Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/7532--19-9928

Earth System Prediction Capability (ESPC) Initial Operational Capability (IOC) Ensemble System

Neil Barton
Matthew Janiga
Justin McLay
Carolyn Reynolds

Naval Research Laboratory
Monterey, CA

Clark Rowley
Patrick Hogan
Prassad Thoppil

Naval Research Laboratory
Stennis Space Center, MS

October 3, 2019

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
03-10-2019	NRL Memorandum Report	

4. TITLE AND SUBTITLE

Earth System Prediction Capability (ESPC) Initial Operational Capability (IOC) Ensemble system

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
75-4726-N9-5

6. AUTHOR(S)

Neil Barton, Clark Rowley*, Patrick Hogan*, Matthew Janiga, Justin McLay, Carolyn Reynolds, and Prassad Thoppil*

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
4726

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
1 University Circle
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/7532--19-9928

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
One Liberty Center
875 North Randolph Street, Suite 1425
Arlington, VA 22203-1995

10. SPONSOR / MONITOR'S ACRONYM(S)

ONR

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

*Naval Research Laboratory, 1005 Balch Boulevard, Stennis Space Center, MS 39529

14. ABSTRACT

The Navy Earth System Prediction Capability (Navy-ESPC) is the new global coupled atmosphere-ocean-sea ice prediction system developed at the Naval Research Laboratory (NRL) for operational forecasting for timescales of days to subseasonal. This report describes and evaluates the Navy-ESPC model. When the Navy-ESPC model becomes operational, it will be the first time NRL's operational partner, Fleet Numerical Meteorology and Oceanography Center, will provide coupled atmosphere ocean sea-ice forecasts, atmosphere forecasts past 16 days, and ocean and sea ice ensemble forecasts. A unique aspect of the Navy-ESPC model is that the global ocean model is eddy resolving at 1/12° in the ensemble and at 1/25° in high-resolution deterministic configurations. The component models consist of currently operational systems and include: atmosphere - NAVy Global Environmental Model (NAVGEM); ocean - HYbrid Coordinate Ocean Model (HYCOM); and sea ice - Community Ice CodE (CICE). Physics updates to improve the simulation of equatorial phenomena, particularly the Madden Julian Oscillation (MJO), and to provide more consistent air-sea fluxes, were introduced into NAVGEM. No physics updates to accommodate coupling were included in HYCOM or CICE. Data assimilation is loosely coupled between the NRL Atmospheric Variational Data Assimilation System - Accelerated Representer (NAVDAS-AR) for the atmosphere and the Navy Coupled Ocean Data Assimilation (NCODA) for the ocean/ice components. A 16 member ensemble configuration and high-resolution deterministic configuration are evaluated based on analyses and forecasts run from February 2017 to January 2018. When compared to other operational centers, the Navy-ESPC model performs in a similar manner when examining large-scale atmospheric characteristics, such as the MJO, North Atlantic Oscillation (NAO), Antarctic Oscillation (AAO), and other indices. Forecasts of ocean sea surface temperatures perform better than climatology in the tropics out to 60 days. In addition, the Navy-ESPC Pan-Arctic and Pan-Antarctic sea ice extent predictions perform better than climatology out to about 45 days, although the skill is dependent on season.

15. SUBJECT TERMS

Navy-ESPC NAVGEM HYCOM CICE
Ensemble Global coupled forecasting Model validation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Neil Barton
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unclassified Unlimited	73	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited			(831) 656-4733

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

1. Introduction

Environmental situational awareness is a key element of the Naval Research and Development Framework to ensure battlespace dominance in the 21st century (https://www.onr.navy.mil/our-research/naval-research-framework). High fidelity forecasts of the global environmental state are critical to a wide range of naval operations including mission planning, accurate delivery of precision ordnance, theater air and missile defense, communications, and security. For extended-range timescales, a global ensemble forecast system (EFS) is necessary to produce skillful forecasts and information essential to risk assessment, such as forecast reliability and uncertainty.

Here in this validation test report (VTR) we evaluate the first global coupled EFS developed for Navy operations for forecasts from 0 to 60 days.

2. Navy-ESPC Description

2.1. Forecast Model

The Navy Earth System Prediction Capability (Navy-ESPC) model is built upon global models that currently run operationally in stand-alone configurations. The NAVy Global Environmental Model (NAVGEM) (Hogan et al. 2014) is the global atmosphere model in Navy-ESPC. The ocean and sea ice models in Navy-ESPC are part of the operational Global Ocean Forecasting System version 3.1 (GOFS 3.1) (Metzger et al. 2014). GOFS 3.1 is an ocean and sea ice coupled model consisting of the HYbrid Coordinate Ocean Model (HYCOM) for ocean and the Community Ice CodE (CICE) for sea ice. These models, which are also referred to as components, are coupled together using Earth System Module Framework (ESMF) tools used in conjunction with the National Unified Operational Prediction Capability (NJUOPC) layer (Theurich et al. 2016). Figure 1 shows a schematic of the Initial Operational Capability (IOC) configuration of the Navy-ESPC system as tested in this VTR.

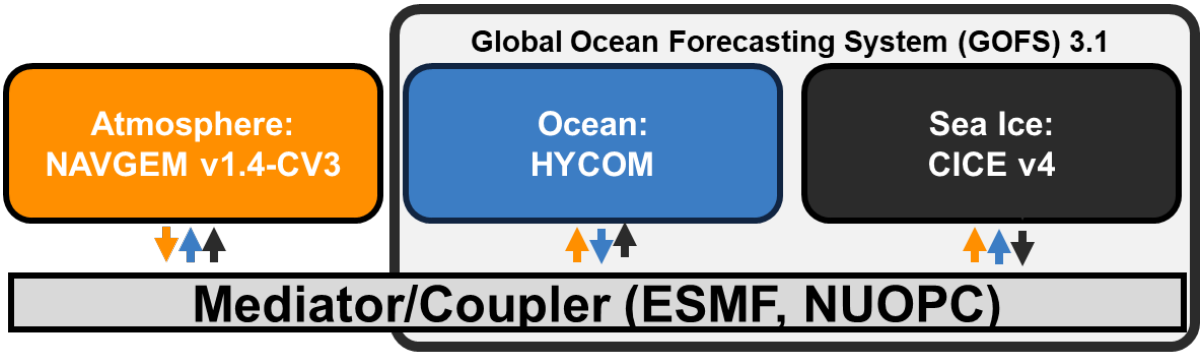

Figure 1: Schematic of the Navy-ESPC coupled system that will be used for Initial Operational Capability (IOC). The colored arrows represent variables being passed to and from each component (e.g., NAVGEM, HYCOM, CICE) to and from the mediator. The mediator uses the NUOPC standards on top of the ESMF tools. NAVGEM v1.4-CV3 is NAVGEM v1.4 with changes described in section 2.1.3. The GOFS 3.1 model is HYCOM and CICE coupled without NAVGEM.

2.1.1. HYCOM

Manuscript approved August 29, 2019.

The HYbrid Coordinate Ocean Model (HYCOM) and the Community Ice CodE (CICE) are coupled for the operational GOFS 3.1 model (Metzger et al. 2014). HYCOM is a primitive equation ocean general circulation model capable of nowcasting and forecasting the 3-dimensional temperature, salinity and current structure of the global ocean. It employs potential density referenced to 2000 m and includes the effects of thermobaricity (Chassignet et al. 2003). The HYCOM horizontal grid is uniform cylindrical from 78.64°S to 66.0°S , on a Mercator projection from 66.0°S to 47°N, and curvilinear north of 47°N as it employs an Arctic dipole patch by which the poles are shifted over land to avoid a singularity at the North Pole. Note, there is no grid south of 78.64°S due to the land mask. This grid is referred to as a tripole grid.

HYCOM is capable of combining different techniques to define the vertical coordinate (hence the name HYbrid Coordinate). Vertical coordinates can be (1) isopycnal layers (density tracking), often the best coordinate in the deep stratified ocean; (2) levels of equal pressure (nearly fixed depths), best used in the mixed layer; and (3) unstratified ocean and sigma-levels (terrain-following), often the best choice in shallow water. HYCOM combines all three approaches by choosing the optimal distribution at every time step. The model makes a dynamically smooth transition between coordinate types by using the layered continuity equation. The hybrid coordinate extends the geographic range of applicability of traditional isopycnic coordinate circulation models into shallow coastal seas and weakly stratified parts of the world ocean. It maintains the significant advantages of an isopycnal model in stratified regions while allowing more vertical resolution near the surface and in shallow coastal areas, hence providing a better representation of the upper ocean physics.

HYCOM is configured with options for a variety of mixed layer sub-models (Halliwell 2004), and a more complete description of HYCOM physics can be found in Bleck (2002) and an application of global HYCOM within the Indonesian Sea can be found in Metzger et al. (2010).

2.1.2. CICE

The Los Alamos-developed CICE model (Hunke and Lipscomb 2008) is the sea ice component of Navy-ESPC. CICE is also currently used in GOFS 3.1. The horizontal grid of CICE is on the same grid as HYCOM for conservation of variables when coupling. CICE includes sophisticated ice thermodynamics such as multiple ice thickness layers, multiple snow layers and the capability to forecast multi-categories of ice thickness according to World Meteorological Organization definitions. In addition, CICE has several interacting components including a thermodynamic model that computes local growth rates of snow and ice due to snowfall; vertical conductive, radiative and turbulent fluxes; a model of ice dynamics that predicts the velocity field of the ice pack based on a model of the material strength of the ice; a transport model that describes advection of the areal concentration, ice volumes and other state variables; and a ridging parameterization that transfers ice among thickness categories based on energetic balances and rates of strains. The CICE version number delivered for IOC is 4.0.

2.1.3. NAVGEM

Description

NAVGEM (Hogan et al. 2014) is the Navy's high-resolution global weather prediction system, run operationally at the Fleet Numerical Meteorology and Oceanography Center (FNMOC) and is used as the atmospheric component in Navy-ESPC. The forecast model uses both grid point and spectral (i.e., spherical harmonic) representations to perform the forecast. Grid point calculations are performed for

the Semi-Lagrangian (SL) advection and in all physical parameterizations. Calculations in spectral space are performed for the Semi-Implicit (SI) corrections to the divergent component of the winds, virtual potential temperature, and surface pressure. The horizontal computational grid of NAVGEM is a quadratic Gaussian grid, which for T359, results in 1080 x 540 grid points and a horizontal grid point resolution of about 37 km. The T359 grid is used in the Navy-ESPC IOC ensemble. The model has 60 vertical levels, with a model top of 0.04 hPa, which is approximately 70 km above sea-level. The vertical coordinate is a hybrid pressure coordinate, which is terrain-following in the troposphere then smoothly transitions to a pure pressure coordinate at 85 hPa, with 20 isobaric levels between 85 hPa and 0.04 hPa (Eckermann 2009; Eckermann et al. 2014).

The dynamical core of NAVGEM is a 3-time level, SL/SI numerical integration of the hydrostatic equations of motion and the first law of thermodynamics. The dynamical variables are the east-west wind, the north-south wind, the virtual potential temperature, the specific humidity, the surface pressure, ozone, cloud liquid and ice water. The SL/SI formulation is based on the work of Ritchie (1987, 1988, 1991); Ritchie et al. (1995). This is described in a European Centre for Medium-range Weather Forecasts (ECMWF) documentation manual, (http://www.ecmwf.int/research/ifsdocs/CY36r1/index.html). The major difference between the Ritchie and NAVGEM formulations is that NAVGEM uses potential temperature while ECMWF uses temperature.

Spectral representations of vorticity and divergence are computed from the horizontal winds. The hydrostatic approximation, which is accurate down to 10 km horizontal resolutions, is assumed and used to calculate geopotential heights and vertical motion. Rain/snow rates are computed from the stratiform and cumulus parameterizations. For stratiform clouds, cloud fraction is computed based on the relative humidity, vertical motion, and lapse rate (Teixeira and Hogan 2002) assuming cloud water is present. For cumulus clouds, the cloud fraction is computed based on the cumulus precipitation reaching the ground and the parameterized cloud base mass flux.

Operational NAVGEM includes: orographic gravity-wave and flow-blocking drag (Webster et al. 2003), the EDMF vertical mixing (Louis et al. 1982; Suselj et al. 2013; Suselj et al. 2014), the Simplified Arakawa-Schubert (SAS) cumulus parameterization (Moorthi et al. 2001), a shallow cumulus parameterization (Han and Pan 2011), a convective cloud fraction parameterization based on a cloud base mass flux scaling modification of the Slingo (1987) scheme, a stratiform cloud fraction parameterization (Slingo 1987; Teixeira and Hogan 2002), a cloud water parameterization based on an extension of the scheme of Zhao and Carr (1997) to include prognostic representations of both cloud liquid and cloud ice, the Rapid Radiative Transfer Model for General Circulation Models (RRTMG) for solar and longwave radiation fluxes (Clough et al. 2005), a land surface parameterization (Hogan 2007), and ozone photochemistry (McCormack et al. 2006).

In addition to the atmospheric variables, canopy temperature, ground temperature, ground liquid water, and ground ice water are computed down to a depth of 2 meters into the ground surface (Hogan 2007).

Physics Updates for Navy-ESPC

The implementation of the NAVGEM forecast model in the Navy-ESPC system retains most of the features of the forecast model in NAVGEM 1.4.3, which is the current stand-alone configuration at FNMOC. An important exception being the addition of a new suite of model physics referred to as

Coupled Version Physics (CVP) (Table 1). The CVP physics has been tailored for the coupled system with two primary objectives: 1) to improve the representation of the Madden Julian Oscillation (MJO), which is recognized as a key contributor to extended range predictability, and 2) to improve the consistency between surface fluxes computed in NAVGEM and in the ocean model HYCOM. The physic differences are largely associated with atmospheric convection and air-sea fluxes.

Representation of the MJO in the coupled system is enhanced significantly through the implementation of a new treatment of atmospheric convection as part of the CVP. This modified version of the Kain-Fritsch scheme (Kain and Fritsch 1990; Kain and Fritsch 1993) incorporates both turbulence- and dynamically forced modes, and is an extension of the work of Ridout et al. (2005). As in the previous version, the new scheme includes a modified closure formulation for the Kain-Fritsch dynamically forced mode based on an assumed quasi-balance of updraft parcel buoyancy at the cloud base level. This closure formulation requires the scheme to be called at every time step to adjust the cloud base mass flux in a similar manner as in the Emanuel convection scheme (Emanuel 1991; Emanuel and Zivkovic-Rothman 1999). The scheme similarly includes from the 2005 version a modified updraft source level selection procedure, which was adopted from the implementation of the Emanuel scheme developed by Peng et al. (2004), which became the operational NOGAPS version of the scheme. Several other enhancements have been added for the current scheme, including addition of a treatment of convective momentum transport, a modified representation of the rate of updraft-environment mixing adapted from Peng et al. (2004), and a modified cloud top condition that enhances the sensitivity of convection to dry layers and associated thermal inversions. Convective momentum transport is treated using the adaptation of the treatment of Gregory et al. (1997). The turbulence-forced mode of convection is represented using a slightly modified version of the scheme employed for the dynamically forced mode. For this mode the precipitating downdraft code is turned off, and the cloud base mass flux is parameterized based on the mass flux at the lifting condensation level of boundary layer plumes in a slightly modified version of the NAVGEM EDMF scheme. Triggering of the turbulence-forced mode by the plumes is represented using the mixed-layer Richardson number convective trigger formulation described by Ridout and Reynolds (1998).

In regards to surface fluxes in the coupled system, issues with consistency between the component models arise in part due to the hourly exchange of variables (i.e., coupling frequency) currently implemented between NAVGEM and HYCOM and CICE. Given this coupling frequency, to provide for sufficiently responsive updates to the surface fluxes over the ocean in NAVGEM and HYCOM, rather than exchange fluxes between the models, each model computes its own fluxes, with NAVGEM providing 10-m winds and near surface (2-m) temperature and moisture to HYCOM, and HYCOM providing sea surface temperatures (SSTs) and surface currents to NAVGEM. The CVP helps to mitigate resultant differences in surface fluxes felt by the ocean and atmosphere by implementing the HYCOM surface flux scheme in NAVGEM. This scheme is an adaptation Kara et al. (2005) of the COARE 3.0 scheme of Fairall et al. (2003). Notably, the coupled system framework enables for the first time the inclusion of the surface current contribution to the near-surface shear in the NAVGEM surface flux computation. This treatment is consistent not only with the implementation of the COARE 3.0 scheme in HYCOM, but also with the development of the scheme itself, in which surface current corrections were made to the near surface wind speed (Fairall et al. 2003).

As noted, Navy ESPC coupled system modifies the current standalone version of NAVGEM (see summary in Table 1). In addition, within the ESPC IOC there are differences between the NAVGEM version used deterministically, and that used for ensemble forecasts (i.e., as discussed in this VTR). Resolution differences between the deterministic and ensemble versions are listed in Table 2. There are also slight

physic changes between these resolutions. The dynamics framework utilized for the deterministic (T681L60) (18 km equivalent horizontal grid) configuration follows that of NAVGEM 1.4.3 in its use of the same three-time level semi-Lagrangian, semi-implicit dynamical core, perturbation virtual potential temperature for the prognostic temperature variable and use of the reduced Gaussian grid. The ensemble (T359L60) (37 km equivalent horizontal grid) configuration follows NAVGEM 1.2 (T359L50) in retaining the virtual potential temperature as the prognostic temperature variable, and using the full Gaussian grid. The adiabatic correction option introduced in NAVGEM 1.2.2 is turned off for the ensemble resolution. Configurations for both NAVGEM resolutions used in Navy-ESPC retain the 60-level vertical grid used in NAVGEM versions 1.3 – 1.4.3.

The NAVGEM version used in the Navy-ESPC coupled system benefits from the CVP, which improved the consistency between surface fluxes in the coupled system and improved the representation of the MJO. Prior to this physics change, NAVGEM was not competitive with the best MJO forecast models as shown in the model intercomparision study of Jiang et al. (2015). The CVP led to a very significant improvement in the MJO such that the Navy-ESPC model compares favorably to other state-of-the-art forecast models in representation of the MJO as seen through evaluation of the 17 years of Navy-ESPC forecasts produced for the NOAA SubX project (Janiga et al. 2018).

The CVP suite has a default setting and optional versions that can be specified by setting a new environmental variable "CVP_TYPE". The current default setting is "CVP_TYPE=3", which selects the version of the suite used for the current VTR testing. Setting "CVP_TYPE=2" selects the physics code that is being used in the Navy-ESPC coupled system for the ongoing North American Multi-Model Ensemble for the Subseasonal eXperiment (SubX) runs, and "CVP_TYPE=0" selects the NAVGEM 1.4.3 version physics.

Table 1: Summary of NAVGEM changes between the NAVGEM v1.4 version and the NAVGEM v1.4-CV3 used in the Navy-ESPC system. The main changes between the two atmospheric models include changes in the convection parameterization and boundary layer scheme.

Parameterization/Scheme	NAVGEM v1.4	NAVGEM v1.4-CV3
Convection Parameterization	SAS (Moorthi et al. 2001)	Modified Kain- Fritsch
Boundary Layer Scheme	Louis et al. (1982)	COARE (Kara et al. 2005)
Prognostic Temperature Variable	Perturbation Virtual Potential Temperature	Virtual Potential Temperature
Grid	Thinned Gaussian	Full Gaussian
Adiabatic Correction	Used	Turned off for ensemble resolution

Table 2: Resolution of the Navy-ESPC components. NAVGEM's T0359 resolution (approximately 37 km). HYCOM and CICE are on the same tripole grid, and the 1/12° degree grid corresponds to about a 9km resolution south of 60°N and about a 4 km resolution in the Arctic Ocean.

Model Run	NAVGEM	HYCOM	CICE
Navy-ESPC Ensemble	T0359 L60	1/12° tripole grid with 41 layers	1/12° tripole grid with 4 ice categories
Navy-ESPC Deterministic	T0681 L60	1/25° tripole grid with 41 layers	1/25° tripole grid with 4 ice categories

2.1.4. Coupling

The Navy-ESPC system is coupled via the Earth System Modeling Framework (ESMF) (Theurich et al. 2016). This framework allows the components to dynamically interact at a specified coupling time step allowing the different components to adjust for the processes described by the various components. Furthermore, by keeping these large data sets in-memory there is reduction of the file I/O in the whole system compared to running each component separately. As mentioned above, a one hour coupling time step is implemented in Navy-ESPC. A more frequent time step was briefly tested, and the results did not change significantly. Note the stand-alone atmospheric system NAVGEM typically holds the ocean surface temperature and sea-ice fixed over a given forecast period. So the coupled system provides a substantial improvement to this forcing particularly for forecasts longer than a few days. A summary of the fields passed is in Figure 2.

