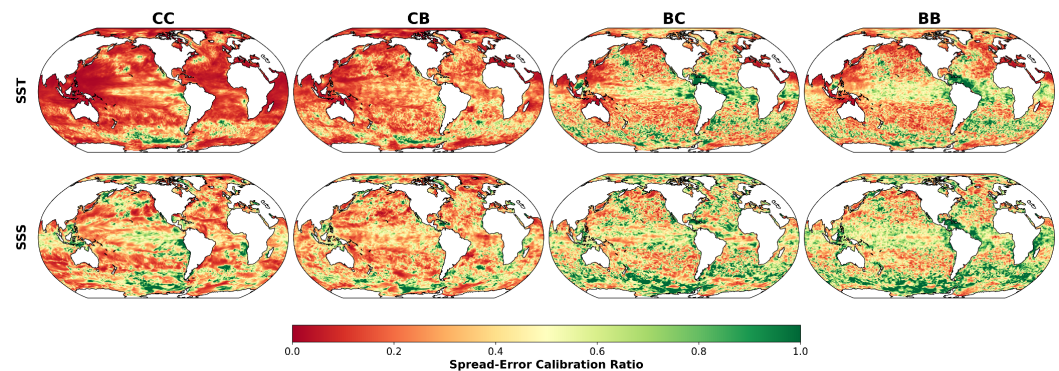


Figure 15. Time-averaged spread–error calibration ratios C_R for SST (**top**) and SSS (**bottom**) from the four OSSEs. The averaging window is April through December to avoid the effects of the initial ensemble bias.

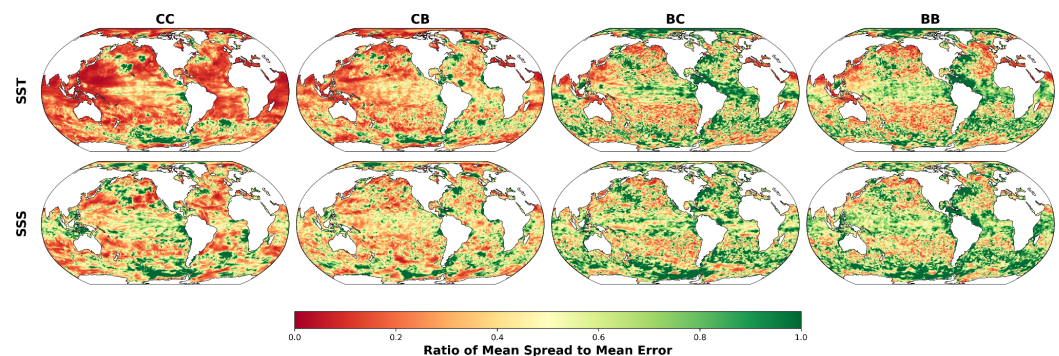


Figure 16. Traditional spread-to-error ratio (Equation (4)), computed using a time average, for SST (**top**) and SSS (**bottom**) from the four OSSEs. The averaging window is April through December to avoid the effects of the initial ensemble bias.

4. Discussion

In this paper we explore the potential for backscatter and stochastic parameterizations to increase ensemble spread and improve performance in global ocean data assimilation, targeting initialization for global coupled prediction systems on S2S2D timescales and global coupled reanalyses. These applications commonly use ocean models at non-eddy resolving resolutions, and such models are often woefully deficient in terms of their variability. This lack of variability can be mitigated in a data assimilation context (for producing reanalyses or initial conditions for S2S2D forecasts) using inflation schemes, of which there are many. Alternatively, the lack of variability can be addressed within the model itself through the addition of stochastic parameterizations and backscatter schemes.

There are few previous investigations comparing inflation to stochastic parameterizations, and none of them target global ocean models at non-eddy resolving resolutions. Previous investigations have been performed in the context of atmospheric synoptic scales [31] and mesoscales [34], and of oceanic mesoscales [33,37], with inflation schemes from simple tuned multiplicative inflation [33,37] to spatio-temporally varying and adaptive multiplicative and additive inflation schemes [31,34]. The results are mixed. In some cases variance-increasing parameterizations are clearly superior, while in other cases there is no apparent benefit compared to inflation. Clearly the decision depends on the parameterizations and inflation schemes in question.

The investigation here uses the MOM6 ocean model in the CESM framework at a nominal $2/3^\circ$ resolution, forced by an atmospheric reanalysis product. We investigate the effects of three combined parameterizations, which we call the Backscatter Package (BackPack):

1. A stochastic correction to the model density, applied in the computation of the hydrostatic pressure and in the mixed-layer eddy restratification scheme [21];
2. The stochastic GM+E backscatter scheme [15];
3. The Leith+E backscatter scheme [56,57].

We use the DART data assimilation package with synthetic observations of temperature and salinity, whose times and locations are taken from in situ observations in the WOD. The assimilation scheme uses the adaptive inflation method of Gharamti [77] applied to the forecast ensemble. Future work might explore applying inflation to the analysis ensemble or to both the forecast and analysis, following Gharamti et al. [79].

We find that using the Backpack model configuration clearly increases variability and ensemble spread, leading to lower inflation factors. The Backpack leads to reduced errors in the ensemble mean state, especially in salinity, but the difference compared to adaptive inflation without the Backpack is modest at best. The Backpack's main benefit is that it significantly improves model calibration, which we measure using a novel metric called the spread–error calibration ratio. Although all of the experiments conducted here have under-dispersed ensembles, the Backpack simulations have much larger spread and hence much better spread–error calibration.