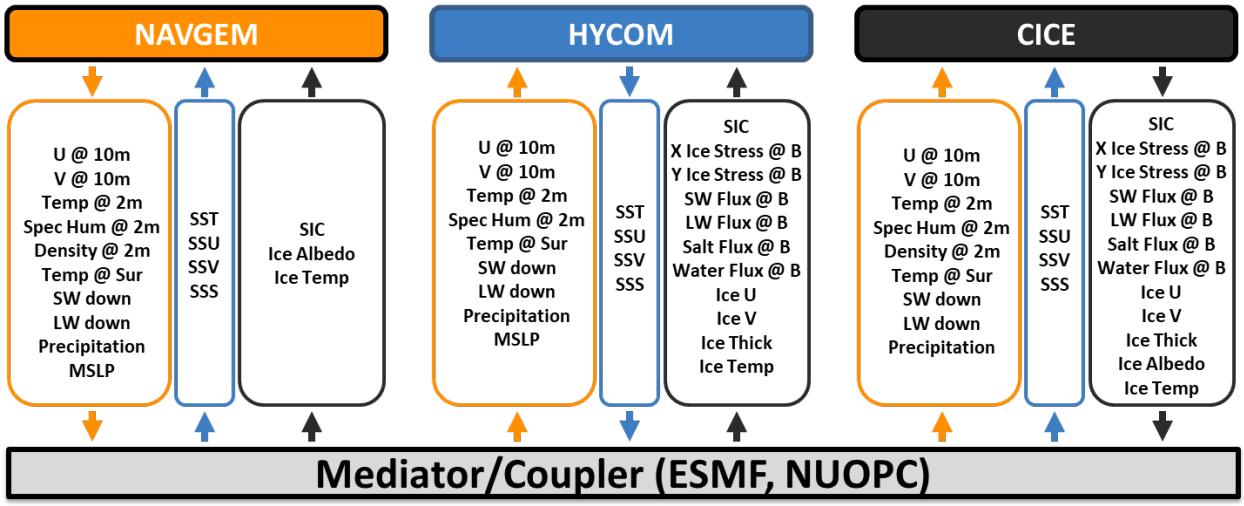

Figure 2: List of fields passed to and from the components (i.e., variables that are coupled). The color represents the component/model the field is from and the arrows represent if the fields are going into or out of the component. Long name for the atmospheric fields include: U @ 10m-> zonal winds at 10 meters, V @ 10m -> meridional winds at 10 meters , Temp @ 2m -> air temperature at 2 meters, Spec Hum @ 2m -> specific humidity at 2 meters, Density @ 2m -> air density at two meters, Temp @ Sur -> temperature of the surface (ground, ice, ocean), SW down -> downward surface shortwave radiative flux, LW down -> downward surface longwave radiative flux, Precipitation -> total precipitation, MSLP -> mean sea level pressure. Long names for the ocean variables include: SST -> sea surface temperature, SSU -> zonal sea surface current, SSV -> meridional sea surface current, SSS -> sea surface salinity. Long names for the sea ice variables include: X Ice Stress @ B -> x direction ice stress at sea ice basal, Y Ice Stress @ B -> y direction ice stress at sea ice basal, SW Flux @ B -> shortwave radiative flux at sea ice basal, LW Flux @ B -> longwave radiative flux at sea ice basal, Salt Flux @ B -> flux of salt at sea ice basal, Water Flux @ B -> flux of fresh water at sea ice basal, Ice U -> zonal velocity of sea ice at the surface, Ice V -> meridional velocity of sea ice at the surface, Ice Thick -> ice thickness, Ice Albedo -> sea ice albedo, Ice Temp -> temperature of sea ice at surface.

2.2. Data Assimilation

The Navy-ESPC system utilizes two mature assimilation code bases that have been developed for the GOFS 3.1 and NAVGEM forecast models separately. The Naval Research Laboratory (NRL) Atmospheric Variational Data Assimilation System Accelerated Representer (NAVDAS-AR) is used for the atmosphere

and the Navy Coupled Ocean Data Assimilation (NCODA) is used for the ocean, sea-ice and wave models and.

2.2.1. NAVDAS-AR: Atmosphere Data Assimilation

NAVDAS-AR (Rosmond and Xu 2006; Xu et al. 2005) has the ability for both weak and strong constraint variational assimilation, and is formulated in the terms of dual variables (i.e., observation space), the dimension of which is generally much smaller than the corresponding state (i.e., model) space. This restriction of the observation space is accomplished by discarding the unobservable degrees of freedom in the system (Bennett 2002).

NAVDAS-AR can process well over 100 million observations in every 6 hour data assimilation window. In a typical cycle, after quality control and data thinning, approximately 4 million observations are assimilated to create a final analysis. Observations routinely assimilated are listed in Table 3. These include: conventional observations (e.g. radiosondes., dropsondes, buoys, etc.); aircraft observations; feature tracked winds; radiances from AMSU-A, MHS, SSMIS, AIRS, ATMS, IASI, CrIS, SAPHIR, GMI, AMSR/2 and a suite of geostationary sensors which provide a clear-sky radiance product; bending angles from Global Navigation Satellite System Radio Occultation (GNSS-RO); SSMIS and WindSat wind speeds, scatterometer winds, and synthetic observations of tropical storms. NAVDAS-AR approximates the full non-linear Navy-ESPC coupled forecast system with a series of linear coupled Euler-Lagrange equations, which are in essence a simplified version of the stand-alone NAVGEM system and allow for a computationally efficient solve of the assimilation problem.

The resolution of the adjoint and tangent-linear models, and thus the effective resolution of the analysis increments, is approximately 100 km (T119). The same vertical model resolution is used for both the full Navy-ESPC system and the adjoint and tangent-linear models and the resulting analysis. The solution, which is the system's best estimate of the atmospheric state, is obtained through a series of Tangent Linear Model (TLM) and adjoint model integrations. The number of integrations depends on the number of observations and their fit to the full resolution Navy-ESPC coupled model. Typically about 70 iterations are required for the current operational system to meet the pre-specified convergence criterion. This is the most computationally intensive part of the 4D-Var analysis. The computation of the convolution of the background error covariance matrix with the adjoint sensitivity field at the initial time represents a substantial portion of each model integration computational cost.

2.2.2. NCODA: Ocean and Sea Ice Data Assimilation

The Navy Coupled Ocean Data Assimilation (NCODA) is a fully three-dimensional, multivariate (3DVar) data assimilation scheme (Cummings 2005; Cummings and Smedstad 2013) for the following ocean/ice variables: temperature, salinity, geopotential, vector velocity components, and ice concentration; all are analyzed simultaneously. Data are selected for assimilation based on receipt time (i.e., the time the observation is received at the center) instead of the observation time so, any data received since the previous NCODA analysis are used in the next analysis. For each data type, the user defines the maximum age of data to be used in the analysis. All data will not necessarily be from synoptic times, so they can be compared against a time dependent background field using the First Guess at Appropriate Time (FGAT). Hourly forecast fields are used in FGAT for assimilation of Sea Surface Temperatures (SSTs) to maintain the diurnal cycle, whereas daily averaged forecast fields are used in FGAT for profile data, both synthetic and real. NCODA is cycled with HYCOM and CICE to provide updated initial conditions for the next model forecast using an incremental analysis update procedure (Bloom et al. 1996).

In GOFS 3.1, the NCODA ocean analysis increments are inserted into HYCOM over a six hour window, whereas the NCODA ice analysis is directly inserted into CICE. The analysis corrections to the HYCOM and CICE forecasts are based on all observations that have become available since the last analysis, which may include observations made prior to the previous analysis data window. These include surface observations from satellites, including altimeter Sea Surface Height (SSH) anomalies, SST, and sea ice concentration, plus in situ SST observations from ships and buoys as well as temperature and salinity profile data from XBTs, CTDs and Argo floats (Table 3). See Table 13.1 in Cummings and Smedstad (2013) for a complete list of assimilated observations along with typical data counts. All observations are first quality controlled, and this is done via NCODA_QC (Quality Control), which is operational at FNMOC.

2.2.3. Loosely Coupled Data Assimilation

A loosely coupled data assimilation (DA) system is used for the Navy-ESPC model. Loosely coupled DA, which is also referred to as weakly coupled DA, is defined by using separate DA systems for the ocean and atmosphere; and using a coupled model for the first guess (i.e., forecast model). In the Navy-ESPC system, the Navy-ESPC coupled forecast model is used for the first guess while NCODA and NAVDAS-AR produce increments for the ocean-ice and atmosphere respectively. The atmosphere and ocean increments are inserted using Incremental Analysis Update (IAU) for 3 hours before the forecast target time (e.g., 00Z, 06Z, 12Z, 18Z). The 3 hour IAU for NCODA and NAVDAS-AR is new for Navy-ESPC, as the GOFS 3.1 stand-alone system uses a 6 hour IAU, and NAVDAS-AR does not use IAU with NAVGEM. NAVDAS-AR runs every 6 hours for the forecast times, and NCODA runs once a day at 12Z. Figure 3 shows a schematic of the Navy-ESPC loosely coupled data assimilation system.

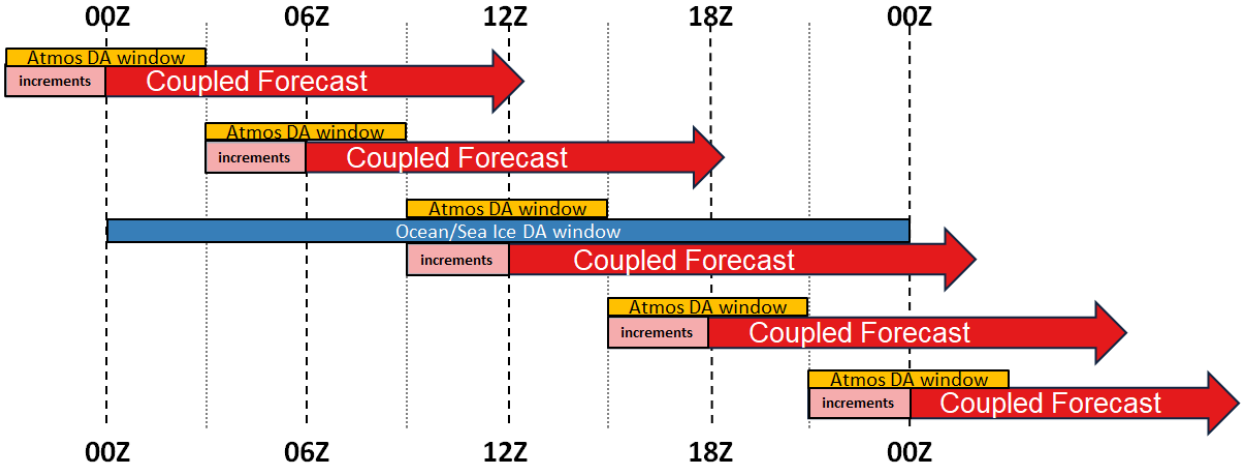

Figure 3: Schematic of the loosely coupled DA system for the Navy-ESPC system. The red arrows represented the Navy-ESPC first guess/forecast from the coupled model. The yellow bars represent the observational window used for NAVDAS-AR and the blue bar represents the nominal ocean and sea ice observational window used for NCODA. The light red boxes represent when the increments produced by NCODA and NAVDAS-AR are inserted in the Navy-ESPC coupled model. Note, there is a 3 hour Incremental Analysis Update (IAU) for insertion of the increments.

2.3. Ensemble Design: Perturbed Observations

For this report and proposed for IOC, the Navy-ESPC ensemble is configured as a 16-member ensemble, based on a perturbed-observation approach. Note, the Navy-ESPC ensemble perturbed observations approach is different from the method used in the operational NAVGEM ensemble, or the regional ocean ensemble tested at NAVOCEANO. In the Navy-ESPC perturbed-observation approach, each ensemble member maintains 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.

NAVDAS and NCODA have the capability to introduce perturbations to the observations assimilated in the atmosphere, ocean, and ice. 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 the NCODA, satellite observations of SSH, SST, and ice concentration and in situ surface observations are perturbed before thinning, synthetic profiles are generated using perturbed SSH and SST predictors, and in situ profile observations are perturbed individually independent of the profile thinning.

In the present configuration, 15 members add perturbations to the observations and one "control" member does not. Note, that this control member is still part of a random draw from the unknown true distribution of errors.

The forecast model resolutions used for the ensemble runs are described in Table 2. As shown, the ensemble resolution is lower than the deterministic run because of computational limitations – the lower computational cost at lower resolution allows for a larger ensemble with the resources available.

3. Computing Details

3.1. Input Data Streams

FNMOC collects, quality controls, and delivers to the operational systems the required atmosphere, ocean, and ice observations data. The IOC Navy-ESPC ensemble system does not require any additional data other than those assimilated in the current operational uncoupled systems. The total volume of the data for the full system is currently about 100Gb/day. Table 3 lists a summary of the input observations used by NCODA and NAVDAS-AR. Each ensemble member uses the same input data streams.

Table 3: List of observations used in NCODA and NAVDAS-AR. In-Situ and satellite observations used in each system are listed.

System	Observations	NCODA
NCODA	In-situ	a. Temperature, salinity, profile data from XBT, CTD, Argo, TAO moorings, gliders, and marine mammals b. Current observations from HF radar, drifting buoy c. Optical data from gliders, AUVs d. Naval Ice Center ice edge

	Satellite	a. Sea Surface Temperature (SST) b. Sea Surface Salinity (SSS) c. Altimeter Sea Surface Height Anomaly (SSHA) d. Altimeter Significant Wave Height (SWH)
		e. Sea surface color (optical) data f. Microwave and visible sea ice concentration g. Satellite-based heat flux estimates
NAVDAS-AR	In-situ	a. Radiosondes; Pibals; Downsondes; Dropsondes; and Driftsonde b. Land and ship surface observations c. Aircraft observations d. Synthetic observations
	Satellite	a. Surface Winds i. Scatterometer, ASCAT and ERS-2 ii. SSMI/SSMIS iii. WindSat b. Feature Tracked Winds i. Geostationary (6 satellites) ii. Polar Orbiters (AVHRR and MODIS) iii. Combined polar/geo winds (CIMSS) c. GNSS-RO Bending Angle i. GRACE-A, -B; Terra and TanDEM SAR-X ii. COSMIC FM1-6, COSMIC 2, KOMPSAT-5 iii. GRAS MetOp-series d. IR Sounding Radiances i. 3 IASI, AIRS, 2 CrIS ii. Geostationary Clear-Sky Radiances (6 satellites) e. MW Sounding Radiances i. AMSU-A; 3 SSMIS; 4 MHS; 2 ATMS f. Ozone retreivals i. SBUV/2; GOME-2; OMPS-Nadir Profiler and Total Column g. Aerosol Optical Depth i. 2 MODIS (MOD04) ii. 4 AVHRR (ACSPO) iii. VIIRS h. Fire Biomass i. 2 GOES, Meteosat i. Soil Moisture i. AMSR-2, SMOS j. Ocean Altimeter i. Sea Surface Height Anomaly (SSHA) ii. Significant Wave Height (SWH)

3.2. Output Data Streams

Output data stream types are: restart files, model history output, and post-processing output for each modeling component. In operations, there is no output from the mediator/coupler in the Navy-ESPC system.

For NAVGEM, the model history output is the same as the model restart files. During the forecast and after the forecast completes, these model history files are used to create flatfiles (i.e., FORTRAN write binary files) for post-processing tools. Output is written every 3 hours for the short forecast lengths and less frequently (either 6 or 12 hours) after forecast lengths of 5 days. For HYCOM, the restart files are separate from the model output files. While the model runs and after completion, the HYCOM binary native grid model output are further interpolated in space to a uniform 0.08° latitude/longitude grid between ±80° latitude and in the vertical to 40 z-levels for the SSH, temperature, salinity, zonal and meridional velocity components; these are output in netCDF format (the z-levels and file format are defined by the Naval Oceanographic Office/FNMOC-SSC). CICE also has independent restart files, in addition to model history files. CICE model history files include multiple variables for the entire globe, and these files are re-written for selected variables for the Arctic and Antarctic regions. To better describe the volumes of data output by the system the history files which describe the system state are presented in Table 4. Each component has namelist options in which the amount of output can be controlled. Table 4 represents the amount of data produced for the VTR testing. Note, data processing, data transfer, and data removing are being performed while the model is running to minimize the total amount of data stored at one time.

Table 4: Output file size estimate for each component of the Navy-ESPC System and the total. Estimates represent the configuration used for the VTR testing and the amount of output can be changed through varies namelist options. Total size in the last row is the output plus postoutput times 45 days times 16 members plus two restart files.

	HYCOM	CICE	NAVGEM	Total
Restart file(s) Size	24 GB	7.6 GB	4.4 GB	36 GB
Estimated output for a 24 hour forecast	168 GB	5.4 GB	9.6 GB	183 GB
Estimated postoutput for a 24 hour forecast	19 GB	1.5 GB	5.2 GB	25.7 GB
Total Size for a 45 day 16 member ensemble forecast	134.68 TB	4.9 TB	10.7 TB	150.28 TB

3.3. Computational Requirements

Running a coupled model requires a synchronization between each component (e.g., NAVGEM, HYCOM, CICE, MEDIATOR) to complete computations near the same time for one coupling time-step. The number of processors for each modeling component can be increased or decreased as needed for the times-to-completion to be similar. The processor count for NAVGEM and HYCOM can be changed without a re-compile. However, CICE's processor count depends on resolution and is set during the compile. For CICE, the processor count needs to be an integer divisor of the number of points in the longitudinal direction. CICE also allows for different processor shapes, which determines how the grid arranges with the processors, and the Navy-ESPC system only uses the slenderX1 shape in CICE. For the GLBb0.08, which is the resolution of HYCOM's computational grid and the ensemble resolution configuration, the number of points in the longitudinal direction in CICE is 4500. For testing, we selected

CICE processors counts of 150, 180, and 225; and then changed the processor counts for NAVGEM and HYCOM so the computations take similar times to completion and to decrease total walltime for the forecast. Table 5 shows the processors (e.g., core) counts used for these testings for all components, and shows the time to completion for these core counts on the Navy SGI machines. Table 5 only shows times when using 225 processors for CICE. Note that the mediator/coupler shares cores with each component. The time to completion is faster with lower total core counts because each component requires a similar amount of time for completion. This results in components not needing to wait, which requires additional time to re-start the simulation. VTR testing used 2227 cores for the ensemble resolution testing on the Cray XC40s, but the below results are for the SGI machines.

Table 5: Timings of completing a 45 day Navy-ESPC forecast at the ensemble resolution on multiple core counts on the Navy SGI machines. The total core count is the sum of the core counts in HYCOM, CICE, and NAVGEM.

Total Core Count	HYCOM	CICE	NAVGEM	Time to Completion (in hours)	Model Day per Hour
1717	1332	225	160	7.6	5.9
2016	1631	225	160	8.2	5.5
2267	1882	225	160	8.4	5.3
1285	900	225	160	8.8	5.1
2227	1882	225	120	10	4.1

4. Methodology

4.1. VTR Test Period

The testing for the ensemble period was separated largely into two categories: (1) running of the analysis data assimilation system, and (2) execution of 60 day long forecasts.

The analysis ensemble runs started on 15 December 2016 (Table 6). 16 members started on this date using the same restart files and ran until 31 January 2018. 15 members had perturbed observations in the DA cycle, and one control member did not. Since the members started with the same initial conditions, the spread increase by time is examined in section 5.

The 60-day forecasts were started after a 45-day "spin-up" period in order to give the parallel update cycles time to diverge. All long forecasts were executed on Wednesdays at 12Z and the first ensemble of forecasts occurred on 01 February 2017 (Table 6). The last ensemble of forecasts completed on 24 January 2018. Only one long forecast did not complete due to an instability in HYCOM. Some forecasts failed with the normal HYCOM time step and had to be re-run using a reduced HYCOM time step.

Table 6: Simulation period for the analyses and long forecasts. All long forecasts start on Wednesdays at 12Z. The ensemble consist of 16 members of the analysis and long forecasts.

Simulation	First Forecast Date	Last Forecast Date
Analyses	15 December 2016	31 January 2018
Long Forecast	01 February 2017 (First Wednesday)	24 January 2018 (Last Wednesday)

4.2. Verification Methods

The Navy-ESPC model is the first Navy coupled model transitioned to operations. Specifically, this is the first time FNMOC will have a global ocean/ice ensemble, ocean/ice predictions past 7 days, and atmospheric predictions past 16 days. The Navy-ESPC model differs from other coupled models in operations in that (1) an eddy resolving ocean is used for the ensemble and (2) forecasts for each component matter to the customer. e.g., ECMWF's coupled operational system mainly focuses on atmospheric forecast while the Navy-ESPC model focuses on atmosphere, ocean, and sea ice forecasts.

Given the lack of a current operational baseline system for comparison, our main verification metric is to compare the Navy ESPC forecasts against a forecast of climatology. Following the methodology of McLay et al. (2016a), the comparison against climatology is considered a measure of the limits of predictability. The limit of predictability is dependent on the choice of metric, region, and/or variable. Note that no bias correction is used for the results presented in this VTR.

4.2.1. Atmosphere

Madden-Julian Oscillation

As noted above, a goal of the NAVGEM physic updates were to obtain skillful long forecast of the Madden-Julian Oscillation (MJO). The MJO is a mode of atmospheric variability in the tropics with a timescale of 20-100 days and is one of the few phenomena affecting global predictability on extended-range timescales (Lim et al. 2018). The phase of the MJO is known to have substantial impact on many other phenomena that impact sensible weather important to both DoD and civilian interests, such as: tropical cyclone genesis and frequency, atmospheric rivers, and even the atmospheric circulation in the Arctic and Antarctic. As such, it is important for extended-range forecasting systems to be capable of realistic simulations and accurate predictions of the MJO.

We compare the Navy-ESPC system's prediction of the MJO during the VTR time period to other state-of-the art systems. In order to evaluate the skill of the MJO forecasts, we use the Real-time Multivariate MJO (RMM) index (Wheeler and Hendon 2004). This commonly used index has two components, which we evaluate. RMM1 is positive when the MJO is active over the Maritime Continent and negative when the MJO is active over South America to Africa. RMM2 is positive when the MJO is active over the Pacific Ocean, and negative when the MJO is active over the Indian Ocean. As with the other atmospheric indices, we compare forecasts from the NAVGEM ensemble transform atmosphere-only ensemble (NAVGEM ET), Australian Bureau of Meteorology (BOM), Meteo-France (CNRM), the European Centre for Medium-range Weather Forecasts (ECMWF), and the NOAA Climate Forecast System version 2 (CFSv2). The forecasts from the other centers were downloaded from the Subseasonal to Seasonal (S2S) database (Vitart et al. 2017). In addition we show results for the NAVGEM deterministic forecasts as well. Models from the other centers vary in ensemble size as noted in Table 7. We also include scores for a persistence forecast starting from the Navy ESPC ensemble mean initial state, and a "climatological" forecast derived from an ensemble constructed using analyzed values of the indices from the same month taken from the prior 18 years. Verification indices (and the indices used in the climatological ensemble forecasts) were derived from NOAA satellite-derived Outgoing Longwave Radiation, and ECWMF ERA-Interim reanalyses.

Table 7: Description of different modeling systems used in the MJO and teleconnection intercomparisons.

Model Acronym	Modeling Center	# of Ensemble Members

BOM	Australian Bureau of Meteorology	33
CFSv2	NOAA Climate Forecast System Version 2	16
CNRM	Meteo-France	51
EC16	ECMWF	16
EC51	ECMWF	51
NAVGEM-D	Navy NAVGEM Operational Deterministic System	1
NAVGEM-E	Navy NAVGEM Operational Ensemble System	20
ESPC-D	Navy ESPC Deterministic Forecasts	1
ESPC-E	Navy ESPC Ensemble Forecasts	16

Atmospheric Teleconnection Indices

In addition to the MJO, it is anticipated that teleconnection patterns may be predictable on longer time scales compared to fields at one location. Because of their large scale, teleconnection patterns evolve on relatively slow time scales and as a result are more predictable than the atmospheric state at any single point.

The teleconnection patterns considered include the Arctic Oscillation (AO) and Antarctic Oscillation (AAO), which provide information on the zonal symmetry of the northern and southern jet streams, and the Pacific North American Oscillation (PNA) and North Atlantic Oscillation (NAO), which provide information on dominant modes of variability over these respective regions.

We evaluate the skill of the Navy-ESPC ensemble in both a deterministic (individual forecast) and ensemble mean sense, and benchmark the performance of the Navy system with other state of the art prediction systems from NOAA and operational centers. The other centers considered here are the same as described in Table 7. The verifying scores were computed using the National Centers for Environmental Prediction (NCEP) reanalysis. The climatological forecasts were produced using an ensemble derived from NCEP reanalysis states from the verifying month selected over the previous 68-year time period. Persistence forecasts were produced from the Navy ESPC ensemble mean analysis.

Evaluating the skill of the forecasts for these teleconnection patterns is a useful way to gauge the ability of the model to capture these relatively slowly varying leading modes of atmospheric variability, which are tied to anomalous sensible weather conditions. In addition, the Naval Information Warfighting Development Center (NIWDC), responsible for developing tactics, techniques, and procedures for situations including meteorology and oceanography impacts, has indicated that Navy forecasters are interested in forecasts of the AO to provide information as to the magnitude of anomalies in weather patterns that might be expected in the Northern Hemisphere mid and high latitudes.

Ensemble Synoptic Score Card

The MJO and teleconnection pattern analysis diagnostics are designed to understand the S2S predictability of the Navy-ESPC model. In addition to these new metrics, we examine synoptic forecast metrics that have been developed for the NAVGEM-ET scorecard (McLay and Whitcomb 2017; McLay et al. 2016a; McLay et al. 2016b).

The atmospheric synoptic ensemble scorecard provides a scalar summary measure of the difference in performance between a new candidate ensemble system and some baseline system, e.g. the current

operational ensemble system or a climatological ensemble. Because the Navy-ESPC model is new, the main comparison is against the climatological ensemble.

The scorecard methodology used for the Navy-ESPC transition test follows those used in previous years' ensemble transition tests, e.g. McLay and Whitcomb (2017); McLay et al. (2016a). Here climatological forecasts are derived from a 20-year ECMWF ERA-Interim reanalysis dataset. This dataset is in the form of 1°x1° flat files for 0000UTC and 1200UTC of each day in the years 1991-2010. The climatological 20-member forecast ensemble for a given date, say 2014013112, is obtained by collecting together the 20 different ERA-Interim analyses for the dates 1991013112, 1992013112, …, 2009013112, 2010013112.The scorecard encompasses the following variables, regions, lead times, and metrics defined in Table 8. Many metrics are used for a holistic diagnostics of the ensemble. Hence, one skill score metric does not dominant the ensemble performance. ECMWF analysis is used at the validating truth in all metrics using the synoptic score card.

Table 8: List of variables, regions, lead times and metrics used in the atmospheric synoptic ensemble scorecard.

Variables	500 hPa geopotential height ($Ø_{500}$) 10m wind speed (V_{10m}) 250 hPa wind speed (V_{250}) 2m air temperature (T_{2m}) 850 hPa air temperature (T_{850})
Regions	Tropics (TR) [20°S,20°N] Southern Hemisphere Extratropics (SE) [90°S,20°S] Northern Hemisphere Extratropics (NE) [20°N,90°N]
Lead times	24h-1452h, at intervals of 24h
Metrics	BIAS: Average error of the ensemble mean RMSE: Root-mean-square error of the ensemble mean CRPS: Continuous ranked probability score of the raw ensemble cumulative distribution function VARR: Average ratio of ensemble variance and squared-error of the ensemble mean ACOR: Uncentered Anomaly correlation of the ensemble-mean

Scorecard entries are obtained through the following steps:

1) Choose a particular combination of variable, region, lead time, and metric.
2) For both the new system and the baseline system, calculate the metric for each date-time group (DTG) in the test period. There is a total of 47 DTGs in the Navy-ESPC test period.
3) Calculate the paired difference between the metric value of the new system and the metric value of the baseline system for each DTG in the test period.
4) Calculate the average of the paired differences over all DTGs in the test period.
5) Test the statistical significance of the average paired difference. The statistical significance is tested using a parametric t-test with an effective sample size that accounts for temporal correlation of the paired difference values (Wilks 2011).
6) Calculate the relative improvement of the average metric value of the new system as compared to the average metric value of the baseline system.

7) Given the statistical significance and the relative improvement, calculate the score using Table 9.

Table 9: Atmospheric scorecard values dependent on statistical significance between cases. For the Navy-ESPC system, the Navy-ESPC ensemble is a case and climatology is another case.

	Score
Case #1 Statistical significance exceeds 95% threshold AND relative improvement ≥ +5%	+1
Case #2 Statistical significance exceeds 95% threshold AND relative improvement ≤ -5%	-1
All other cases	0

To summarize the vast amount of information in the synoptic scorecard, a grand cumulative score is obtained by summing the scorecard entries over all possible combinations of variable, region, lead time, and metric. All the scorecard entries are equally weighted in the summation. These cumulative scores include a grand cumulative score, obtained by summing the scorecard entries for each lead time over all possible combinations of variable, region, and metric. All the scorecard entries are equally weighted in the summation. Various other cumulative scores are obtained by summing all the scorecard entries for each lead time as a function of metric, region, or variable. For further insight, various other cumulative scores are obtained for each season by summing all the scorecard entries for a given choice of metric, region, or variable.

4.2.2. Ocean

The extended-range ensemble-mean ocean forecasts of temperature and salinity are evaluated using in situ profile and surface measurements. During the model forecast runs, daily-mean temperature and salinity fields were calculated in HYCOM and saved in native format, and postprocessed into the FNMOC-standard netCDF format described above. For each netCDF file, a corresponding 24 h 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, both on the observed depths and on the standard 40 levels. Forecast BIAS and RMSE are calculated using the matchups as a function of depth and of forecast lead time. The ocean model prediction of selected isotherm depths is evaluated using the observation matchup files. The BIAS and RMSE of the forecast depth for the 26C, 20C, and 15C isotherms are calculated as a function of forecast lead time. The ocean ensemble forecast spread skill for temperature and salinity are evaluated using the same matchup dataset. Matchups are binned by the error, and bin means and their slope are calculated as a function of depth and forecast lead time.

The primary reference comparison for the ocean forecast is climatology, because there is presently no extended-range (more than 7 day) global ocean forecast system. The climatology used for ocean temperature and salinity is the Generalized Digital Environmental Model (GDEM) v4(Carnes et al. 2010).

4.2.3. Sea Ice

Extended-range forecasts of sea ice extent are evaluated using two metrics. The first is the Integrated Ice Edge Error (IIEE) (Goessling et al. 2016), which is the total area over which the model has over-forecast the extent of the ice edge plus the total area over which the model has under-forecast the

extent of the ice edge. Here, we define the ice edge as 15% ice concentration. We also consider a Brier Skill Score (BSS) (Brier 1950) for ice concentration below 15%. The Brier Skill Score is the degree of improvement of the Brier Score (BS) of the Navy-ESPC ensemble forecast over the BS produced using a persisted analysis or climatology. The BSS of 1 is a perfect forecast, while a BSS of 0 indicates there is no improvement as compared to persistence or climatology. The BSS was also used in Wayand et al. (2019) as a sea ice prediction metric.

The Navy-ESPC model assimilates sea ice concentrations from a blended Multisensor Analyzed Sea Ice Extent (MASIE) (Meier et al. 2015) and passive satellites. This blended product can differ from passive remotely sensed sea ice concentrations; particularly during the summer months when the passive sea ice concentrations may classify melt ponds as open water. The blended product may also differ from the initial sea ice concentrations used in the model because the CICE model relaxes changes in the concentration before the simulation to minimize energy imbalances introduced during the assimilation process. For these reasons, we use self-analysis for verification, which follows Hebert et al. (2015).

The climatological forecasts are produced using SSMR SSM/I data from the National Snow and Ice Data Center (NSIDC) (NSIDC 0051) from the 2007-2017 time period. Including analyses from years prior to 2007 may bias the climatology high given the recent observed trends in sea ice. We calculate these scores for both the Arctic and Antarctic, and consider the scores for the entire 2017 period, and for each individual season.

5. Results

5.1. Ensemble Spin Up

This section attempts to answer the question: How long does it take to saturate the ensemble spread at analysis time? The goal is to determine how long of a "spin-up" period is required before ensemble spread at analysis time saturates.

5.1.1. Atmosphere

The ensemble spread at analysis time of many atmospheric variables did not increase after a few days after the start of the experiment (Figure 4). After this short spin-up period, the tau=0 ensemble spread showed little variation and, in general, did not track excursions in the value of the bias corrected ensemble-mean error using ECMWF reanalysis as verification. The atmospheric component of the Navy-ESPC system was significantly under –dispersive, that is, the ensemble spread was smaller than the ensemble mean error (between a factor of 1.6 for tropical winds and 6.5 for 850 hPa Temperatures, Table 10). This under-dispersion was not unexpected as it has been observed in other perturbed observation systems (e.g. UKMO and ECMWF). Plans to address this under-dispersion are discussed in section 7.

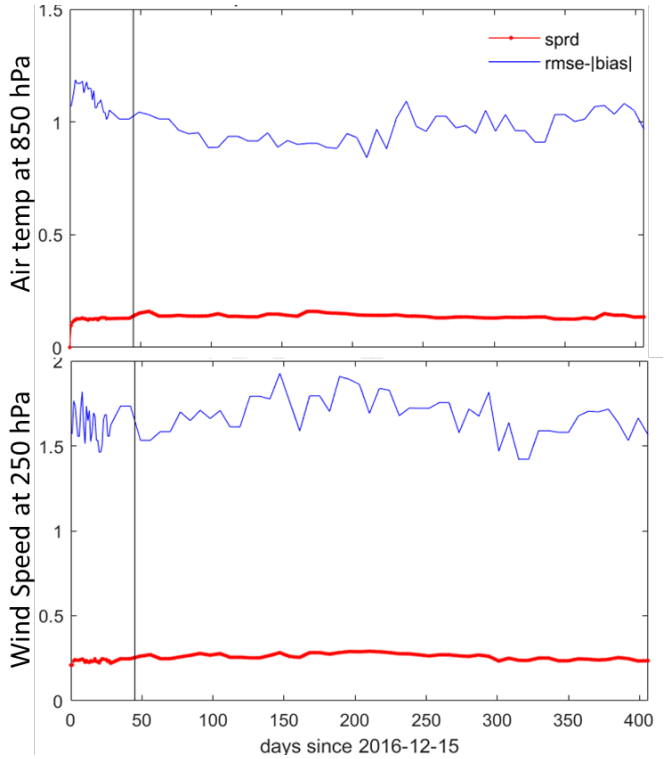

Figure 4: Analysis (tau=0) ensemble (red) spread and (blue) RMSE – |bias| for (top) air temperature at 850 hPa and wind speed at 250 hPa. The vertical black line represents when the long forecasts started. BIAS is computed against the ECMWF analysis. 16 ensemble members are used in the analysis.

Table 10: Analysis (tau=0) ensemble "spread factor". The spread factor is defined as the ratio of the bias-corrected ensemble-mean RMSE to the ensemble standard deviation.

Variable	NE	SE	TR
Geopotential Height at 500 hPa	3.1	3.1	2.0
Winds at 10 meters	2.9	2.8	1.6
Winds at 250 hPa	4.9	4.4	3.1
Temperature at 850 hPa	6.4	6.5	5.3
Temperature at 2 meters	4.7	5.5	2.3

Though the Navy-ESPC is under-dispersive at analysis time, the 14 day forecast is performing in a similar manner as the NAVGEM-ET for some cases. In particular, the 10 meter winds in the Navy-ESPC ensemble and NAVGEM-ET ensemble are similar (Figure 5). In addition the analysis bias for 10 meter winds in ESPC (ensemble and high resolution deterministic) is smaller than for NAVGEM-ET over the oceans at analysis time (Figure 6). In some regions this improvement is on the order of 1.5 to 2 m s^{-1} in regions of Naval importance. These distinct spatial characteristics are not captured in the overall spatial averaging.

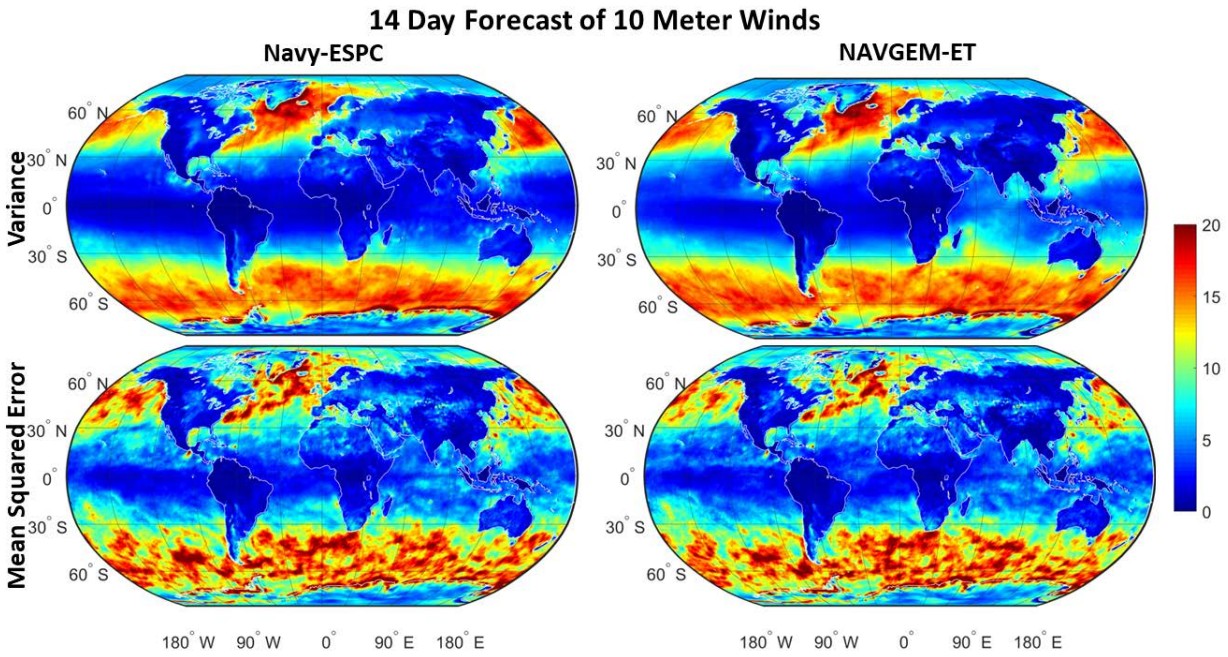

Figure 5: 14 day forecast (top) variance and (bottom) mean squared error of the (left) Navy-ESPC ensemble and the (right) NAVGEM-ET ensemble of 10 meter winds. The metrics are computed with respect to ECMWF analysis.

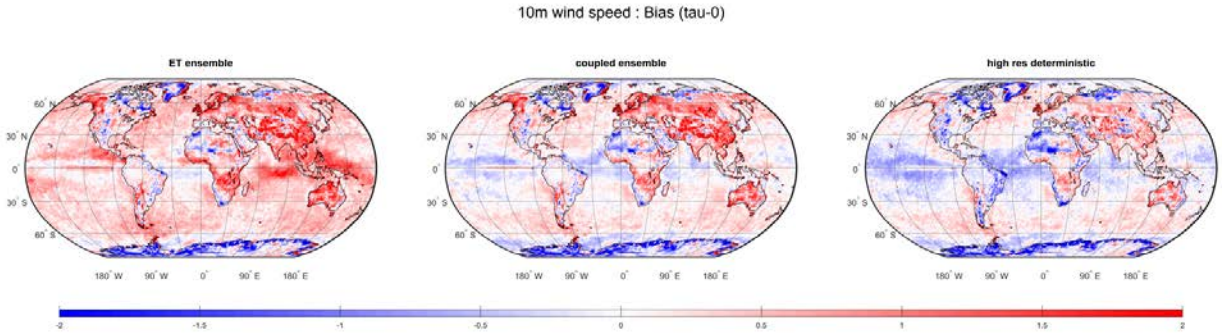

Figure 6: 10 meter wind bias at analysis time of the (left) NAVGEM-ET ensemble, (middle) Navy-ESPC ensemble, and (right) Navy-ESPC high resolution deterministic run relative to ECMWF.

5.1.2. Ocean and Sea Ice

Unlike the atmosphere that spun-up in a few days, it took more than a month for the ocean and ice spread to reach equilibrium (Figure 7). The longest spin-up was for the ocean sea surface height (over 100 days) and the shortest was SST (on order of a week). Ice variables exhibited a strong seasonal cycle in the spread metrics but were spun-up after about a month. As a compromise between different variables, we chose to start long forecast integration after 45 days of initial spread spin-up.

To measure the efficiency of this strategy, we evaluated upper ocean (surface to 200 m, generally including the mixed layer and upper thermocline) spread for the entire archive of the 52 long forecasts (Figure 8 and Figure 9). The ensemble spread measured at observed profiles during the 52 weekly

forecasts shows seasonal variation (and some artifacts of the data availability) but no noticeable increase in spread due to ensemble spin-up (Figure 8 and Figure 9).

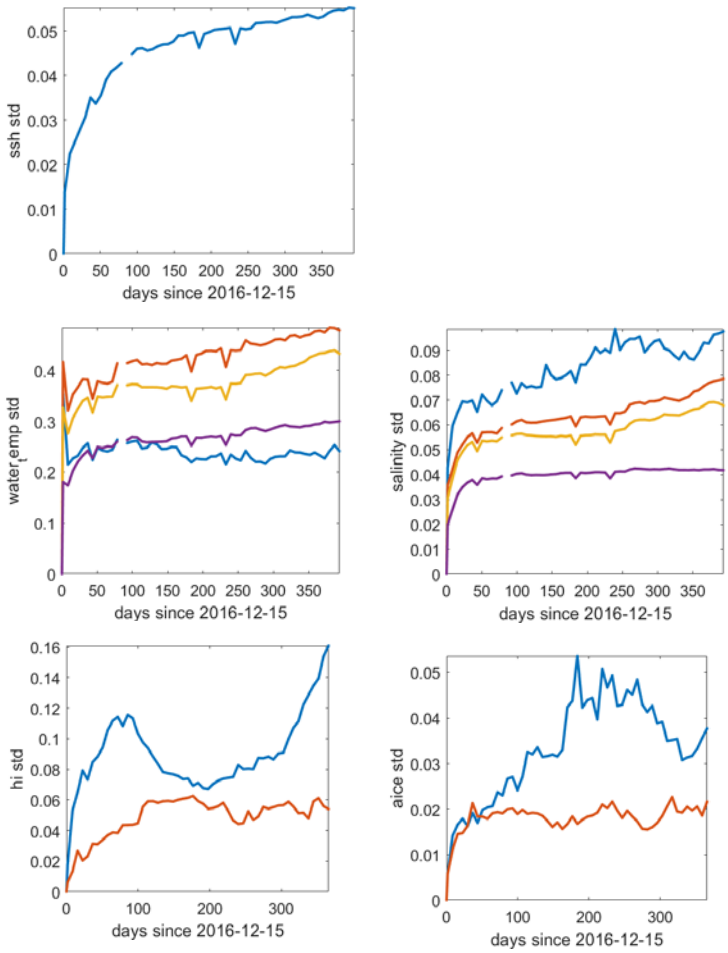

Figure 7: Initial-time spread of the 16 member ensemble. The cycling system was initialized with identical initial conditions on Dec 15th, 2016. The Navy-ESPC 60 day forecasts started around day 45 from 2016-12-15. The colors of the lines represent depth: blue -> surface, red -> 100 meters, yellow -> 200 meters, purple -> 500 meters.

As noted above, the long forecast runs were initiated about 45 day after the ensemble analysis cycle to allow for the ocean and ice spread to grow through the perturbed-observation approach and the nonlinear growth of perturbations in the cycling system.

For the upper ocean (surface to 200 m, generally including the mixed layer and upper thermocline), the ensemble spread measured at observed profiles during the 52 weekly forecasts shows seasonal variation (and some artifacts of the data availability) but no noticeable increase in spread due to ensemble spin-up (Figure 8 and Figure 9).

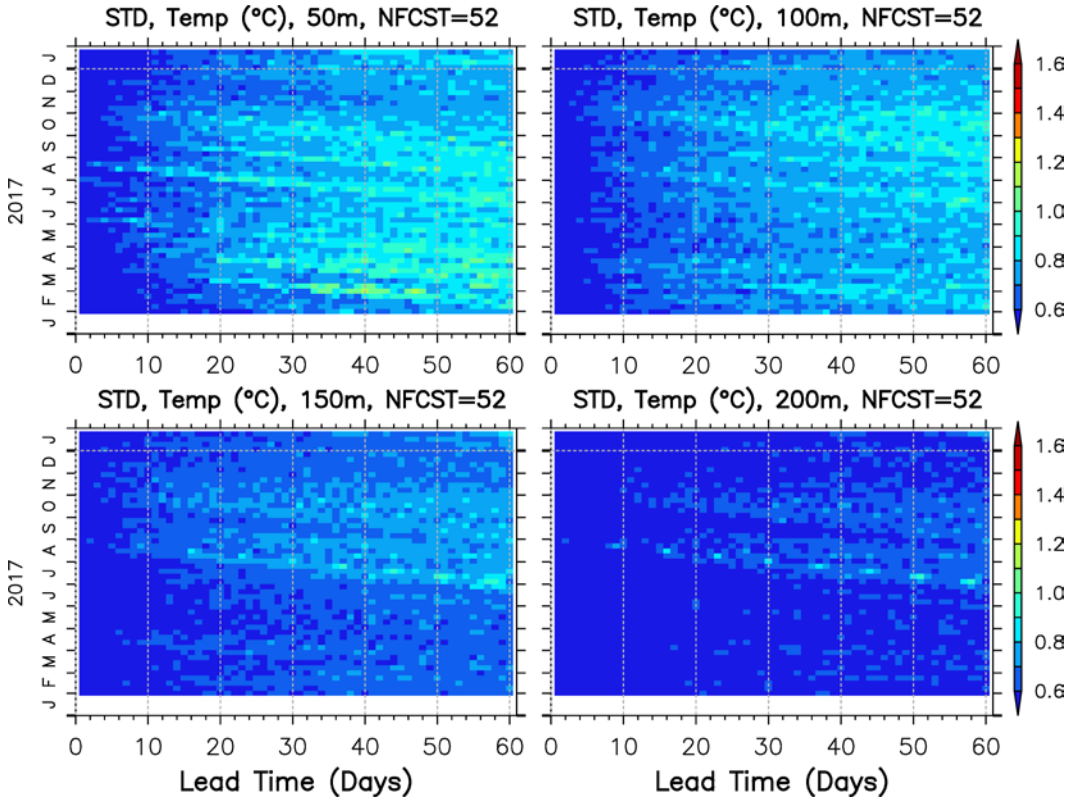

Figure 8: The Navy-ESPC ocean ensemble temperature standard deviation for the validation forecast dataset as a function of forecast lead time and forecast week.