All experiments exhibit under-dispersion. One might initially suspect this to result from model errors, but both the CC and BB experiments exhibit under-dispersion even in the absence of model errors. It is possible that in some regions the initial bias in the ensemble (cf. Figure 5) might not be fully corrected by the end of the first year, perhaps due to the sparsity of the in situ ocean observing system, resulting in under-dispersion despite a lack of model errors. Errors in the data assimilation scheme, especially those associated with spurious correlations, which tend to excessively reduce ensemble spread, might also play a role. In particular, under-dispersion in these experiments could be related to lack of localization in the vertical direction; the use of vertical localization would result in smaller updates to the state variables, and hence smaller reductions in ensemble spread.

The experiments using a Control ensemble use a larger horizontal localization radius than the experiments using a Backpack ensemble, and thus (in principle) allow for more spurious updates and as a result more spurious ensemble deflation than the Backpack experiments. This could also partially explain why the experiments with a Control ensemble are more under-dispersed than the experiments with a Backpack ensemble. However, Figure 6 suggests that the Control ensemble simply has longer-range correlations than the Backpack ensemble, and thus leads to more state-variable updates and greater reductions in ensemble spread at each assimilation cycle. Tighter localization radii might mitigate this effect in future investigations: in addition to removing long-range correlations associated with sampling errors, localization might also be used to remove long-range correlations stemming from model errors.

Modelers considering adopting the Backpack may wish to know which component is most responsible for the improvements. Figure 1 shows that the three components of the Backpack are relatively separated in space, suggesting that improvements in any given region are driven by whichever component of the Backpack is active in that region. That said, the three components may have nonlinear interactions along energy pathways (for example, after Leith+E modifies the vorticity field in the tropics, it may affect the mean flow in GM+E-active regions through wave propagation or teleconnections), leading to an effect from the combined Backpack that may be greater than the sum of its parts. In the future, single-scheme experiments could be used to isolate the effect of each component and to identify any interactions between components.

Although we do not perform S2S2D ensemble forecasts, we expect that an ensemble initialized using data assimilation with a Backpack ensemble would have a larger and

better-calibrated spread during the forecast phase, providing better uncertainty quantification than an ensemble without the Backpack (cf. [25]). The spread in the non-BackPack ensemble would have been, during the data assimilation phase, supported by the adaptive inflation algorithm, but during the forecast phase this adaptive inflation would not be active and we would expect the ensemble spread to slowly decay, leading to increasingly poor calibration of the forecast at longer lead times. In contrast, the spread of the Backpack ensemble is primarily supported by the Backpack parameterizations, which do not switch off in the transition from data assimilation to S2S2D forecasting. We expect an ensemble S2S2D forecast performed using the Backpack parameterization suite might be better calibrated than a forecast without the Backpack. Results from non-assimilating ensemble simulations suggest that the Backpack ensemble does maintain a higher spread than the Control ensemble at lead times up to one year (Figure 7). Verifying these expectations experimentally is a topic for future investigation.

Although the Backpack suite clearly helps increase the internal variability and ensemble spread of the ocean model, even the Backpack ensemble remains significantly under-dispersed in our experiments, despite the use of adaptive inflation. Future work might attempt to further reduce the variability deficit through the development of stochastic parameterizations targeting processes and areas not covered by the methods in the Backpack suite. Figure 15 shows, for example, that ensemble SST remains significantly under-dispersed in many regions, including the eastern and subpolar North Atlantic, the Mediterranean, the Arabian Sea and the Bay of Bengal, the southwestern and northeastern North Pacific, and the subtropical South Pacific. Ensemble SSS is also under-dispersed, though to a lesser extent, in these locations. One option to enhance surface variability could be to use stochastic and machine learning parameterizations of air–sea fluxes [95–99].

Author Contributions: Conceptualization, K.B., D.E.A. and I.G.; methodology, K.B., D.E.A., J.L.A., F.S.C., M.E.G. and I.G.; software, K.B., F.S.C., M.E.G. and I.G.; validation, K.B.; formal analysis, K.B. and I.G.; investigation, K.B.; resources, I.G.; data curation, K.B. and I.G.; writing—original draft preparation, I.G.; writing—review and editing, K.B., D.E.A., J.L.A., F.S.C., M.E.G. and I.G.; visualization, K.B. and I.G.; supervision, I.G.; project administration, I.G.; funding acquisition, I.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NOAA Weather Program Office grant number NA22OAR4320151-T3-01S020, which also funded the APC.

Data Availability Statement: This research used the Modular Ocean Model, version 6 (MOM6 [46,100]), the Data Assimilation Research Testbed (DART [71]) software package, version 11.19.2 [101], and the gcm-filters software package [64,65], version 0.5.1. Data and scripts associated with this research can be found in [102].