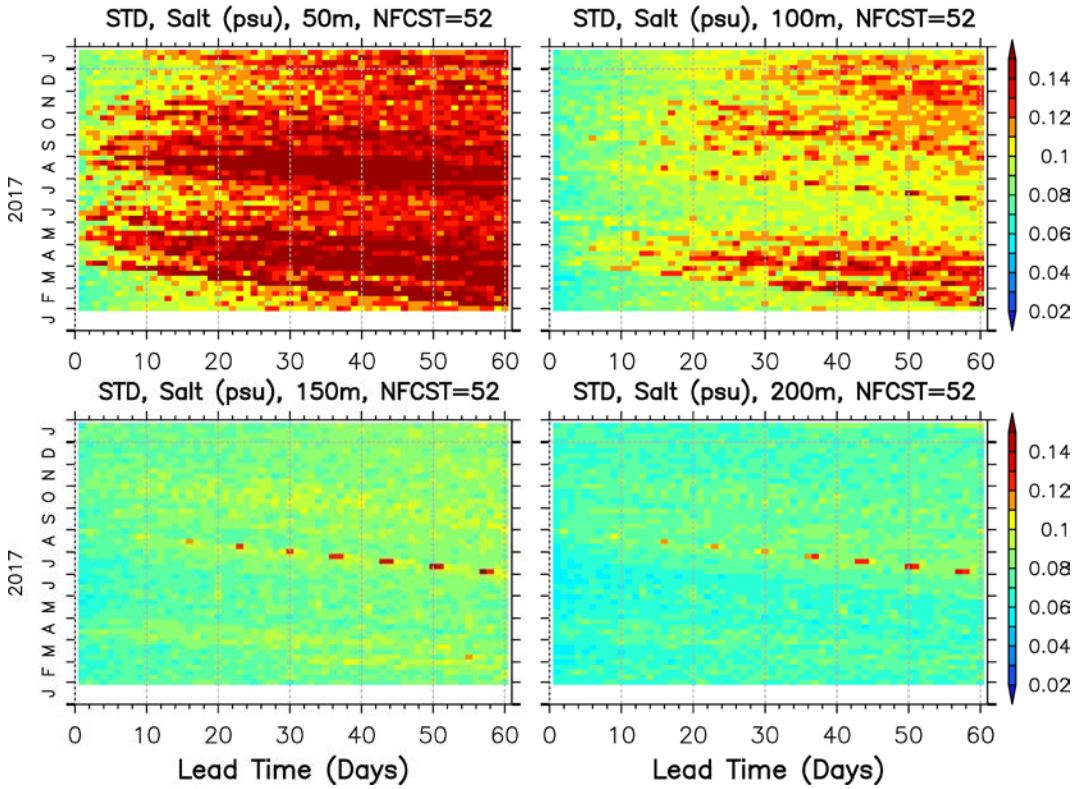

Figure 9: The Navy-ESPC ocean ensemble salinity standard deviation for the validation forecast dataset as a function of forecast lead time and forecast week.

5.2. Long Forecasts

5.2.1. Atmosphere

Madden-Julian Oscillation: Performance

Figure 10 shows the ACOR and RMSE for RMM1 and RMM2 evaluating the skill of the ensemble member forecasts individually, and the skill of the ensemble mean. For RMM1, the Navy ESPC ensemble individual members have the top performance for ACOR. In terms of RMSE, the individual Navy ESPC ensemble members are comparable to the best model, ECMWF, out to about 15 days, although as will be discussed below, some of the superior performance in terms of the average skill of the individual members is attributed to the ensemble underdispersion. When considering the ensemble mean, because of under-dispersion issues, the Navy ESPC ensemble does not gain as much skill as the other systems. It is still a very competitive model, substantially better than CFSv2, though not quite as good as ECMWF for ACOR or ECMWF and the French model for RMSE.

For RMM2, the ACOR performance is again relatively very good, being competitive with the top model (ECMWF) in terms of individual forecast performance, and competitive with all the other models in terms of ensemble mean performance, (ECMWF pulls ahead here). In terms of RMSE, the Navy-ESPC errors grow more rapidly than most of the other centers initially, although performance is competitive with CFSv2 at longer lead times. This is due to a strong negative bias in RMM2 that develops in the Navy-ESPC over the first week.

As the Navy-ESPC physics differ from the NAVGEM physics in part to enhance performance for the MJO, it is useful to compare Navy-ESPC and NAVGEM in terms of MJO forecast skill. When considering the skill of the individual ensemble forecasts, the Navy-ESPC ensemble (black curve) outperforms the NAVGEM ensemble (red curve) by a wide margin for RMM1, and by a smaller but still substantial margin for RMM2 for ACOR (the two are comparable for RMM2 in terms of RMSE). When considering ensemble mean scores, the Navy-ESPC ensemble still outperforms the NAVGEM ensemble for RMM1, but the scores are comparable for RMM2 AC, and NAVGEM outperforms the Navy-ESPC ensemble for long longer lead times for RMM2 RMSE. As noted in the previous paragraph, this is due to a strong negative bias in Navy ESPC for RMM2 that develops initially and eventually corrects.

It should be noted that the sample size here is quite small given that the MJO is not always active and the time period considered is only a year. Thus these results should be used to give a general idea of performance only. When examining 17 years of Navy ESCP forecasts produced for the SubX experiment, Janiga et al. (2018) found that the Navy-ESPC system overestimates the amplitude of the MJO whereas most other models underestimate the amplitude of the MJO. The ACOR of the RMM1 and RMM2 is more indicative of predictions of the phase of the MJO where the RMSE of RMM1 and RMM2 also reflects errors in MJO amplitude.

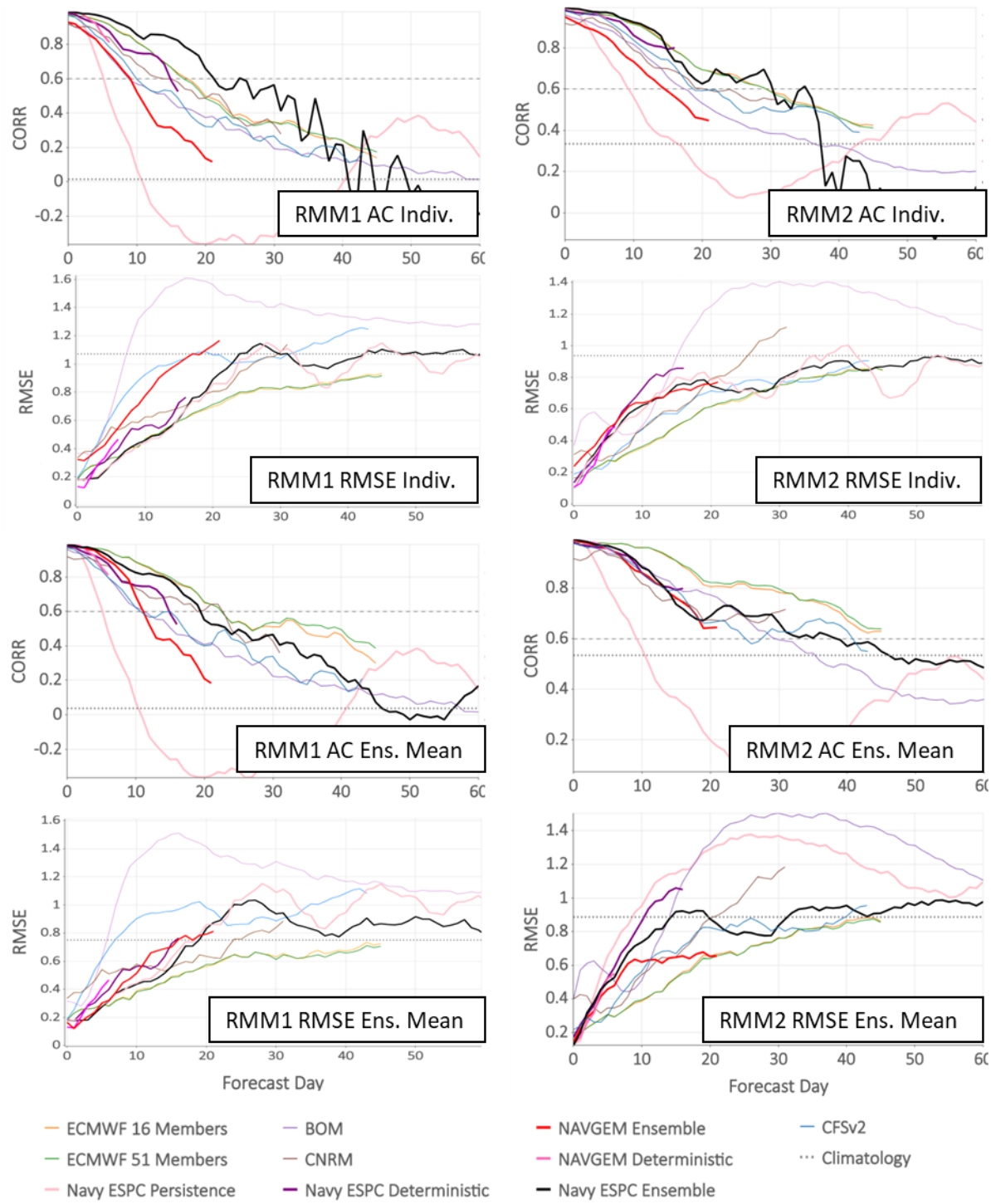

Figure 10: ACOR and RMSE scores for RMM1 (left) and RMM2 (right) as denoted in figure panels. Scores are calculated for individual forecast members and averaged for each center ("Indiv"), and calculated for ensemble means ("Ens. Mean"). Center colors are denoted in Key and include Australian Bureau of Meteorology (BOM), Meteo-France (CNRM), ECMWF 16-member and 51-member ensembles, NOAA (CFSv2), Navy-ESPC high-resolution deterministic, Navy-ESPC Ensemble, NAVGME deterministic, and NAVGEM Ensemble.

Figure 11 shows violin diagrams indicating the distributions of the day when the individual ensemble member ACOR falls below 0.6 for RMM1 and RMM2. As noted in the discussion of Figure 10, Navy-ESPC ensemble individual members perform exceptionally well in terms of RMM1 ACOR compared to the other centers. In terms of ensemble mean ("X"s), the Navy-ESPC performance is slightly below that of ECMWF, comparable to CNRM, and better than the other centers and the NAVGEM ensemble. The Navy-ESPC deterministic model also performs relatively well. For RMM2, the Navy-ESPC individual ensemble performance is not quite as good as ECMWF, but on average better than the other centers. For the ensemble mean ("X"s), the Navy-ESPC outperforms CFSv2 and BOM. For several of the forecast systems, the average RMM2 ACOR does not fall below 0.6 before the end of their forecast range, hence the lack of "X"s for these systems in the bottom panel of Figure 10. The large spread in the performance of the individual ensemble members, indicated by the large range of days covered by the violins, attests to sampling issues given the relatively small forecast data set. Presumably a larger forecast set would result in more similar skill amongst ensemble members.

As mentioned above, we wish to highlight a caveat when interpreting the mean scores of the individual ensemble members. We illustrate this point with Figure 12, which shows the RMSE for RMM1 and RMM2 for the least skillful (top) and most skillful (bottom) ensemble members. The skill range for the Navy ESPC (black curve) is considerably smaller than the skill range for the NAVGEM ensemble (red curve) and most other systems. That is, ensemble systems that are not as severely underdispersive as Navy ESPC will have individual members that are considerably more skillful and considerably less skillful than the Navy ESPC members. As the distribution about the average is not symmetric (that is, the RMSE score is bounded by zero on the skillful side), the ensembles that have larger spread suffer a penalty compared to Navy ESPC when computing the average skill score of the individual members. This should be kept in mind when interpreting the mean skill of the individual ensemble members.

The MJO forecasts have ACOR skill both above 0.6 and above the skill of a climatologically-derived forecast out to much longer lead times. In terms of ACOR, the forecasts can remain relatively skillful for a month. Future work will explore how to leverage this extended-range forecast skill in the MJO for sensible weather that is correlated with MJO phase.

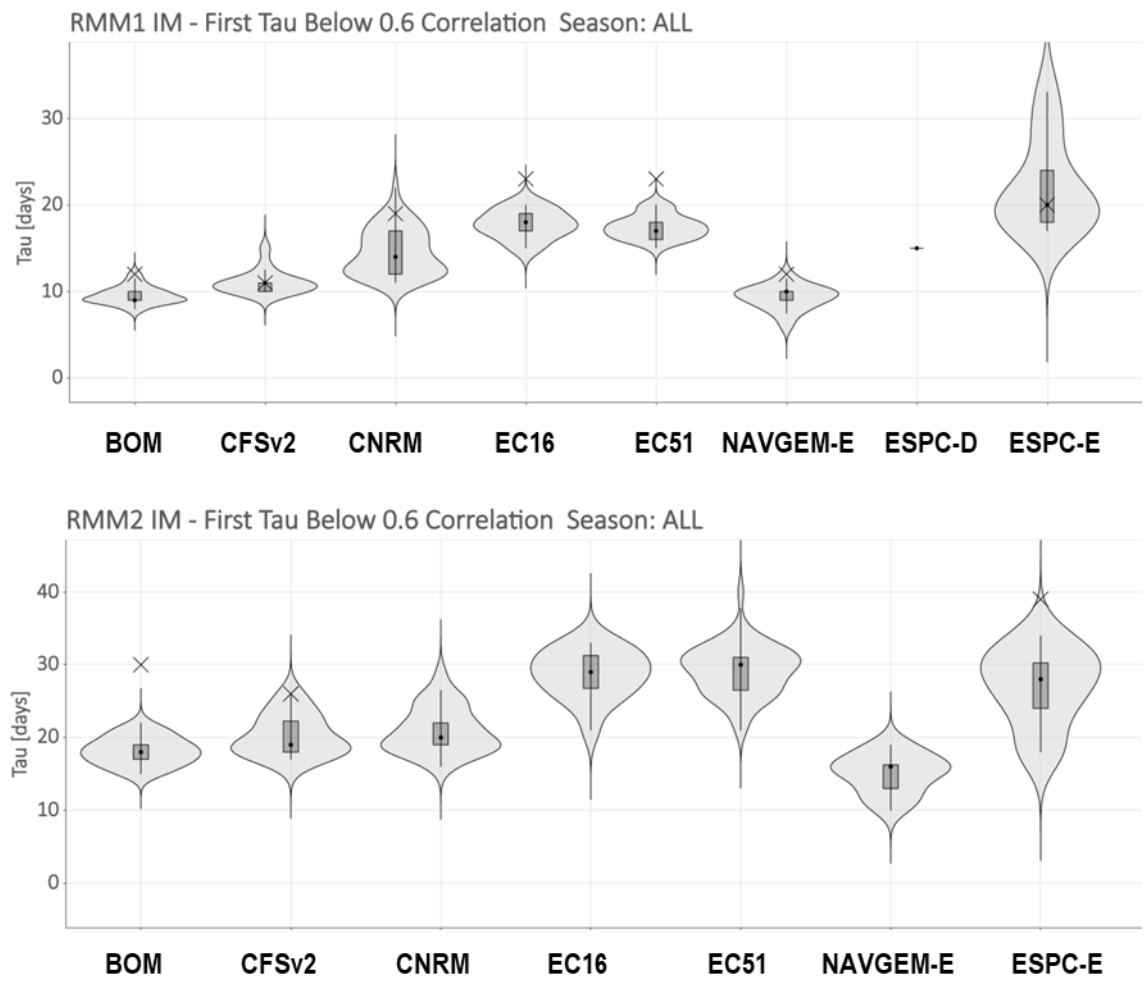

Figure 11: Violin plots showing the distribution of the day that the forecast of RMM1 (top panel) and RMM2 (bottom panel) ACOR falls below 0.6. Each ensemble member is treated as an individual (deterministic) forecast, and the width of the violin is proportional to the number of ensemble members that fall below 0.6 on that day. The forecasts considered come from the Australian Bureau of Meteorology (BOM), NOAA (CFSv2), Meteo-France (CNRM), ECMWF 16-member (EC16) and 51-member (EC51) ensembles, NAVGEM ET ensemble (NAVGEM-E), Navy-ESPC high-resolution deterministic forecasts (ESPC-D), and the Navy-ESPC ensemble (ESPC-E). The day on which the ensemble mean forecast ACOR falls below 0.6 is denoted by the "x". Note that for RMM2, not all center ensemble mean ACs falls below 0.6 before the end of the forecast. ESPC-D only shown for RMM1 as score does not fall below 0.6 for RMM2 before the end of the forecast.

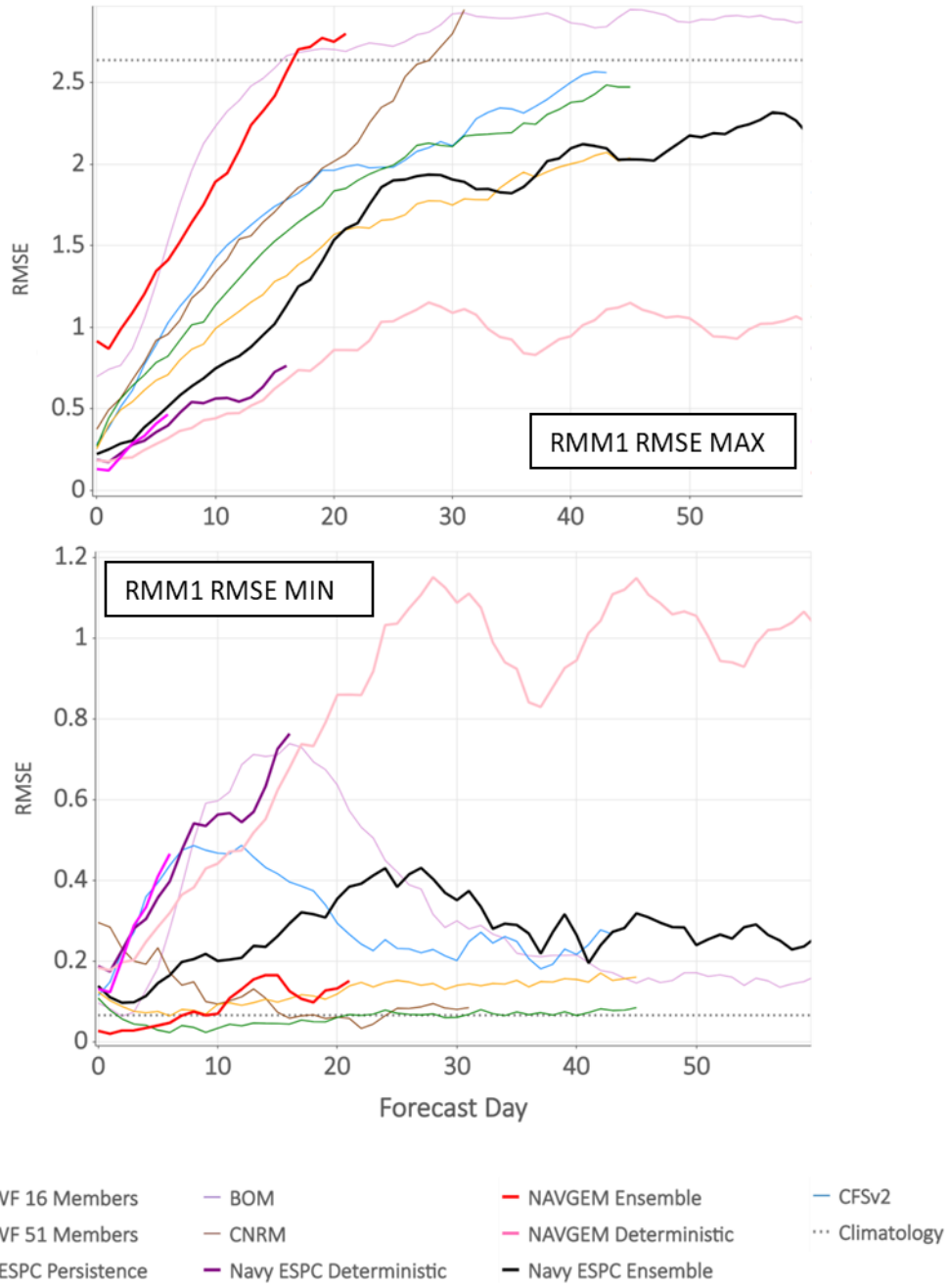

Figure 12: RMSE scores for RMM1 for the ensemble member with the largest RMSE (top) and smallest RMSE (bottom). Center colors are denoted in Key and include Australian Bureau of Meteorology (BOM), Meteo-France (CNRM), ECMWF 16-member and 51-member ensembles, NOAA (CFSv2), Navy ESPC high-resolution deterministic, Navy ESPC Ensemble, and NAVGEM deterministic, and NAVGEM Ensemble. Scores for persistence and climatological forecasts are also included.

Atmospheric Anomaly Patterns: Skill of Forecasts

Figure 13 shows the ACOR and RMSE for the forecast as a function of forecast lead time for the oscillations and forecast models described in Section 4.2.1. Also included are the forecast scores for a persistence forecast and for a climatological forecast. In Figure 13, the ensemble forecast members are

scored individually. Ensemble mean scores are shown in Figure 14. In this "deterministic" perspective, where the ensemble forecast members are scored individually, the Navy-ESPC ensemble-resolution and high-resolution deterministic forecasts are generally comparable to or better than the forecasts from the other centers as well as the NAVGEM ET, and substantially outperform persistence forecasts. The one exception is the AAO RMSE, where ECMWF, NOAA, and CNRM have better scores for the first two weeks. This is due to a bias in the prediction of the AAO. Note, the ACOR score, less affected by bias, is very good. Using the rule-of-thumb threshold of 0.6 for useful ACs, the predictive skill for these patterns typically dips below that threshold between 8 and 10 days. However, comparing dynamic forecasts to "climatological" forecasts randomly selected from a multi-decadal reanalysis archive (for the appropriate time of year), Navy ESPC forecasts outperform the climatological forecasts (dotted gray line) out past two weeks.