Acknowledgments: The authors gratefully acknowledge support from Helen Kershaw and Kevin Raeder. HK and KR provided software support for both DART and CESM. The comments from two anonymous peer reviewers led to improvements in the clarity and depth of the presentation. We would like to acknowledge high-performance computing support from the Derecho system [103] provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APB	Autonomous pinniped bathythermograph
CAFE60v1	Climate Retrospective Analysis and Forecast Ensemble system, version 1
CESM	Community Earth System Model

DA	Data assimilation
DART	Data Assimilation Research Testbed
EAKF	Ensemble adjustment Kalman filter
EnKF	Ensemble Kalman filter
GM	Gent and McWilliams
MOM6	Modular Ocean Model, version 6
NorCPM	Norwegian Climate Prediction Model
OSSE	Observing System Simulation Experiment
psu	Practical salinity unit
RTPP	Relaxation-to-prior perturbations
RTPS	Relaxation-to-prior spread
S2S2D	Subseasonal to seasonal to decadal
SKEB	Stochastic kinetic energy backscatter
SPEAR	Seamless System for Prediction and Earth System Research
sQG	Surface quasi-geostrophic
SSS	Sea surface salinity
SST	Sea surface temperature
TAO	Tropical atmosphere ocean
TRITON	Triangle Trans-Ocean Buoy Network
WOA18	World Ocean Atlas 2018
WOD	World Ocean Database
WRF	Weather Research and Forecasting
XBT	Expendable bathythermograph

Appendix A

Proof of Theorem 1. The proof is closely related to standard analysis of hypothesis testing using Student's t -distribution. The sample mean $\bar{X}$ is distributed as

$$\bar{X} \sim N\left(\mu_e, \frac{\sigma_e^2}{N}\right). \quad (\text{A1})$$

The scaled sample variance,

$$U = \frac{N-1}{\sigma_e^2} V \quad (\text{A2})$$

is independent of $\bar{X}$ and has a chi-square distribution with $N-1$ degrees of freedom. Consequently, the quantity

$$T = \frac{\bar{X} - \mu}{\sqrt{V/N}} \quad (\text{A3})$$

has a Student's t -distribution with $N-1$ degrees of freedom.

Rearranging this expression yields

$$\bar{X} = \mu_e + T \sqrt{\frac{V}{N}} \quad (\text{A4})$$

which allows us to express R as

$$R = \frac{X_0 - \mu_t}{\sqrt{V}} - \frac{T}{\sqrt{N}} + \frac{\mu_t - \mu_e}{\sqrt{V}}. \quad (\text{A5})$$

By assumption, X_0 is independent of both T and V . Therefore, the expected value of R^2 is

$$\mathbb{E}[R^2] = \mathbb{E}\left[\frac{(X_0 - \mu_t)^2 + 2(X_0 - \mu_t)(\mu_t - \mu_e) + (\mu_t - \mu_e)^2}{V}\right] - \frac{2}{\sqrt{N}}\mathbb{E}\left[\frac{(X_0 - \mu_e)T}{\sqrt{V}}\right] + \frac{1}{N}\mathbb{E}[T^2]. \quad (\text{A6})$$

Independence of X_0 from T and V implies both that the second term is

$$\mathbb{E}\left[\frac{(X_0 - \mu_t)T}{\sqrt{V}}\right] = \mathbb{E}[X_0 - \mu_t]\mathbb{E}\left[\frac{T}{\sqrt{V}}\right] = 0, \quad (\text{A7})$$

and that the first term is

$$\mathbb{E}\left[\frac{(X_0 - \mu_t)^2 + 2(X_0 - \mu_t)(\mu_t - \mu_e) + (\mu_t - \mu_e)^2}{V}\right] = (\sigma_t^2 + (\mu_t - \mu_e)^2)\mathbb{E}\left[\frac{1}{V}\right]. \quad (\text{A8})$$

Because $U \sim \chi_{N-1}^2$, its multiplicative inverse has an inverse chi-square distribution, with $\mathbb{E}[U^{-1}] = 1/((N-1) - 2) = 1/(N-3)$. Thus,

$$\mathbb{E}\left[\frac{1}{V}\right] = \frac{N-1}{\sigma_e^2}\mathbb{E}\left[\frac{1}{U}\right] = \frac{N-1}{\sigma_e^2(N-3)} \quad (\text{A9})$$

and therefore the first term in Equation (A6) simplifies to

$$\frac{N-1}{N-3} \frac{\sigma_t^2 + (\mu_t - \mu_e)^2}{\sigma_e^2}. \quad (\text{A10})$$

Finally, recognizing that $\mathbb{E}[T^2] = \text{Var}[T] = (N-1)/(N-3)$ for a t -distribution with $N-1$ degrees of freedom, we obtain the expected value of R^2 :

$$\mathbb{E}[R^2] = \frac{N-1}{N-3} \left(\frac{\sigma_t^2 + (\mu_t - \mu_e)^2}{\sigma_e^2} + \frac{1}{N} \right). \quad (\text{A11})$$