When evaluating the forecasts from an ensemble means perspective (Figure 14), the Navy forecast systems are still competitive, but the Navy ESPC system does not gain as much skill over the deterministic forecasts as is seen in some of the other center ensembles, which we presume is due to the under-dispersion of the ensemble forecasts. In particular note that while the NAVGEM ET individual forecasts are not as skillful as the Navy-ESPC individual ensemble forecasts (Figure 13), when considering the ensemble mean forecast skill (Figure 14), the two systems are usually of comparable skill. We also wish to note a caveat when interpreting the average skill of the individual members. The relatively good performance of Navy ESPC individual ensemble members is in part due to a relative lack of spread in the ensemble. As was illustrated in the previous subsection describing the MJO forecasts, this lack of ensemble spread means that the best members in the Navy ESPC ensemble are not as skillful as the best members at other centers, and likewise the worst members in the Navy ESPC ensemble have smaller errors than the worst members of the other centers. As the distribution of skill about the average score is not symmetric (scores can only get so good), this results in what is effectively a penalty against the systems that have larger ensemble spread when considering the mean skill of the individual members

Figure 15 shows Violin plots describing the distribution of the forecast day when the ACOR skill score of these ensemble forecasts (considered individually, that is, in a deterministic sense) drops below 0.6. The higher the violin plot, the longer the forecast system has skill, on average, above this somewhat arbitrary level. The Navy-ESPC ensemble forecast performance, in a deterministic sense, is for the most part comparable to or better than the performance of the other systems and outperforms the NAVGEM ET. The day that the ensemble means forecast skill falls below 0.6 is also included for these systems, and is indicated by the "X". The Navy-ESPC ensemble mean remains skillful usually about a half day or day longer than the median of the individual forecasts. Other centers are usually able to gain two, sometimes three days in skill using the ensemble mean. NAVGEM ET gain in skill between deterministic and ensemble for the AO and PNA is particularly large, attesting to the merits in the ensemble design. Thus the ensemble mean skill between NAVGEM ET and Navy-ESPC is comparable (NAVGEM ET has n slight edge for the AO, and Navy-ESPC has an edge for the AAO). We expect we will be able to realize even more benefit from the Navy-ESPC ensemble forecasts once issues with underdispersion have been addressed, employing techniques successfully used in NAVGEM ET, such as Stochastic Kinetic Energy Backscatter and other methods. The skill of the Navy-ESPC high-resolution deterministic forecast is also shown, and is comparable to the skill of the Navy-ESPC ensemble forecasts considered individually, and generally within a day difference of the ensemble mean score.

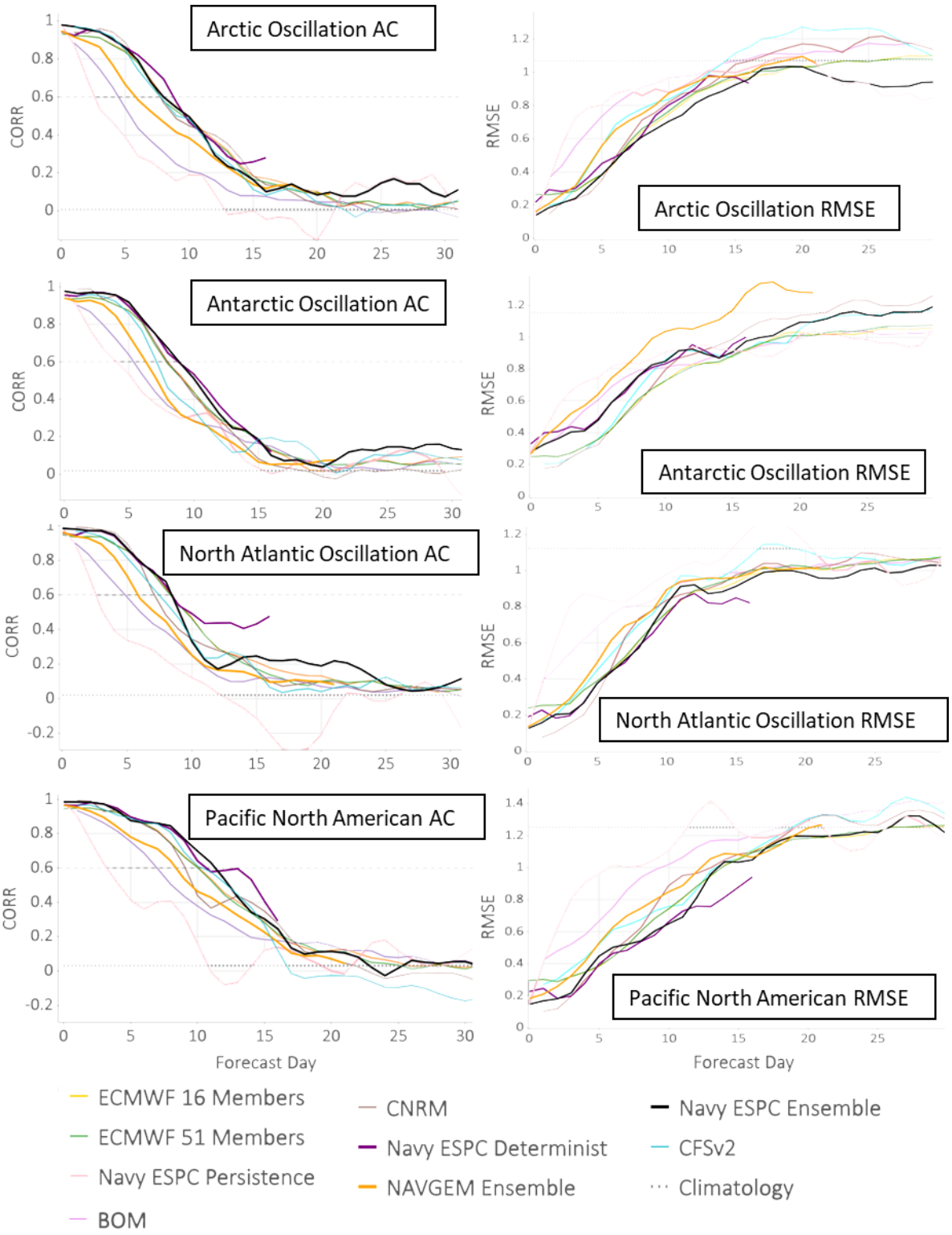

Figure 13: ACOR scores, left, and RMSE scores, right, for the AO, AAO, NAO and PNA. Scores are calculated for individual forecast members and averaged for each center. Center colors are denoted in Key and include Australian Bureau of Meteorology (BOM), Meteo-France (CNRM), ECMWF 16-member and 51-member ensembles, NOAA (CFSv2), Navy-ESPC high-resolution deterministic, Navy-ESPC

nsemble, and NAVGEM Ensemble. Scores for persistence and climatological forecasts are also included. The RMSE values are defined against each index.

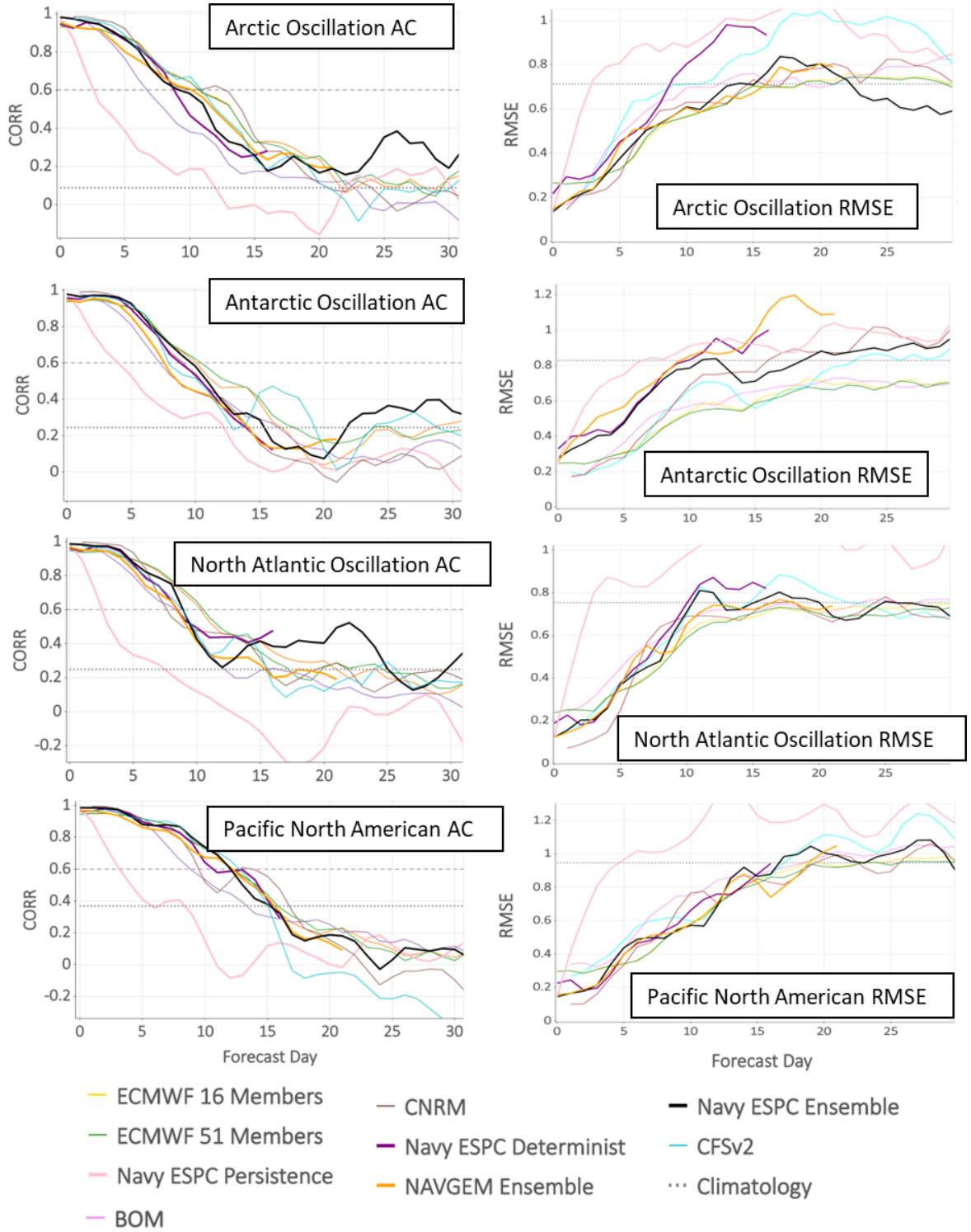

Figure 14: Same as Figure 13 except that scores are for the ensemble mean forecasts. Navy-ESPC high-resolution deterministic forecast is also shown.

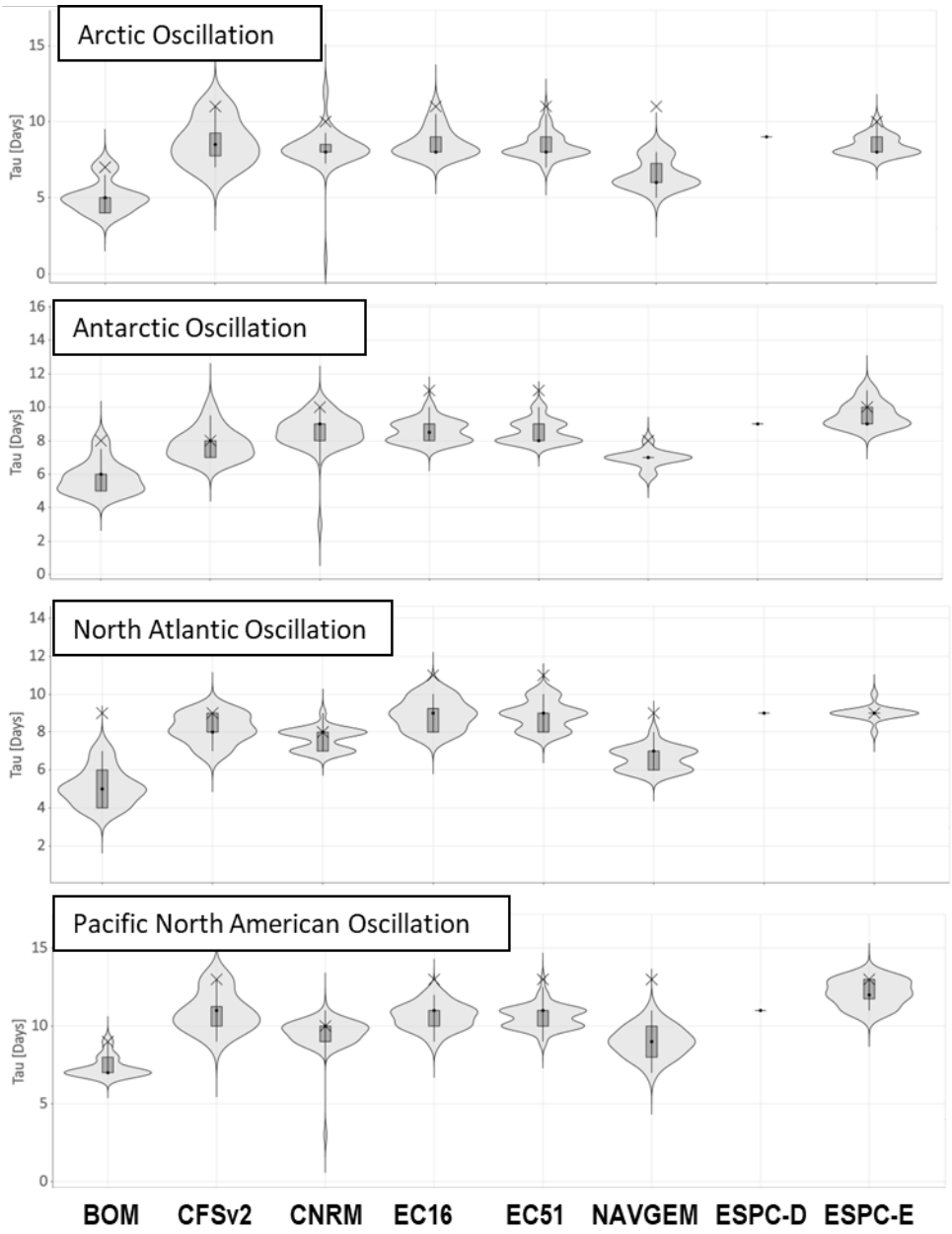

Figure 15: "Violin" plots showing the distribution of the day that the forecast of the large-scale oscillation (as denoted in panels) ACOR falls below 0.6. Each ensemble member is treated as an individual (deterministic) forecast, and the width of the violin is proportional to the number of ensemble members that fall below 0.6 on that day. The forecasts considered come from the Australian Bureau of Meteorology (BOM), NOAA (CFSv2), Meteo-France (CNRM), ECMWF 16-member (EC16) and 51-member (EC51) ensembles, NAVGEM ET (NAVGEM), Navy-ESPC high-resolution deterministic forecasts (ESPC-D), and the Navy-ESPC ensemble (ESPC-E). The day on which the ensemble mean forecast ACOR falls below 0.6 is denoted by the "x".

Ensemble Synoptic Score Card

Here we consider the synoptic scorecard sum by metric, including all regions and all variables (Figure 16). We can see wide variations in how long the forecast skill exceeds that of climatology (defined in section 4.2.1) for a particular metric. Values ranging from 6.5 days for CRPS, 7.0 days for BIAS, 8 days for RMSE, to 15 days for ACOR. Note that VARR remains below the skill of the climatological ensemble for the entire time period considered due to the under-dispersion of initial conditions in the Navy-ESPC ensemble (section 5.1). In addition, a climatology based on different years represents a high amount of variability for initial conditions.

Evaluating performance by variable (Figure 16b), also indicates large differences in how long the forecast remains more skillful than the climatological ensemble. 10-m wind speed is skillful above climatology out to about 4.5 days, while 250-hPa wind speed is skillful out to 6 days. 2-m temperature remains below the skill of climatology for the period, due to temperature biases, while 850-hPa temperature is above or at climatological skill levels out to 7.0 days. 500-hPa geopotential height retains the most skill above climatology for the longest time period, out to 9 days, reflecting that 500-hPa geopotential height represents larger-scale phenomena compared to the other variables. There is no discernable difference by region (Figure 16c) with the Navy-ESPC forecast skill beating the climatological ensemble out to between 5.5 and 6.0 days for all three regions considered.

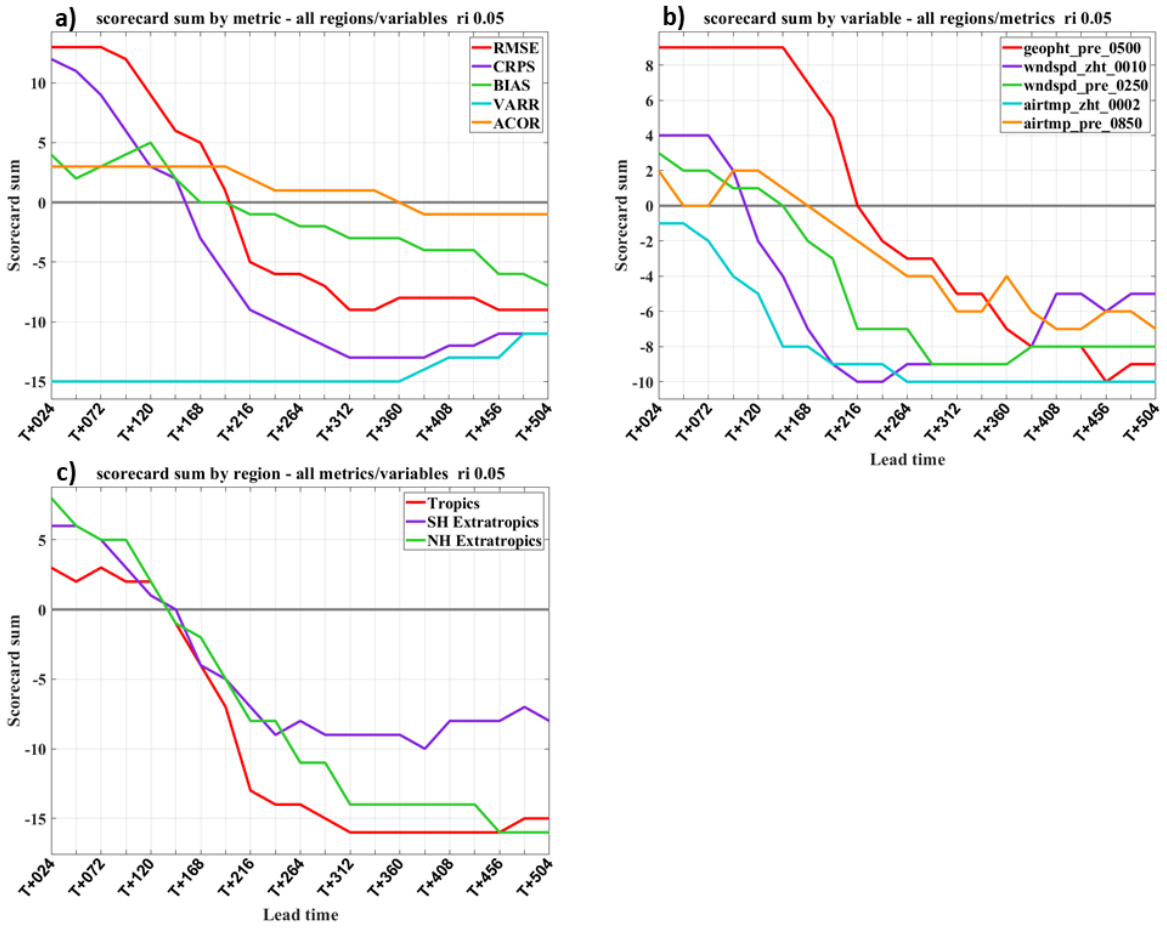

Figure 16: Scorecard sum for the Navy-ESPC ensemble relative to a climatological ensemble in a) as a function of metric, aggregating all regions and all variables; b) as a function of variable, aggregating all

regions and metrics; c) as a function region, aggregating all metrics and variables. The different color curves correspond to the different metrics, variables, or regions as given in the keys in each panel.

In Figure 17a, we compare the grand total scorecard sum ensemble skill relative to the climatological ensemble, aggregating over all regions, metrics and variables, for the Navy-ESPC ensemble and the NAVGEM operational ET ensemble. Figure 17b again shows forecast skill over climatology for the Navy-ESPC and NAVGEM ET, but in addition also shows results for three "deterministic" forecasts; the Navy-ESPC deterministic high-resolution forecast, the Navy-ESPC ensemble control member, and the NAVGEM ET ensemble control member. Another difference between Figure 17a and Figure 17b is that Figure 17b only includes scores for the "deterministic" metrics, that is ACOR, BIAS, and RMSE, while Figure 17b also includes VARR and CRPS. Several important conclusions can be drawn from this plot. First, the difference in performance between the Navy-ESPC ensemble and the NAVGEM ET ensemble is smaller when excluding the VARR and CRPS metrics, which are the metrics most greatly impacted by ensemble design. Second, the performance of the Navy-ESPC and NAVGEM ET control members is very similar, indicating that the gap in performance is due to ensemble design, not the forecast skill of the individual members. The skill of the coupled high-resolution Navy-ESPC is not quite as good as the ensemble control members, although for other metrics, the coupled high-resolution Navy-ESPC does show advantages over the ensemble-resolution Navy-ESPC.

Unlike the MJO and teleconnection metrics, the NAVGEM ET ensemble outperforms the Navy-ESPC ensemble considerably using this scorecard. This is due to the spread in the Navy-ESPC ensemble compared to the NAVGEM ET ensemble. In particular the initial spread is larger in the NAVGEM ET ensemble, the NAVGEM ET ensemble uses Stochastic Kinetic Energy Backscatter –moisture convergence (SKEB-MC) method to improved spread throughout the forecast, and the NAVGEM ET ensemble has 20 members compared to 16 for the Navy-ESPC ensemble. These differences give an advantage to the NAVGEM ET ensemble particularly when examining the VARR and CRPS metrics. Note that the forecast model in the Navy-ESPC ensemble is much more computationally intense compared to the NAVGEM ET ensemble and running more members than 16 is not trivial.