□

References

1. Keppenne, C.L.; Rienecker, M.M.; Jacob, J.P.; Kovach, R. Error covariance modeling in the GMAO ocean ensemble Kalman filter. *Mon. Weather Rev.* **2008**, *136*, 2964–2982. [\[CrossRef\]](#)
2. Karspeck, A.R.; Yeager, S.; Danabasoglu, G.; Hoar, T.; Collins, N.; Raeder, K.; Anderson, J.; Tribbia, J. An ensemble adjustment Kalman filter for the CCSM4 ocean component. *J. Clim.* **2013**, *26*, 7392–7413. [\[CrossRef\]](#)
3. Penny, S.; Kalnay, E.; Carton, J.; Hunt, B.; Ide, K.; Miyoshi, T.; Chepurin, G. The local ensemble transform Kalman filter and the running-in-place algorithm applied to a global ocean general circulation model. *Nonlinear Process. Geophys.* **2013**, *20*, 1031–1046. [\[CrossRef\]](#)
4. Penny, S.G.; Behringer, D.W.; Carton, J.A.; Kalnay, E. A hybrid global ocean data assimilation system at NCEP. *Mon. Weather Rev.* **2015**, *143*, 4660–4677. [\[CrossRef\]](#)
5. Karspeck, A.R.; Danabasoglu, G.; Anderson, J.; Karol, S.; Collins, N.; Vertenstein, M.; Raeder, K.; Hoar, T.; Neale, R.; Edwards, J.; et al. A global coupled ensemble data assimilation system using the Community Earth System Model and the Data Assimilation Research Testbed. *Q. J. Roy. Meteor. Soc.* **2018**, *144*, 2404–2430. [\[CrossRef\]](#)
6. Lu, F.; Harrison, M.J.; Rosati, A.; Delworth, T.L.; Yang, X.; Cooke, W.F.; Jia, L.; McHugh, C.; Johnson, N.C.; Bushuk, M.; et al. GFDL's SPEAR seasonal prediction system: Initialization and ocean tendency adjustment (OTA) for coupled model predictions. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2020MS002149. [\[CrossRef\]](#)
7. O'Kane, T.J.; Sandery, P.A.; Kitsios, V.; Sakov, P.; Chamberlain, M.A.; Collier, M.A.; Fiedler, R.; Moore, T.S.; Chapman, C.C.; Sloyan, B.M.; et al. CAFE60v1: A 60-year large ensemble climate reanalysis. part I: System design, model configuration, and data assimilation. *J. Clim.* **2021**, *34*, 5153–5169. [\[CrossRef\]](#)