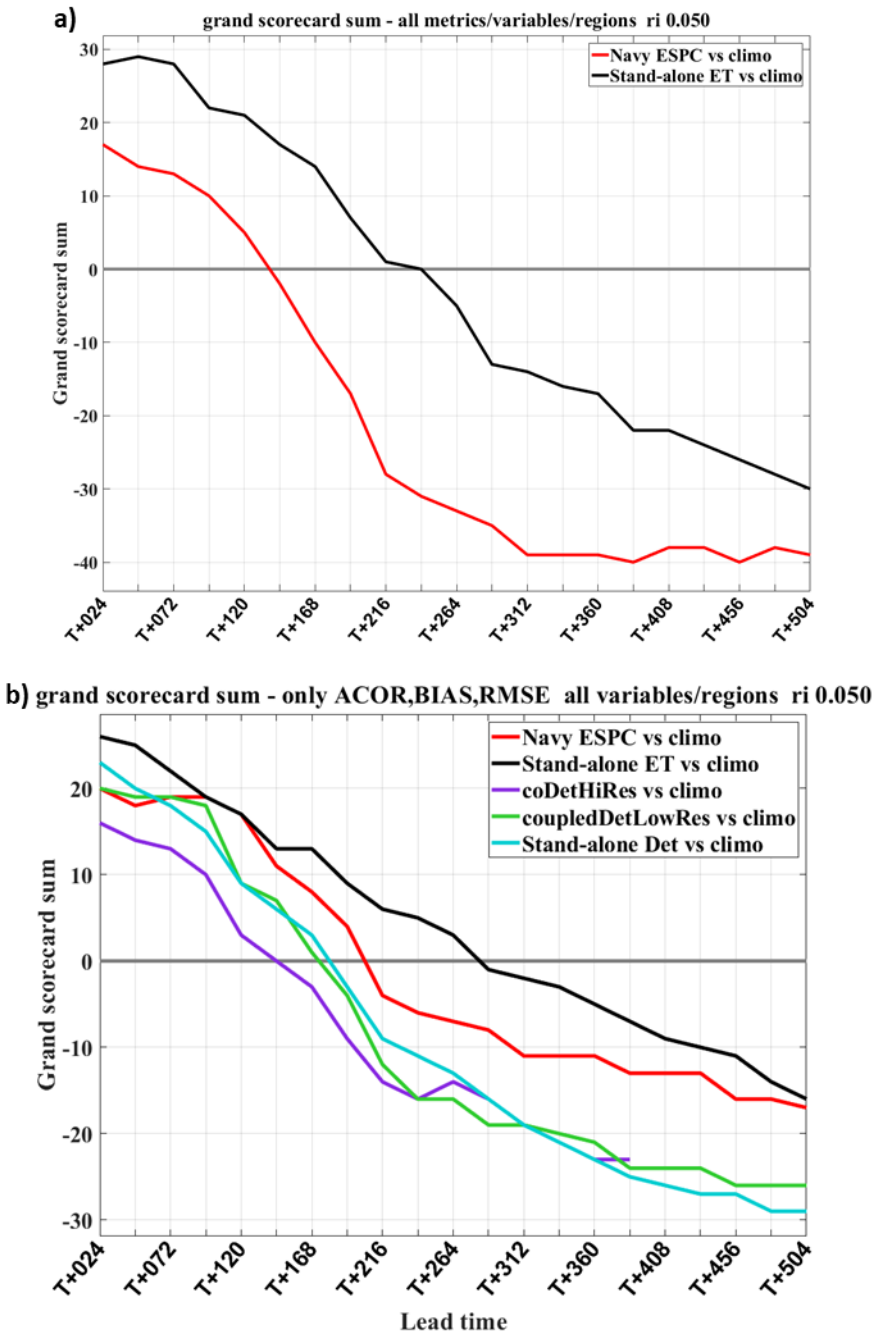

Figure 17: a) Grand total scorecard sum (aggregating over all regions, variables, and metrics) for the Navy-ESPC ensemble (red) and NAVGEM ET ensemble (black). b) Grand scorecard sum aggregated for all variables and regions for deterministic metrics (ACOR, BIAS, RMSE) only for the Navy-ESPC ensemble (red), NAVGEM ET ensemble (black), Navy-ESPC high-resolution deterministic forecasts (purple), Navy-ESPC ensemble control member (green), and NAVGEM ET ensemble control member (cyan).

There are some aspects of the Navy-ESPC ensemble that do clearly outperform the NAVGEM-ET, and this is due to the fact that the physical parameterizations have been tuned to provide better performance at long lead times, particularly in regards to tropical phenomena and the MJO (see section

2.1.4 for a description of the differences). Figure 18 shows the bias as a function of lead time for 10-m winds and 2-m temperature for the Navy-ESPC ensemble, NAVGEM ET ensemble, and the Navy-ESPC high-resolution deterministic forecasts. For the wind speed bias, the Navy-ESPC ensemble and NAVGEM ET ensemble have comparable bias magnitude, although of opposite sign, in the northern extratropics, while in the tropics and the southern extratropics, the magnitude of the wind speed bias in the coupled system, both deterministic and ensemble, is lower than that of the NAVGEM ET system. For 2-m temperature, the Navy-ESPC has a larger bias than that of NAVGEM ET in the extratropics, but a smaller bias in the tropics. We do note that Navy ESPC will not replace the NAVGEM ET due to latency issues.

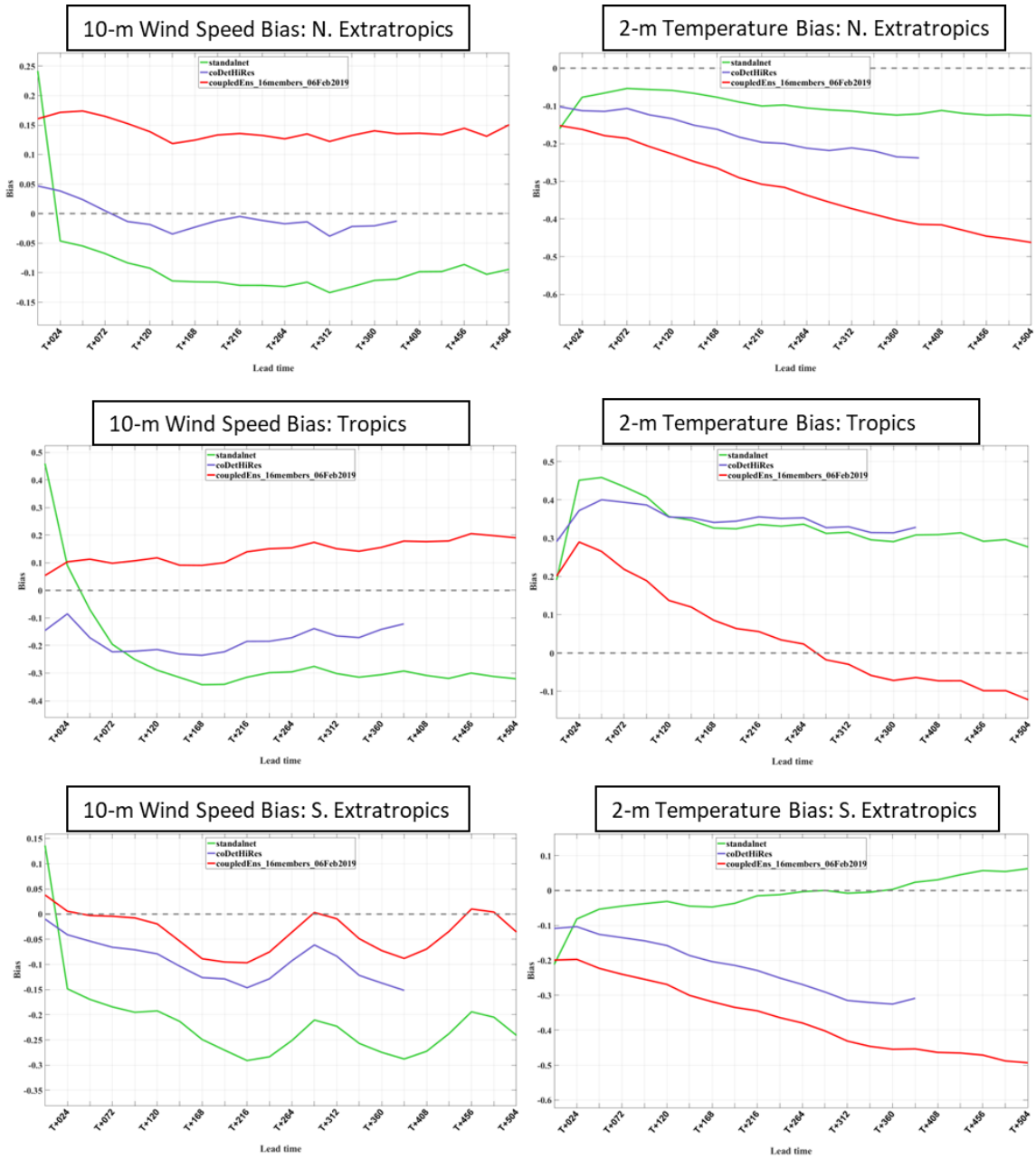

Figure 18: 10-m wind speed biases in m s^{-1} (left) and 2-m temperature biases in °C (right) for the Northern Extratropics (top), Tropics (middle) and Southern Extratropics (bottom) for the NAVGEM ET

ensemble (green), Navy-ESPC ensemble (red), and Navy-ESPC high-resolution deterministic forecast (purple) using ECMWF analyses as verification.

Cloud and Radiation Biases

Figure 19 displays the total cloud cover (i.e., fraction) biases (Navy-ESPC – verifying analysis) for the Navy-ESPC ensemble mean compared against ECMWF ERA-Interim analysis (left column) and the US. Airforce WWMCA (Roadcap et al. 2015) cloud analysis (right column). Rows of the figure show biases for progressively increasing taus: 0 days, 7 days, 14 days, and 60 days. Biases against both datasets suggest that Navy-ESPC has too high total cloud cover in the tropics, over ice, and over the desert regions of the Africa and the Middle East. In the mid-latitudes, Navy-ESPC may not have enough clouds. Note, the ECMWF analysis has too high of total cloud fraction when compared to satellites (Stengel et al. 2018), and the WWMCA analysis also has uncertainties (Cleary 2012), hence biases from both produces are shown.

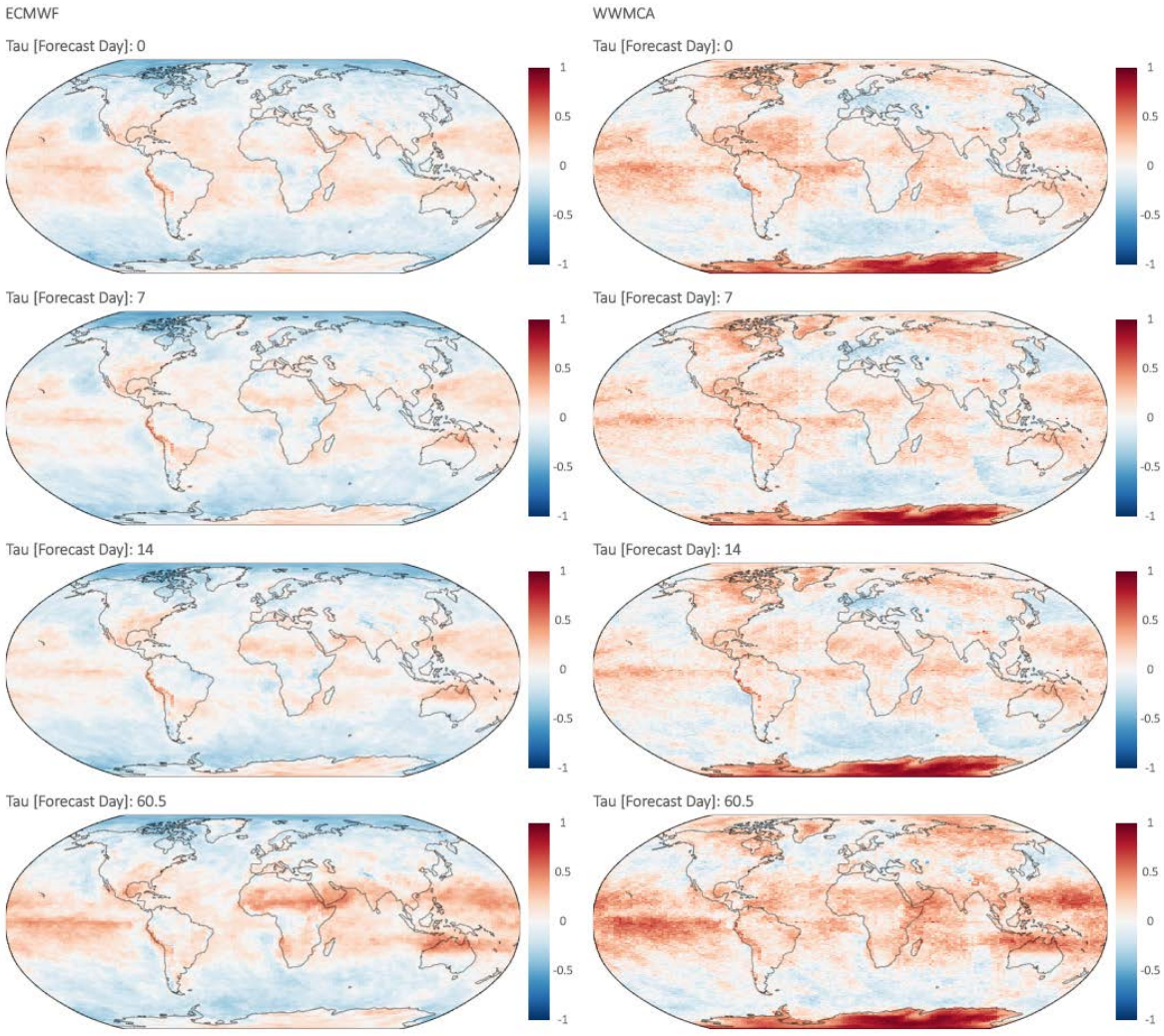

Figure 19: Total cloud BIAS for the Navy-ESPC system for all ensemble forecasts compared to (left) ECMWF analysis and (right) World Wide Merged Cloud Analysis (WWMCA) (Roadcap et al. 2015) for the forecast taus of 0.5 hours, 7 days, 14 days, and 60.5 days.

Figure 20 displays flux biases (Navy-ESPC – verifying analysis) compared against NASA CERES surface flux estimates (Loeb et al. 2018). At initial time the biases in long and short wave mirror cloud cover biases. The largest shortwave biases are located along the ITCZ and along the West coast of China. Positive SW biases exist off the western coast of California and South America. Low-level stratocumulus clouds occur in these regions, and these types of clouds are difficult to simulate in global atmospheric models. The positive SW biases suggest that not enough clouds occur in these regions. Biases along the ITCZ are most likely due to representation of clouds in the model, as shown in Figure 19. The negative shortwave bias over Antarctica suggests that Navy-ESPC system has albedos that are too large. The positive biases in the surface net LW radiation suggest SSTs are becoming too high in the Navy-ESPC system.

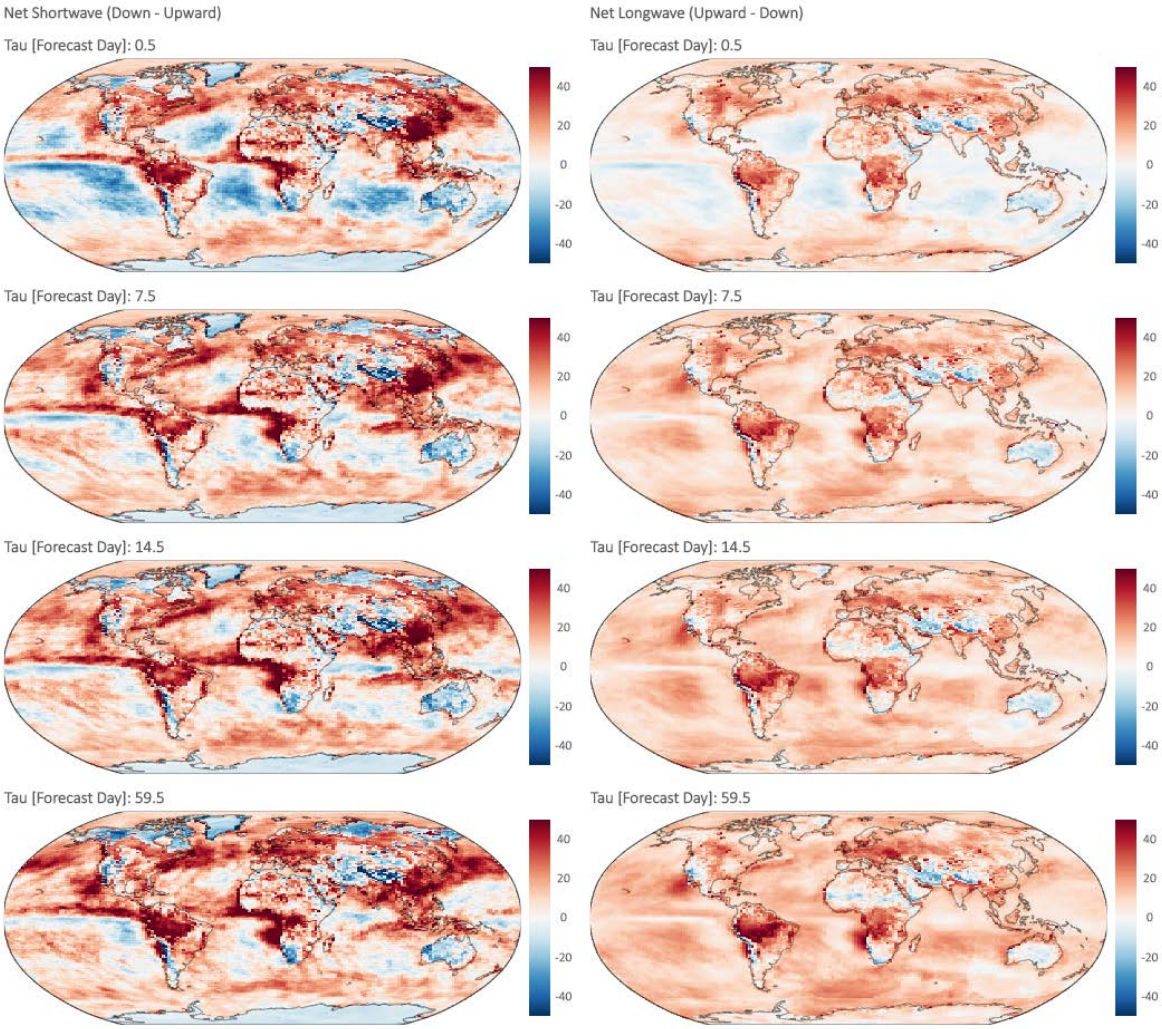

Figure 20: (left) Surface net shortwave and (right) surface net longwave radiation BIAS for the Navy-ESPC system compared to the NASA CERES surface flux estimates for the forecast taus of 0.5 hours, 7 days, 14 days, and 60.5 days.

5.2.2. Ocean

Ensemble Variability

Maps of the ensemble spread are useful as a qualitative assessment of the perturbed-observation technique used to generate the ensemble. Figure 21 and Figure 22 show the ensemble standard deviation in temperature and salinity at 100m for the 30 d and 60 d forecast times. The global distribution of the temperature spread is consistent with areas of western boundary currents, the Arctic Circumpolar Current, and the tropics where there is high temporal and spatial variability. The low temperature variability in ice-covered areas is likely due in part to the relative scarcity of ocean observations in those areas, which limits the impact of perturbations added through observations in the perturbed-observation analysis, but is also not necessarily grossly wrong. Overall, the uncertainty represented by the spread is distributed as expected based on historical observations, shows expected seasonality, and grows with forecast lead time.

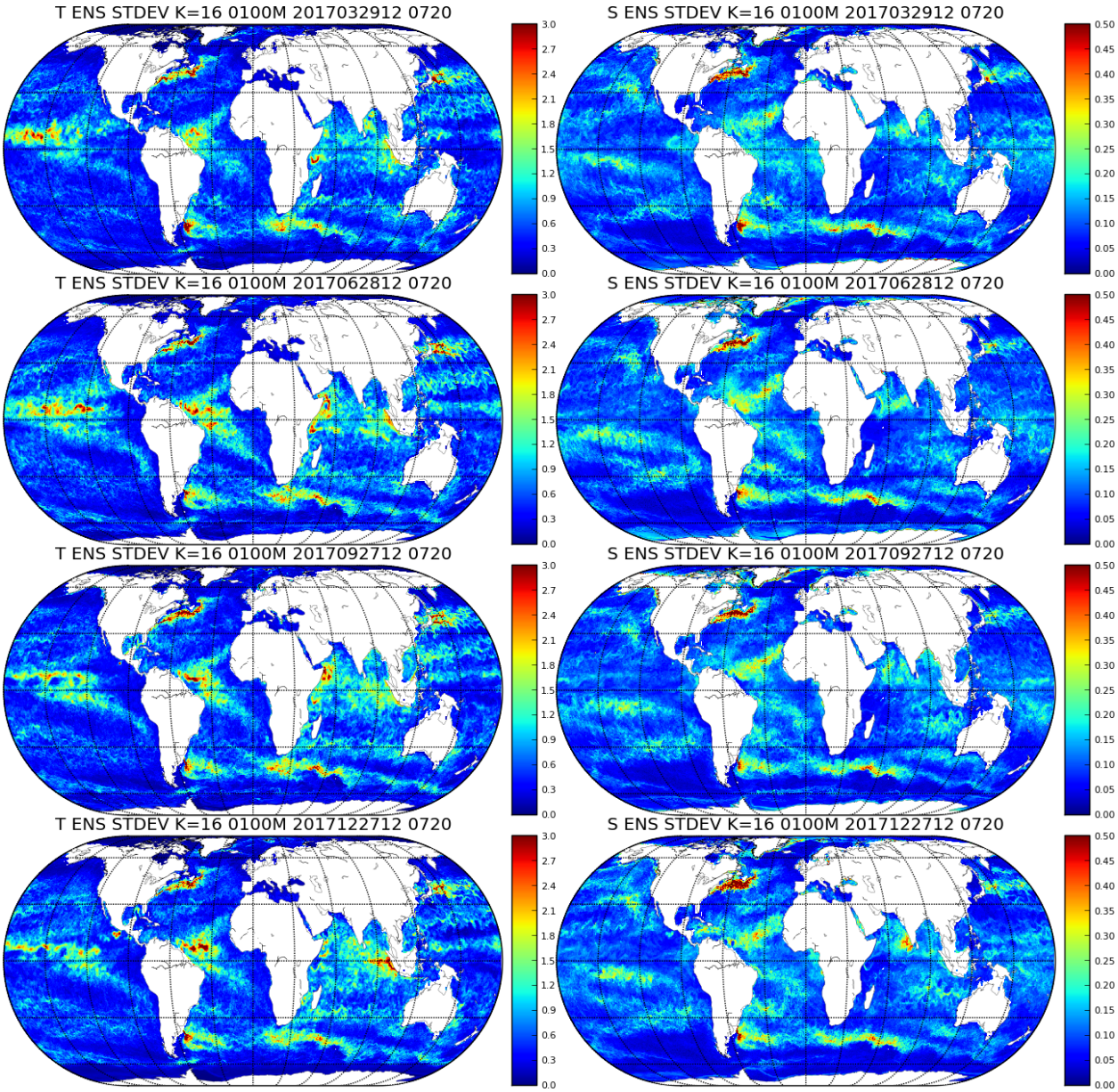

Figure 21: (left) Temperature and (right) salinity ensemble standard deviation at 100m for the 720 h (30 d) forecast lead time for the 29 Mar, 28 Jun, 27 Sep, and 27 Dec 2017 forecasts.

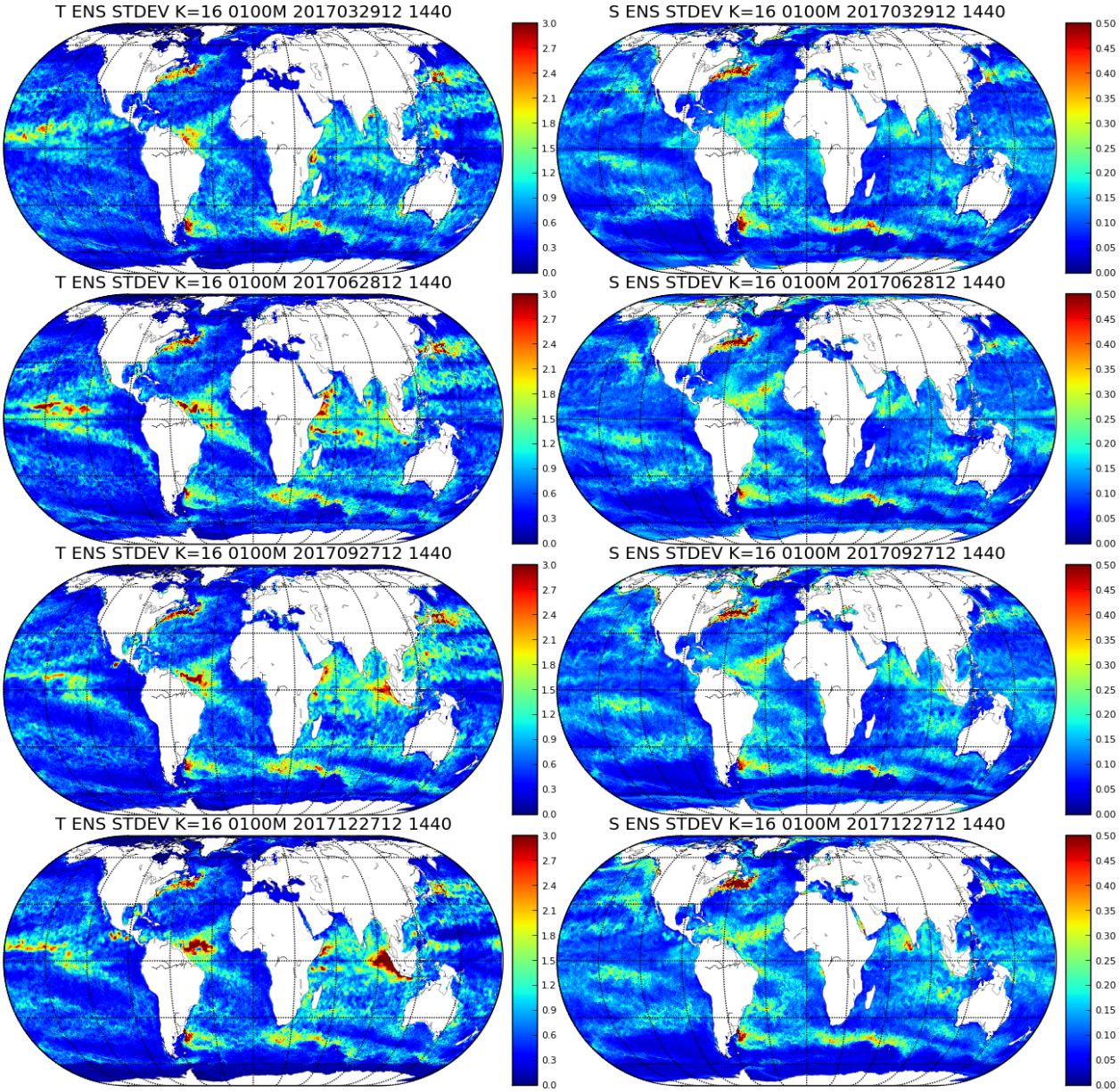

Figure 22: (left) Temperature and (right) salinity ensemble standard deviation at 100m for the 1440 h (60 d) forecast lead time for the 29 Mar, 28 Jun, 27 Sep, and 27 Dec 2017 forecasts.

Ensemble Mean Temperature Forecast

The extended-range ensemble-mean temperature BIAS and RMSE are shown in Figure 23. The Navy-ESPC ensemble mean forecast has skill relative to the climatology — a lower RMSE than the GDEM v4 climatology — out to about 35 days in the forecast, as indicated by the crossing of the RMSE lines in the left panel. The RMSE of the control member (the member without perturbed observations) crosses climatology at a lead time of 10 days. The right panel shows the RMSE, BIAS, and absolute error for the

Navy-ESPC ensemble mean and the GDEM v4 climatology for the upper ocean to 500 m. The Navy-ESPC ensemble mean has a small negative bias at the initial time that decreases overall and becomes positive near the surface (this is consistent with the longwave radiation bias above). The GDEM v4 climatology has a consistent and somewhat larger negative bias. The RMSE of the Navy-ESPC ensemble mean at the start of the forecast is noticeably lower than the GDEM reference, grows quickly in the upper thermocline, and at 100 m becomes larger than climate (> 1.3 °C) between forecast days 26 and 40.

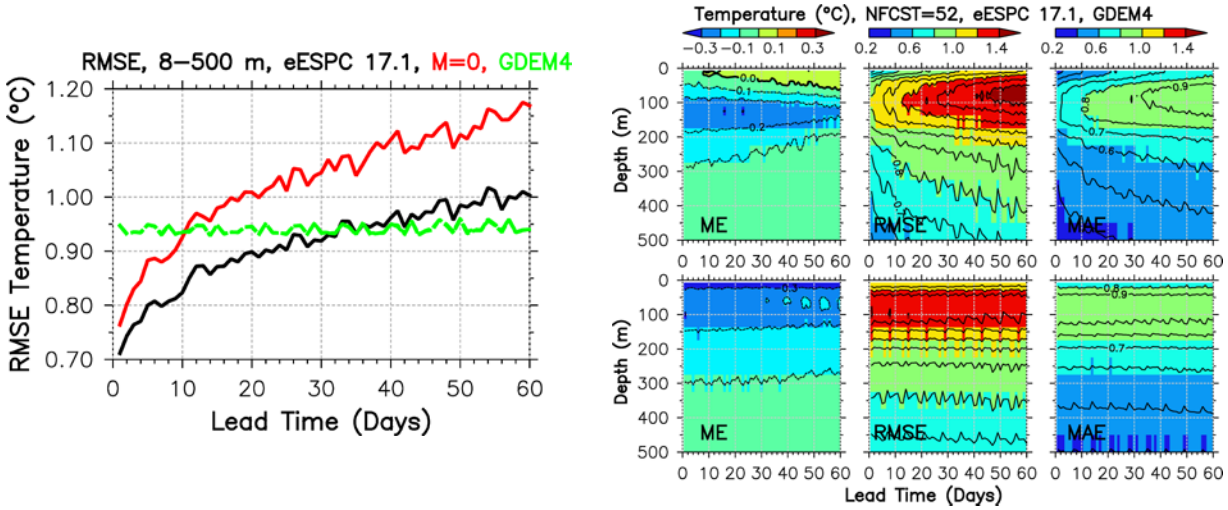

Figure 23: Mean RMSE of ensemble mean (black line), control member (red line), and GDEM v4 climatology (green line) ocean temperature over 8-500 m depth as a function of forecast lead time (left), and temperature BIAS (ME), RMSE, and mean absolute error (MAE) for the upper 500 m for 1 – 60 day forecast time for the ensemble mean (top row) and the GDEM v4 monthly climatology (bottom row).

The time series of the ensemble mean temperature RMSE and BIAS are compared with GDEM in Figure 24, showing the decreasing negative bias and the growth in the RMSE, together with the expected growth in the ensemble temperature standard deviation. The vertical profiles of temperature RMSE and BIAS averaged over the 60 day forecast show the overall better performance of the ensemble mean compared to GDEM (though this includes the period from 35 day to 60 day over which the ensemble mean has lower skill). The time average spread to RMSE ratio for temperature in the upper 500 m is 0.60 and increases to 0.67 after subtracting BIAS from RMSE (RMSE – |ME|). The Navy-ESPC is accounting for 67% of uncertainty.

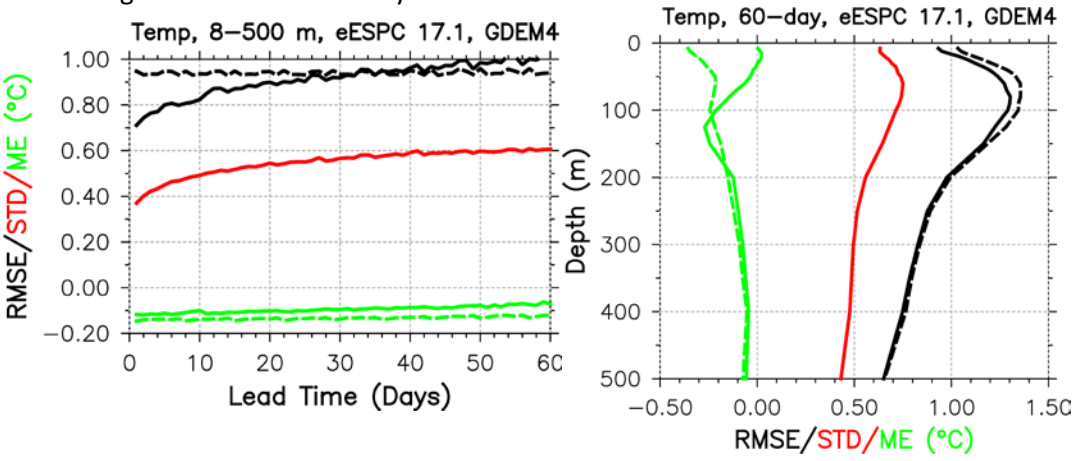

Figure 24: Ocean temperature RMSE, BIAS (ME), and standard deviation of the ensemble mean forecast over 8-500 m depth as a function of forecast lead time (left), and averaged over the 60 day forecast as a function of depth (right). The GDEM BIAS and RMSE are included as dashed lines.

Ensemble Mean Salinity Forecast

The extended-range ensemble-mean salinity BIAS and RMSE are shown in Figure 25. The Navy-ESPC ensemble mean forecast has skill relative to the climatology — a lower RMSE than the GDEM v4 climatology — out to about 20 days in the forecast. The RMSE of the control member (red) has skill up to lead time less than 10 days relative to climatology. The right panel shows the salinity errors are largest near the surface, and the Navy-ESPC ensemble mean develops a positive bias at the surface and a negative bias in the upper thermocline, and the RMSE and absolute error are comparable to GDEM. The time series of the ensemble mean salinity RMSE and BIAS are compared with GDEM in Figure 26, showing the increasing negative bias and the growth in the RMSE, together with the expected growth in the ensemble salinity standard deviation. The vertical profiles of salinity RMSE and BIAS averaged over the 60 day forecast show performance of the ensemble mean is comparable to GDEM (again this includes the period from 20 day to 60 day over which the ensemble mean has lower skill). The time average spread to RMSE ratio for salinity in the upper 500 m is 0.55 and increases to 0.60 after subtracting BIAS from RMSE, (i.e. RMSE – |ME|). The Navy-ESPC is accounting for 60% of the uncertainty (somewhat smaller than for the temperature).

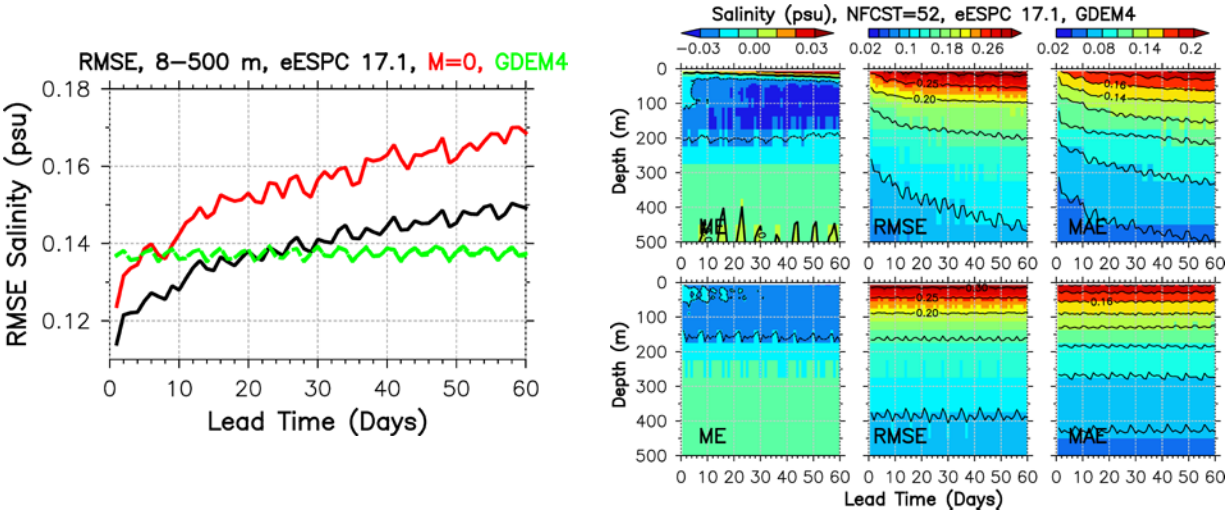

Figure 25: As in Figure 23, for salinity.

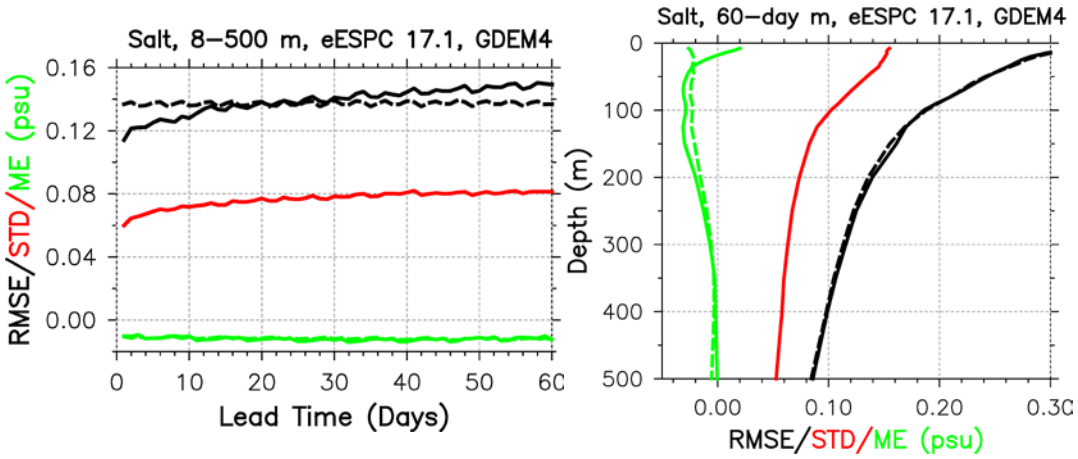

Figure 26: As in Figure 24, for salinity.

Ensemble Mean Temperature Predictability

To evaluate the spatial distribution of ensemble temperature forecast skill, the observations and model are binned into 40°×20° grids. For each bin, the RMSE is calculated with respected to the ensemble mean (EM), control member (M=0), and climatology. Next, the RMSE in each bin is averaged over 8-500 m depth. Last, the forecast day on which model RMSE crosses the climatology is identified for each bin, and plotted in Figure 27. Bins with fewer than 50 matchup comparisons are excluded. Note that the time-series in each bin is smoothed using a 5-day boxcar method before finding the day of crossing, and in the event of multiple crossings, the earliest occurrence is selected. The ensemble mean performs better than climatology out to 60 days in the Pacific Ocean except in the Kuroshio extension east of Japan. The ensemble mean seems to have a slightly lower predictability to 25-35 days relative to climatology in the Atlantic, Northern Indian Ocean and Indonesian Seas. Part of the lower predictability in the Indian Ocean regions likely stems from the dominance of the seasonally reversing monsoon. As expected, the control member (Figure 27, bottom) has lower predictability than the ensemble mean almost everywhere in the globe. Again, the geographical pattern is consistent with the ensemble mean with control member showing lower predictability out to 15-25 days in both Atlantic and Indian Ocean regions. The Gulf Stream region has predictability out to only 10 days relative to climatology.

To compare with the predictability in the deterministic unperturbed control member, Figure 28 shows the RMSE with depth and forecast time for the control member, the ensemble mean forecast, and the GDEM climatology. The forecast lead time after which the forecast RMSE exceeds the GDEM value is a measure of the predictability of the model. The ensemble mean forecast time with skill over climatology is significantly longer than for the single member at the same resolution at all depths. At 100 m, the control member reaches the GDEM RMSE of 1.3 °C after about 14 days, while the ensemble mean forecast exceeds that value after about 26 d.

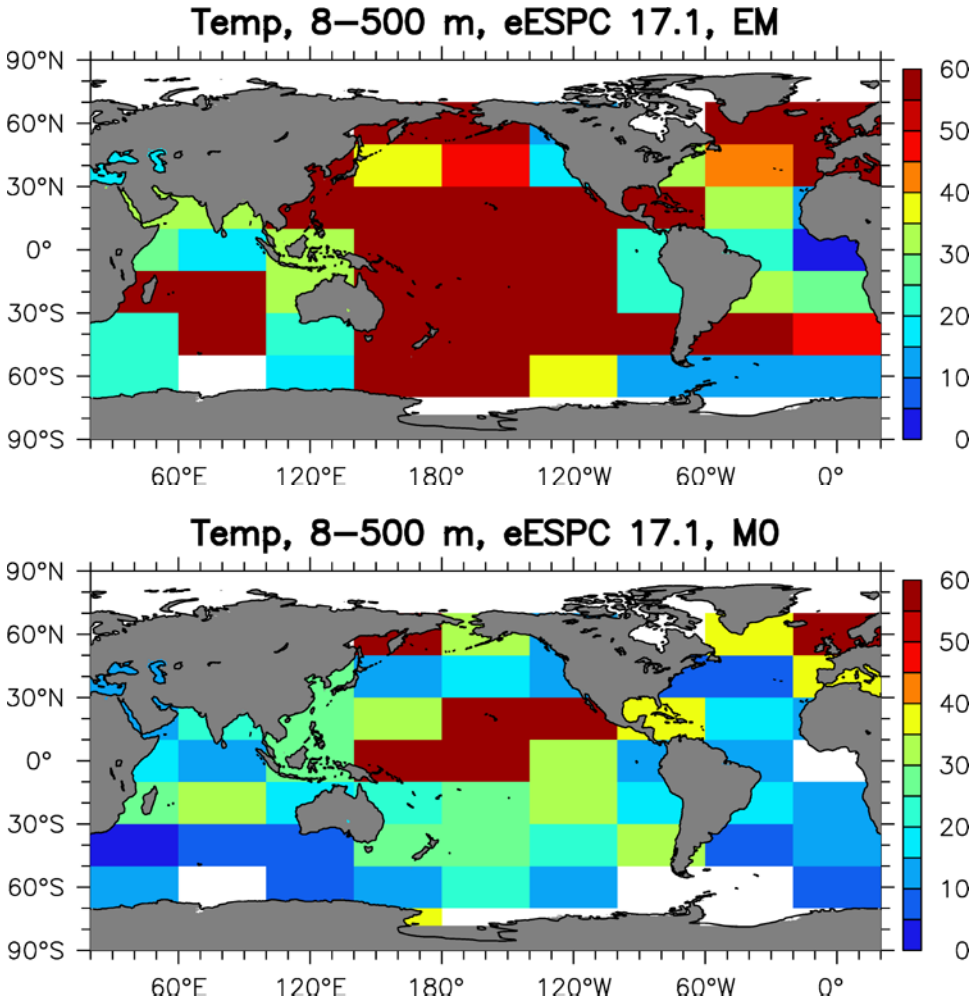

Figure 27: Spatial distribution of the forecast day when the model RMSE crosses the climatological RMSE for (top) the ensemble mean forecast and (bottom) the control member using temperature profile observations. A 60-day crossing suggests the model has predictability out to 60 days or longer.