8. Qin, Y.; Yu, Q.; Wan, L.; Liu, Y.; Mo, H.; Wang, Y.; Meng, S.; Wu, X.; Sui, D.; Xie, J. A Global-Ocean-Data assimilation for operational oceanography. *J. Mar. Sci. Eng.* **2023**, *11*, 2255. [[CrossRef](#)]
9. Garcia-Oliva, L.; Counillon, F.; Bethke, I.; Keenlyside, N. Intercomparison of initialization methods for seasonal-to-decadal climate predictions with the NorCPM. *Clim. Dyn.* **2024**, *62*, 5425–5444. [[CrossRef](#)]
10. Wang, Y.; Counillon, F.; Svendsen, L.; Chiu, P.G.; Keenlyside, N.; Laloyaux, P.; Koseki, M.; de Boisseson, E. An ensemble-based coupled reanalysis of the climate from 1860 to the present (CoRea1860+). *Earth Syst. Sci. Data* **2025**, *17*, 4185–4211. [[CrossRef](#)]
11. Small, R.J.; Bryan, F.O.; Bishop, S.P.; Tomas, R.A. Air–sea turbulent heat fluxes in climate models and observational analyses: What drives their variability? *J. Clim.* **2019**, *32*, 2397–2421. [[CrossRef](#)]
12. Qiu, Z.; Qiao, F.; Jang, C.J.; Zhang, L.; Song, Z. Evaluation and projection of global marine heatwaves based on CMIP6 models. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **2021**, *194*, 104998. [[CrossRef](#)]
13. Agarwal, N.; Small, R.J.; Bryan, F.O.; Grooms, I.; Pegion, P.J. Impact of Stochastic Ocean Density Corrections on Air-Sea Flux Variability. *Geo. Res. Lett.* **2023**, *50*, e2023GL104248. [[CrossRef](#)]
14. Chen, H.; He, Z.; Xie, Q.; Zhuang, W. Performance of CMIP6 models in simulating the dynamic sea level: Mean and interannual variance. *Atmos. Ocean. Sci. Lett.* **2023**, *16*, 100288. [[CrossRef](#)]
15. Grooms, I.; Agarwal, N.; Marques, G.; Pegion, P.; Yassin, H. The stochastic GM+E closure: A framework for coupling stochastic backscatter with the Gent and McWilliams parameterization. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004560. [[CrossRef](#)]
16. Anderson, J.L.; Anderson, S.L. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **1999**, *127*, 2741–2758. [[CrossRef](#)]
17. Chelton, D.B.; Schlax, M.G.; Samelson, R.M. Global observations of nonlinear mesoscale eddies. *Prog. Oceanogr.* **2011**, *91*, 167–216. [[CrossRef](#)]
18. Pegliasco, C.; Delepoulle, A.; Mason, E.; Morrow, R.; Faugère, Y.; Dibarboure, G. META3.1exp: A new global mesoscale eddy trajectory atlas derived from altimetry. *Earth Syst. Sci. Data* **2022**, *14*, 1087–1107. [[CrossRef](#)]
19. Lombardi, B.; Grooms, I.; Kleiber, W. Demographics of mesoscale eddies in an eddy-permitting ocean model and reanalysis. *Prog. Oceanogr.* **2026**, *246*, 103765. [[CrossRef](#)]
20. Brankart, J.M. Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling. *Ocean Model.* **2013**, *66*, 64–76. [[CrossRef](#)]
21. Stanley, Z.; Grooms, I.; Kleiber, W.; Bachman, S.; Castruccio, F.; Adcroft, A. Parameterizing the Impact of Unresolved Temperature Variability on the Large-Scale Density Field: Part 1. Theory. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2020MS002185. [[CrossRef](#)]
22. Brankart, J.M.; Candille, G.; Garnier, F.; Calone, C.; Melet, A.; Bouttier, P.A.; Brasseur, P.; Verron, J. A generic approach to explicit simulation of uncertainty in the NEMO ocean model. *Geosci. Model Dev.* **2015**, *8*, 1285–1297. [[CrossRef](#)]
23. Kenigson, J.; Adcroft, A.; Bachman, S.; Castruccio, F.; Grooms, I.; Pegion, P.; Stanley, Z. Parameterizing the Impact of Unresolved Temperature Variability on the Large-Scale Density Field: 2. Modeling. *J. Adv. Model. Earth Syst.* **2022**, *14*, e2021MS002844. [[CrossRef](#)]
24. Buizza, R.; Milleer, M.; Palmer, T. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. Roy. Meteor. Soc.* **1999**, *125*, 2887–2908. [[CrossRef](#)]
25. Andrejczuk, M.; Cooper, F.; Juricke, S.; Palmer, T.; Weisheimer, A.; Zanna, L. Oceanic stochastic parameterizations in a seasonal forecast system. *Mon. Weather Rev.* **2016**, *144*, 1867–1875. [[CrossRef](#)]
26. Juricke, S.; Palmer, T.N.; Zanna, L. Stochastic Subgrid-Scale Ocean Mixing: Impacts on Low-Frequency Variability. *J. Clim.* **2017**, *30*, 4997–5019. [[CrossRef](#)]
27. Juricke, S.; MacLeod, D.; Weisheimer, A.; Zanna, L.; Palmer, T. Seasonal to annual ocean forecasting skill and the role of model and observational uncertainty. *Q. J. Roy. Meteor. Soc.* **2018**, *accepted manuscript*. [[CrossRef](#)] [[PubMed](#)]
28. Grooms, I.; Kleiber, W. Diagnosing, modeling, and testing a multiplicative stochastic Gent-McWilliams parameterization. *Ocean Model.* **2019**, *133*, 1–10. [[CrossRef](#)]
29. O’Kane, T.J.; Fiedler, R.; Collier, M.A.; Kitsios, V. Ocean model response to stochastically perturbed momentum fluxes. *J. Clim.* **2023**, *36*, 1895–1922. [[CrossRef](#)]
30. Eden, C.; Danilov, S. Backscatter by Negative Bolus Velocity. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2025MS005042. [[CrossRef](#)]
31. Whitaker, J.S.; Hamill, T.M. Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Weather Rev.* **2012**, *140*, 3078–3089. [[CrossRef](#)]
32. Berner, J.; Shutts, G.; Leutbecher, M.; Palmer, T. A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.* **2009**, *66*, 603–626. [[CrossRef](#)]
33. Grooms, I.; Lee, Y.; Majda, A.J. Ensemble filtering and low-resolution model error: Covariance inflation, stochastic parameterization, and model numerics. *Mon. Weather Rev.* **2015**, *143*, 3912–3924. [[CrossRef](#)]
34. Ha, S.; Berner, J.; Snyder, C. A comparison of model error representations in mesoscale ensemble data assimilation. *Mon. Weather Rev.* **2015**, *143*, 3893–3911. [[CrossRef](#)]

35. Anderson, J.L. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A Dyn. Meteorol. Oceanogr.* **2007**, *59*, 210–224. [\[CrossRef\]](#)
36. Anderson, J. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A Dyn. Meteorol. Oceanogr.* **2009**, *61*, 72–83. [\[CrossRef\]](#)
37. Dufée, B.; Mémin, E.; Crisan, D. Stochastic parametrization: An alternative to inflation in ensemble Kalman filters. *Q. J. Roy. Meteor. Soc.* **2022**, *148*, 1075–1091. [\[CrossRef\]](#)
38. Meehl, G.A.; Richter, J.H.; Teng, H.; Capotondi, A.; Cobb, K.; Doblas-Reyes, F.; Donat, M.G.; England, M.H.; Fyfe, J.C.; Han, W.; et al. Initialized Earth System prediction from subseasonal to decadal timescales. *Nat. Rev. Earth Environ.* **2021**, *2*, 340–357. [\[CrossRef\]](#)
39. Yang, Y.; Feng, J.; Zhu, Y.; Wei, Y.; Li, J. Comparison of Subseasonal Predictive Skill of the CFS and GEFS Operational Systems at NCEP and the Impact of Bias Correction. *Weather Forecast.* **2026**, *41*, 1227–1247. [\[CrossRef\]](#)
40. Palmer, T.N. Stochastic weather and climate models. *Nat. Rev. Phys.* **2019**, *1*, 463–471. [\[CrossRef\]](#)
41. Magnusson, L.; Alonso-Balmaseda, M.; Corti, S.; Molteni, F.; Stockdale, T. Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Clim. Dyn.* **2013**, *41*, 2393–2409.
42. Mouatadid, S.; Orenstein, P.; Flaspohler, G.; Cohen, J.; Oprescu, M.; Fraenkel, E.; Mackey, L. Adaptive bias correction for improved subseasonal forecasting. *Nat. Commun.* **2023**, *14*, 3482. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Yan, C.; Zhu, J. A simple bias correction scheme in ocean data assimilation. *J. Mar. Sci. Eng.* **2023**, *11*, 205. [\[CrossRef\]](#)
44. Muhamed Ali, A.; Zhuang, H.; Huang, Y.; Ibrahim, A.K.; Altaher, A.S.; Chérubin, L.M. Ocean currents velocity hindcast and forecast bias correction using a deep-learning approach. *J. Mar. Sci. Eng.* **2024**, *12*, 1680. [\[CrossRef\]](#)
45. Agarwal, N.; Amrhein, D.; Grooms, I. Cross-attractor transforms: Improving forecasts by learning optimal maps between dynamical systems and imperfect models. *Geo. Res. Lett.* **2025**, *52*, e2024GL110472. [\[CrossRef\]](#)
46. Adcroft, A.; Anderson, W.; Balaji, V.; Blanton, C.; Bushuk, M.; Dufour, C.O.; Dunne, J.P.; Griffies, S.M.; Hallberg, R.; Harrison, M.J.; et al. The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *J. Adv. Model. Earth Syst.* **2019**, *11*, 3167–3211. [\[CrossRef\]](#)
47. Danabasoglu, G.; Lamarque, J.F.; Bacmeister, J.; Bailey, D.; DuVivier, A.; Edwards, J.; Emmons, L.; Fasullo, J.; Garcia, R.; Gettelman, A.; et al. The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.* **2020**, *12*, e2019MS001916. [\[CrossRef\]](#)
48. Adcroft, A.; Campin, J.M. Rescaled height coordinates for accurate representation of free-surface flows in ocean circulation models. *Ocean Model.* **2004**, *7*, 269–284. [\[CrossRef\]](#)
49. Redi, M.H. Oceanic isopycnal mixing by coordinate rotation. *J. Phys. Ocean.* **1982**, *12*, 1154–1158. [\[CrossRef\]](#)
50. Gent, P.R.; McWilliams, J.C. Isopycnal mixing in ocean circulation models. *J. Phys. Ocean.* **1990**, *20*, 150–155. [\[CrossRef\]](#)
51. Gent, P.R.; Willebrand, J.; McDougall, T.J.; McWilliams, J.C. Parameterizing eddy-induced tracer transports in ocean circulation models. *J. Phys. Ocean.* **1995**, *25*, 463–474. [\[CrossRef\]](#)
52. Marshall, D.P.; Maddison, J.R.; Berloff, P.S. A framework for parameterizing eddy potential vorticity fluxes. *J. Phys. Ocean.* **2012**, *42*, 539–557. [\[CrossRef\]](#)
53. Jansen, M.F.; Adcroft, A.; Khani, S.; Kong, H. Toward an energetically consistent, resolution aware parameterization of ocean mesoscale eddies. *J. Adv. Model. Earth Syst.* **2019**, *11*, 2844–2860. [\[CrossRef\]](#)
54. Grooms, I. Simulations of eddy kinetic energy transport in barotropic turbulence. *Phys. Rev. Fluids* **2017**, *2*, 113801. [\[CrossRef\]](#)
55. Bodner, A.S.; Fox-Kemper, B.; Johnson, L.; Van Roekel, L.P.; McWilliams, J.C.; Sullivan, P.P.; Hall, P.S.; Dong, J. Modifying the mixed layer eddy parameterization to include frontogenesis arrest by boundary layer turbulence. *J. Phys. Ocean.* **2023**, *53*, 323–339. [\[CrossRef\]](#)
56. Grooms, I. Backscatter in energetically-constrained Leith parameterizations. *Ocean Model.* **2023**, *186*, 102265. [\[CrossRef\]](#)
57. Yassin, H.; Marques, G.; Grooms, I. Balancing backscatter and diffusion in a 1/4° forced global ocean model. *J. Adv. Model. Earth Syst.* **2026**, *18*, e2025MS005462. [\[CrossRef\]](#)
58. Leith, C. Stochastic models of chaotic systems. *Phys. D* **1996**, *98*, 481–491. [\[CrossRef\]](#)
59. Fox-Kemper, B.; Menemenlis, D. Can Large Eddy Simulation Techniques Improve Mesoscale Rich Ocean Models? In *Ocean Modeling in an Eddying Regime*; American Geophysical Union (AGU): San Francisco, CA, USA, 2008; pp. 319–337. [\[CrossRef\]](#)
60. Tsujino, H.; Urakawa, S.; Nakano, H.; Small, R.J.; Kim, W.M.; Yeager, S.G.; Danabasoglu, G.; Suzuki, T.; Bamber, J.L.; Bentsen, M.; et al. JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Model.* **2018**, *130*, 79–139. [\[CrossRef\]](#)
61. Locarnini, R.A.; Mishonov, A.V.; Baranova, O.K.; Boyer, T.P.; Zweng, M.M.; Garcia, H.E.; Reagan, J.R.; Seidov, D.; Weathers, K.W.; Paver, C.R.; et al. *World Ocean Atlas 2018, Volume 1: Temperature*; National Centers for Environmental Information: Silver Spring, MD, USA, 2018.
62. Zweng, M.; Reagan, J.; Seidov, D.; Boyer, T.; Locarnini, M.; Garcia, H.; Mishonov, A.; Baranova, O.; Weathers, K.; Paver, C.; et al. *World Ocean Atlas 2018, Volume 2: Salinity*; National Centers for Environmental Information: Silver Spring, MD, USA, 2019; Volume 2.

63. Reynolds, R.W.; Rayner, N.A.; Smith, T.M.; Stokes, D.C.; Wang, W. An improved in situ and satellite SST analysis for climate. *J. Clim.* **2002**, *15*, 1609–1625. [\[CrossRef\]](#)
64. Grooms, I.; Loose, N.; Abernathey, R.; Steinberg, J.; Bachman, S.D.; Marques, G.; Guillaumin, A.P.; Yankovsky, E. Diffusion-based smoothers for spatial filtering of gridded geophysical data. *J. Adv. Model. Earth Syst.* **2021**, *13*, e2021MS002552. [\[CrossRef\]](#)
65. Loose, N.; Abernathey, R.; Grooms, I.; Busecke, J.; Guillaumin, A.; Yankovsky, E.; Marques, G.; Steinberg, J.; Ross, A.S.; Khatri, H.; et al. GCM-Filters: A Python Package for Diffusion-based Spatial Filtering of Gridded Data. *J. Open Source Softw.* **2022**, *7*, 3947. [\[CrossRef\]](#)
66. Karspeck, A.R. An ensemble approach for the estimation of observational error illustrated for a nominal 1° global ocean model. *Mon. Weather Rev.* **2016**, *144*, 1713–1728. [\[CrossRef\]](#)
67. McDougall, T.J.; Barker, P.M.; Holmes, R.M.; Pawlowicz, R.; Griffies, S.M.; Durack, P.J. The interpretation of temperature and salinity variables in numerical ocean model output and the calculation of heat fluxes and heat content. *Geosci. Model Dev.* **2021**, *14*, 6445–6466. [\[CrossRef\]](#)
68. Deng, S.; Shen, Z.; Chen, S.; Wang, R. Comparison of Perturbation Strategies for the Initial Ensemble in Ocean Data Assimilation with a Fully Coupled Earth System Model. *J. Mar. Sci. Eng.* **2022**, *10*, 412. [\[CrossRef\]](#)
69. Gao, Y. Assessment of inflation schemes on parameter estimation and their application in ENSO prediction in an OSSE framework. *J. Mar. Sci. Eng.* **2023**, *11*, 2003. [\[CrossRef\]](#)
70. Zhao, H.; Zhao, Y.; Deng, X. Application of the most sensitive initial error in sampling scheme design for a mobile platform. *J. Mar. Sci. Eng.* **2023**, *11*, 2080. [\[CrossRef\]](#)
71. El Gharamti, M.; Kershaw, H.; Raeder, K.; Raczka, B.; Johnson, B.; Smith, M.; Anderson, J.; Amrhein, D.; Collins, N.; Hoar, T.; et al. The Data Assimilation Research Testbed: A Robust, Scalable Software Facility with Groundbreaking Capabilities for Model-Data Integration. *Bull. Am. Meteorol. Soc.* **2025**, *106*, E2328–E2345. [\[CrossRef\]](#)
72. Anderson, J.L. A local least squares framework for ensemble filtering. *Mon. Weather Rev.* **2003**, *131*, 634–642. [\[CrossRef\]](#)
73. Anderson, J.L. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **2001**, *129*, 2884–2903. [\[CrossRef\]](#)
74. Wang, R.; Shen, Z. A Deep Neural Network-Ensemble Adjustment Kalman Filter and Its Application on Strongly Coupled Data Assimilation. *J. Mar. Sci. Eng.* **2024**, *12*, 108. [\[CrossRef\]](#)
75. Zhang, S.; Rosati, A. An inflated ensemble filter for ocean data assimilation with a biased coupled GCM. *Mon. Weather Rev.* **2010**, *138*, 3905–3931. [\[CrossRef\]](#)
76. Nadiga, B.T.; Casper, W.R.; Jones, P.W. Ensemble-based global ocean data assimilation. *Ocean Model.* **2013**, *72*, 210–230. [\[CrossRef\]](#)
77. El Gharamti, M. Enhanced adaptive inflation algorithm for ensemble filters. *Mon. Weather Rev.* **2018**, *146*, 623–640. [\[CrossRef\]](#)
78. Su, A.; Zhang, L.; Zhang, X.; Zhang, S.; Liu, Z.; Liu, C.; Zhang, A. A New Scheme of Adaptive Covariance Inflation for Ensemble Filtering Data Assimilation. *J. Mar. Sci. Eng.* **2021**, *9*, 54. [\[CrossRef\]](#)
79. El Gharamti, M.; Raeder, K.; Anderson, J.; Wang, X. Comparing adaptive prior and posterior inflation for ensemble filters using an atmospheric general circulation model. *Mon. Weather Rev.* **2019**, *147*, 2535–2553. [\[CrossRef\]](#)
80. Chen, Y.; Shen, Z.; Tang, Y. On oceanic initial state errors in the ensemble data assimilation for a coupled general circulation model. *J. Adv. Model. Earth Syst.* **2022**, *14*, e2022MS003106. [\[CrossRef\]](#)
81. Tang, Q.; Mu, L.; Goessling, H.; Semmler, T.; Nerger, L. Strongly coupled data assimilation of ocean observations into an ocean-atmosphere model. *Geo. Res. Lett.* **2021**, *48*, e2021GL094941. [\[CrossRef\]](#)
82. Anderson, J.L. Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Weather Rev.* **2012**, *140*, 2359–2371. [\[CrossRef\]](#)
83. Anderson, J.L. Reducing correlation sampling error in ensemble Kalman filter data assimilation. *Mon. Weather Rev.* **2016**, *144*, 913–925. [\[CrossRef\]](#)
84. Xing, X.; Liu, B.; Zhang, W.; Wu, J.; Cao, X.; Huang, Q. An Investigation of Adaptive Radius for the Covariance Localization in Ensemble Data Assimilation. *J. Mar. Sci. Eng.* **2021**, *9*, 1156. [\[CrossRef\]](#)
85. Zingerle, C.; Nurmi, P. Monitoring and verifying cloud forecasts originating from operational numerical models. *Meteorol. Appl. A J. Forecast. Pract. Appl. Train. Tech. Model.* **2008**, *15*, 325–330. [\[CrossRef\]](#)
86. 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.* **2020**, *16*, 831–845. [\[CrossRef\]](#)
87. Shapiro, G.I.; Salim, M. How efficient is model-to-model data assimilation at mitigating atmospheric forcing errors in a regional ocean model? *J. Mar. Sci. Eng.* **2023**, *11*, 935. [\[CrossRef\]](#)
88. Shapiro, G.I.; Gonzalez-Ondina, J.M.; Salim, M.; Tu, J. A comparison of stochastic and deterministic downscaling in eddy resolving ocean modelling: The lakshadweep sea case study. *J. Mar. Sci. Eng.* **2023**, *11*, 363. [\[CrossRef\]](#)
89. Dirkson, A.; Buehner, M. Are we misdiagnosing ensemble forecast reliability? On the insufficiency of spread–error and rank-based reliability metrics. *Q. J. Roy. Meteor. Soc.* **2026**, e70186.
90. Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **2000**, *15*, 559–570. [\[CrossRef\]](#)

91. Anderson, J.L. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **1996**, *9*, 1518–1530. [[CrossRef](#)]
92. Dimitriadis, T.; Gneiting, T.; Jordan, A.I. Stable reliability diagrams for probabilistic classifiers. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016191118. [[CrossRef](#)] [[PubMed](#)]
93. Dirkson, A.; Buehner, M. The Effect of Ensemble Size on the Mean Squared Error and Spread–Error Relationship. *Mon. Weather Rev.* **2025**, *153*, 1219–1229. [[CrossRef](#)]
94. Klenke, A. Moments and Laws of Large Numbers. In *Probability Theory: A Comprehensive Course*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 113–146.
95. Williams, P.D. Climatic impacts of stochastic fluctuations in air–sea fluxes. *Geo. Res. Lett.* **2012**, *39*, L10705. [[CrossRef](#)]
96. Bessac, J.; Monahan, A.H.; Christensen, H.M.; Weitzel, N. Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. *Mon. Weather Rev.* **2019**, *147*, 1447–1469. [[CrossRef](#)]
97. Belyaev, K.P.; Gorshenin, A.K.; Korolev, V.Y.; Osipova, A.A. Comparison of statistical approaches for reconstructing random coefficients in the problem of stochastic modeling of air–sea heat flux increments. *Mathematics* **2024**, *12*, 288. [[CrossRef](#)]
98. Hu, M.; Xie, L.; Li, M.; Zheng, Q.; Zeng, F.; Chen, X. An Improved MG Model for Turbulent Mixing Parameterization in the Northwestern South China Sea. *J. Mar. Sci. Eng.* **2024**, *13*, 46. [[CrossRef](#)]
99. Wu, J.; Perezhogin, P.; Gagne, D.J.; Reichl, B.G.; Subramanian, A.C.; Thompson, E.J.; Zanna, L. Data-driven probabilistic air-sea flux parameterization. *Geo. Res. Lett.* **2026**, *53*, e2025GL120472. [[CrossRef](#)]
100. Hallberg, R.; Adcroft, A.; Ward, M.; Marques, G.; Hedstrom, K.; Altuntas, A.; Harrison, M.; Zadeh, N.; Griffies, S.; Pegion, P.; et al. houssamyassin/MOM6: V1.0.0. 2025. Available online: <https://zenodo.org/records/17901673> (accessed on 17 January 2026).
101. Computational and Information Systems Laboratory. The Data Assimilation Research Testbed, Version 11.19.2; 2025. UCAR/NSF NCAR/CISL/DAReS. Available online: <https://github.com/NCAR/DART> (accessed on 17 January 2026).
102. Grooms, I. Data and Scripts for Beyond Inflation: Backscatter Parameterizations to Address the Variability Deficit in Global Ocean Data Assimilation. 2026. Available online: https://figshare.com/articles/dataset/Data_and_Scripts_for_Beyond_Inflation_Backscatter_Parameterizations_to_Address_the_Variability_Deficit_in_Global_Ocean_Data_Assimilation/32811074/1?file=65990888 (accessed on 17 January 2026).
103. Computational and Information Systems Laboratory. *Derecho: HPE Cray EX System (NCAR Community Computing)*; National Center for Atmospheric Research: Boulder, CO, USA, 2023. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.