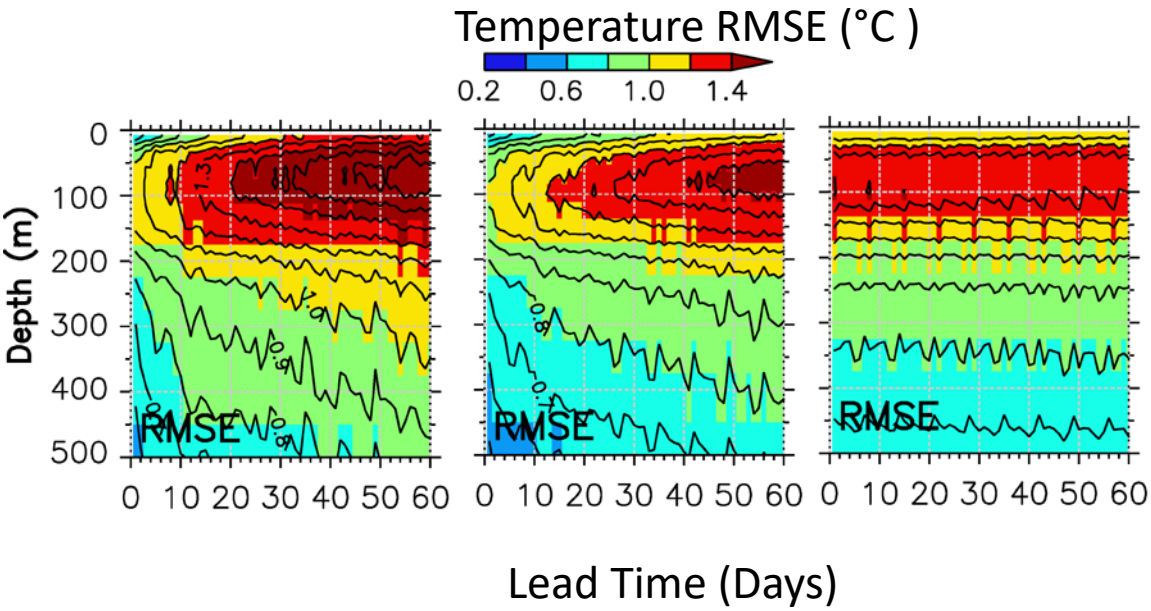

Figure 28: Temperature RMSE for the upper 500 m for 1 – 60 day forecast time for the deterministic control member (left), ensemble mean (center) and the GDEM v4 monthly climatology (right). The lead time beyond which the forecast RMSE at a depth exceeds the GDEM RMSE is an estimate of the model predictability.

Ensemble Mean SST Forecast

Ship surface temperature observations, mostly from drifters and ships tracks, are used to evaluate the ensemble forecasts along with the GDEM v4 climatology. Note that number of ship observations is significantly higher (~30000/day) than profiles observations. The RMSE of the ensemble mean (black line), control member (unperturbed member) (red line) and GDEM v4 climatology are shown in Figure 29 (left) and right panel shows the BIAS and the standard deviation. For comparisons, the errors from 15 ensemble members are shown as grey symbols. The Navy-ESPC has lower BIAS and RMSE than climatology at all lead times out to 60 days. The Navy-ESPC BIAS is near zero at lead times up to 40 days, which is significantly smaller than GDEMv4 at 0.4°C. The control member has skill relative to the climatology out to 20 days in the forecast. The ensemble standard deviation (spread) growth is consistent with the RMSE. However, the Navy-ESPC is under-dispersive as indicated by the spread lower than RMSE, suggesting that dispersion of the ensembles does not account for all the error growth. The time-average spread to RMSE ratio of ~0.72 (ranges between 0.52 and 0.76) suggests that the ensemble is accounting for 72% of uncertainty.

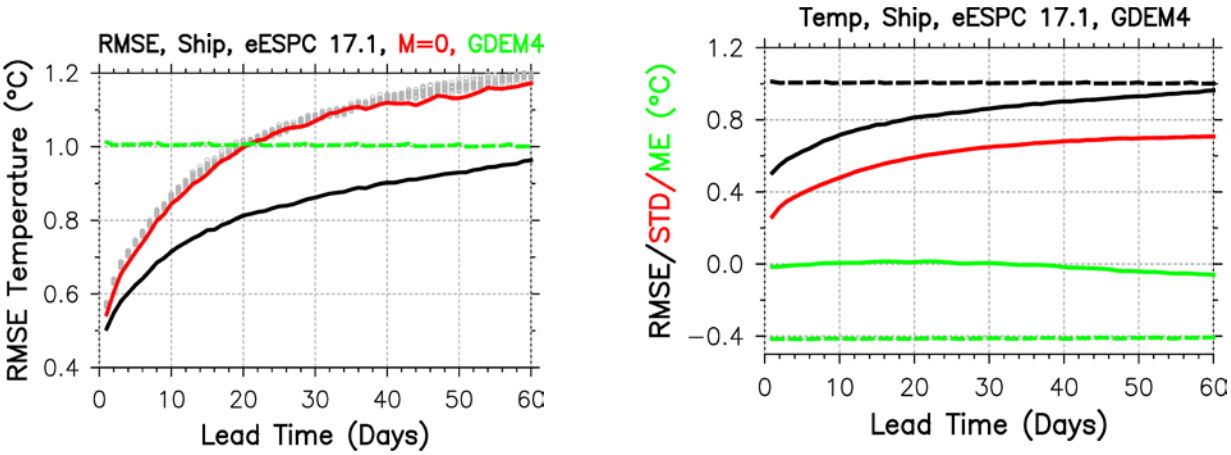

Figure 29: (left) RMSE of ensemble mean using ship surface temperature observations (back line), control member (red line) and GDEM v4 climatology (green line). Grey symbols are RMSE of 15 ensemble members treated independently. (right) The BIAS (ME) and standard deviation (STD) as a function of forecast lead time. Dashed line indicates GDEM v4 climatology.

Ensemble Mean SST Predictability

To evaluate the spatial distribution of ensemble forecast skill for ship based SSTs, the observations and model are binned into 10°×10° grids. For each bin, the RMSE is calculated with respected to the ensemble mean (EM), control member (M=0), and climatology. Next, the RMSE in each bin is averaged over 8-500 m depth. Last, the forecast day on which model RMSE crosses the climatology is identified for each bin, and plotted in Figure 30 (top). Bins with fewer than 100 matchup comparisons are excluded. Note that the time-series in each bin is smoothed using a 5-day boxcar method before finding the day of crossing, and in the event of multiple crossings, the earliest occurrence is selected. The ensemble mean performs better than the climatology out to 60 days in most regions with the exception of regions of western boundary currents, the Antarctic circumpolar current (ACC), and wind-dominated regions. The relatively lower predictability in the western Arabian Sea, Bay of Bengal, and Indonesia regions can be attributed to the monsoon winds. The gap winds along the coast of Mexico appear to influence the SST predictability, lowering predictability out to about 15 days. Lower predictability is also evident in the Agulhas current region (20-25 days), where the generation of mesoscale eddies dominates the SST variability. The control member (Figure 30**,** bottom) has lower predictability than the ensemble mean almost everywhere in the ocean. The tropical Pacific shows improved predictability out to 60 days. As for the ensemble mean forecast, the Gulf Stream, ACC, and Agulhas region have lower predictability, here out to less than 10 days relative to climatology.

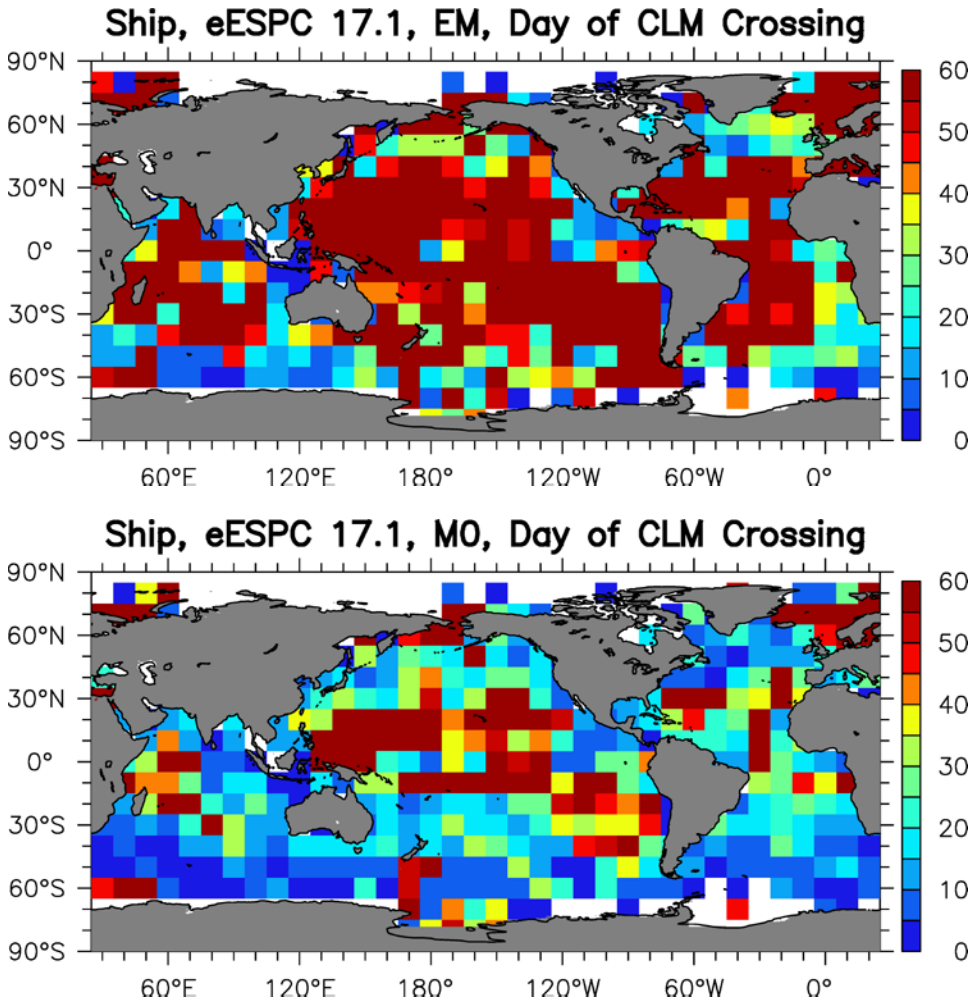

Figure 30: Spatial distribution of forecast day the model RMSE crosses the climatological RMSE for (top) the ensemble mean and (bottom) control member using ship SST observations. A 60-day crossing suggests model has predictability out to 60 days or longer.

Ensemble Mean Isotherm Depth Forecast

The ocean model prediction of selected isotherm depths is evaluated using the observation matchup files. We have evaluated the errors in the three depths of selected isotherms, 26°C, 20°C and 15°C. A snapshot of typical depths for these isotherms is shown in Figure 31 (left). The 26°C isotherm is shallow with depths less than 150 m and surfacing poleward of 30°, and has been used as a proxy for heat content in the hurricane generation and intensification. The 20°C isotherm is typically in the upper thermocline with depths less than 250 m and surfacing poleward of 40°. This isotherm depth is sensitive to the ocean convergence/divergence zones and planetary wave propagation. The 15°C isotherm is typically in the lower thermocline present everywhere in the tropics and subtropics with depths less than 500 m except in the western North Atlantic, where the thermocline is much deeper and this isotherm is found around 700 m.

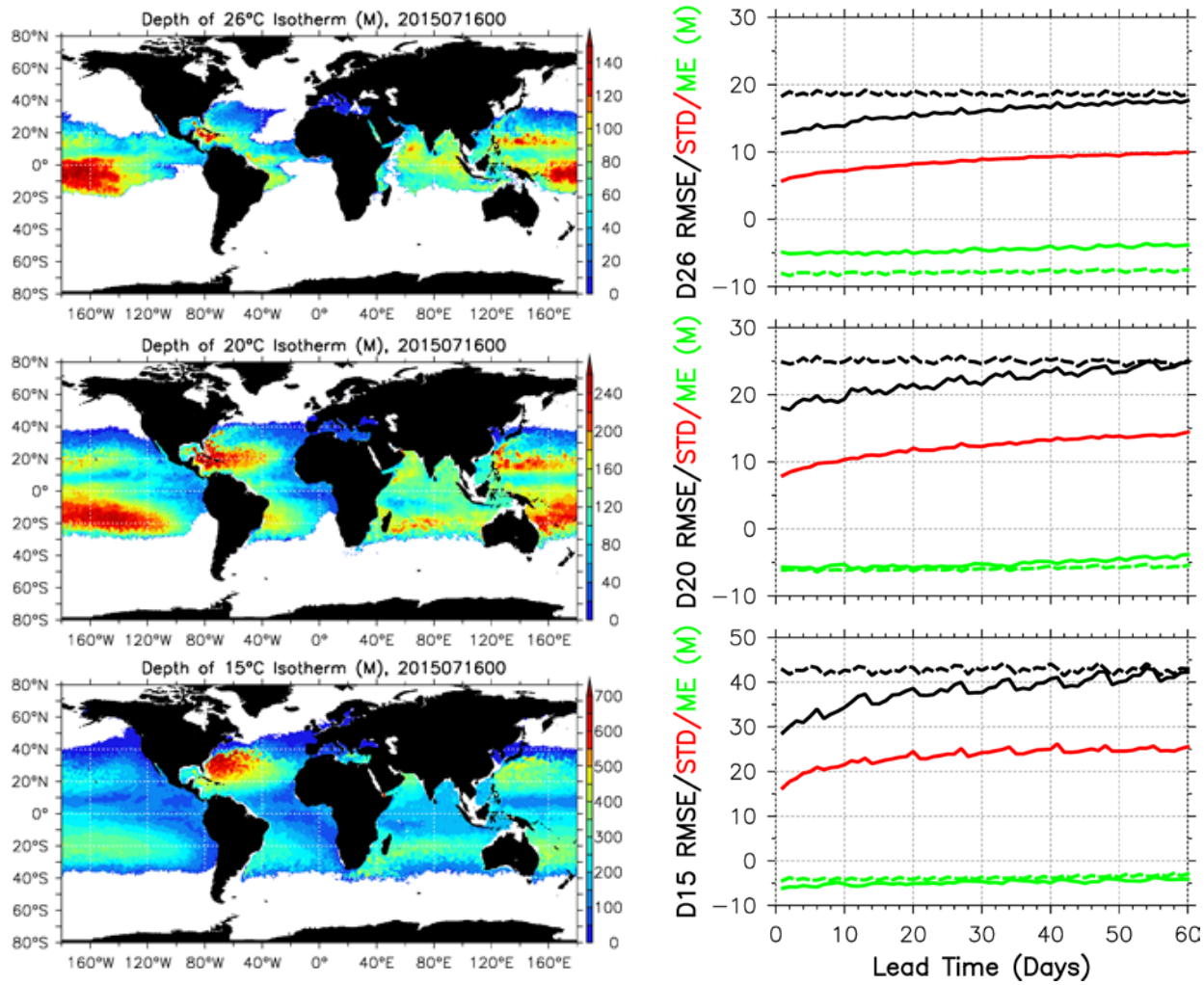

Figure 31: (left) Maps of the typical depths of the upper ocean shallow 26°C isotherm (upper), upper thermocline 20°C isotherm (middle) and lower thermocline 15°C isotherm (bottom) (this example is from uncoupled HYCOM GOFS 3.1 during July 16, 2015). (right) Time series of the RMSE, BIAS (ME), and ensemble standard deviation for the 26°C (top), 20°C (middle), and 15°C (bottom) isotherm depths measured using profile temperature measurements. Dashed lines indicate the RSME and BIAS for the GDEM v4 climatology.

The RMSE of the ensemble mean isotherm depth forecast for the 26°C, 20°C, and 15°C isotherms compared to the GDEM value indicates that the ensemble mean has skill out to 60 day for the 26°C isotherm depth, and to 55 day to 60 day for the 20C and 15C depths (Figure 31, right). The ensemble mean BIAS is about 5 m shallow at all three temperature values, while the GDEM climate BIAS is worse at the 26C isotherm, comparable at 20°C, and somewhat lower at 15°C. The time average spread to RMSE ratio for D26, D20 and D15 are 0.54, 0.55, 0.62 respectively, increasing to 0.76, 0.72 and 0.71 after subtracting BIAS from RMSE, (i.e. RMSE – |ME|). Thus the Navy-ESPC accounts for 70-75% of uncertainty of these selected isotherms depth.

Ensemble Temperature Spread-Skill

The ocean ensemble forecast spread skill for temperature is evaluated using the same matchup dataset. Matchups are binned by the absolute error of the ensemble mean forecast, and the bin medians of error and ensemble standard deviation are calculated. The bin medians (Figure 32) show a positive relationship between the forecast error and the ensemble spread, indicating that the ensemble spread forecast has some skill as a predictor of the forecast error. The slope of the best linear fit of the bin means is an indication of the under- or over- spread of the ensemble, and the results show a significant under-spread in the range 3.0 – 5.3 (these can be compared with the values for the atmosphere in Table 10).

The ensemble spread is a better predictor of the ensemble mean forecast error than the GDEM climatological variance for small errors, but because of the under-spread of the ensemble, the GDEM climatological variance has greater skill as a predictor at larger forecast errors. The distribution of observation errors is shown in the inset figures in Figure 32, and show that the ensemble spread has higher skill for 40% - 50% of observations (the first 2-3 bins). Using the overlap of the quartile ranges as a measure of the ability to distinguish small and large forecast errors, the ensemble spread and the GDEM variance have comparable skill and can likely only distinguish the smallest and largest forecast errors (first 3-4 bins compared to the last 1-2 bins).

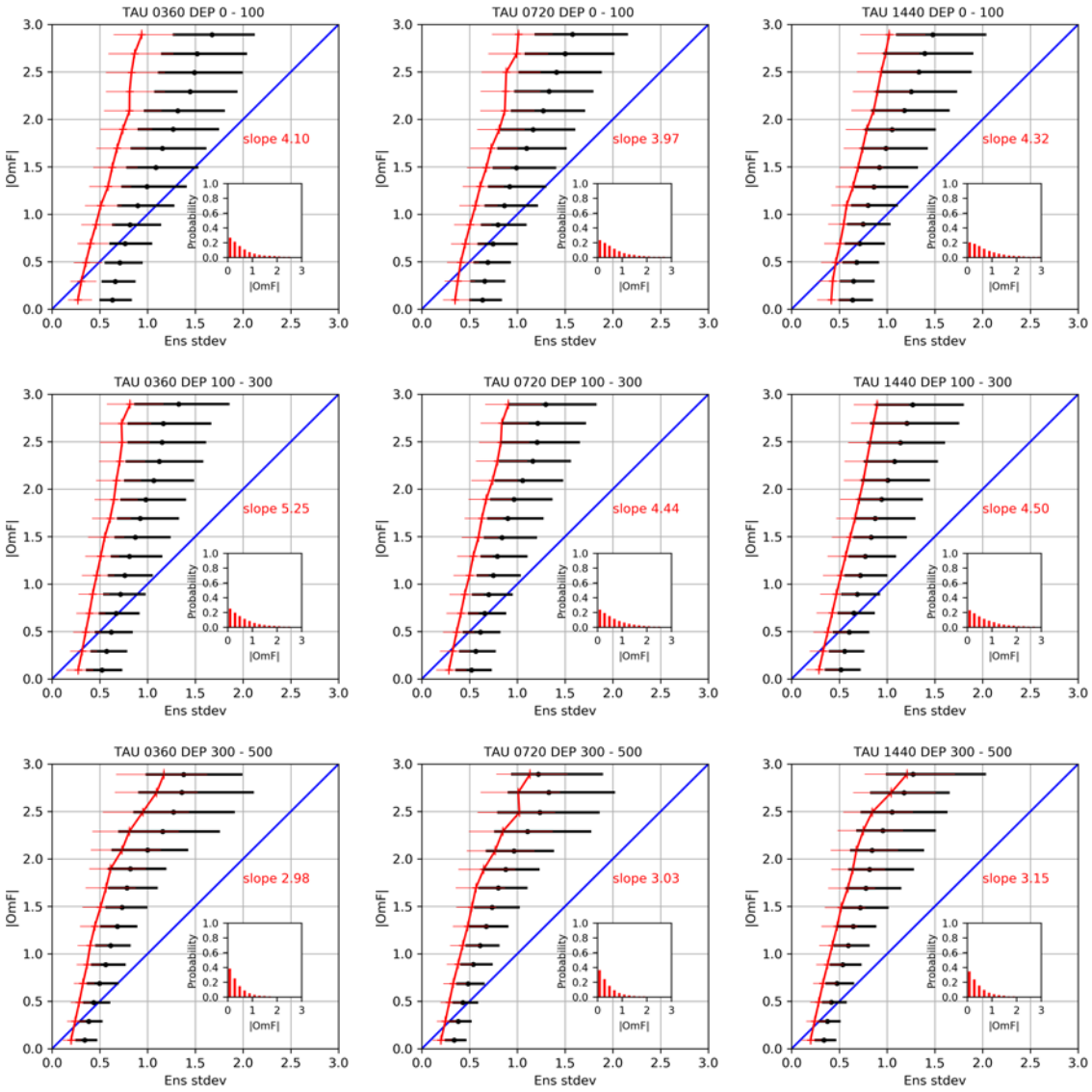

Figure 32: The relationship between bin-median absolute forecast temperature error (|OmF|) and the ensemble forecast spread (ensemble standard deviation) from profile matchups. Matchups are binned by the absolute error and by depth range. The slope of the linear fit to the bin medians is indicated, for forecast times 360 h (15 d), 720 h (30 d), and 1440 h (60 d) (left to right), for depths 0 – 100 m (top row), 100 – 300 m (middle), and 300 – 500 m (bottom). The 1st and 3rd quartile ranges are show by the bars. The ESPC ensemble results are shown in red, and the GDEM climatological variance prediction of the ensemble forecast error in black. Inset shows the frequency of the ensemble mean forecast error by bin (most observations are in the smallest error bins).

Ensemble SST Forecast CRPS and BSS

The Navy-ESPC ocean ensemble performance has been further evaluated in terms of ensemble metrics such as CRPS and BSS. The CRPS measures the squared difference between the forecast and observed cumulative distribution functions (CDFs). The CRPS compares the full distribution with the observation and provides a diagnostic of the global skill. The RMSE, by comparison, measures the squared distance between the ensemble mean and the observations. The CRPS has a negative orientation, and it rewards

concentration of probability around the step function located at the observed value (Wilks 1995). A perfect CRPS score is zero; a higher value of the CRPS indicates a lower skill of the ensemble. One of its advantages is that it has the same units as the predicted variable (so is comparable to the MAE) and does not depend on predefined thresholds. The Brier Score (BS) is the mean squared difference between the forecast probabilities and observed binary outcomes (Brier 1950; Jolliffe and Stephenson 2003) and is analogs to the CRPS except that the probabilities are calculated based on a threshold that defines an event. In other words, CRPS corresponds to the integral of the Brier Score over all thresholds. The Brier Skill Score (BSS) measures the relative skill of forecasts compared to a standard or climatological or persistence forecast (Wilks 2005). The zero value of BSS indicates that the probabilistic forecast skill is equal to the skill of the forecast based on climatology and negative BSS suggests forecast is less accurate than climatology.

The forecast performance of the ensemble prediction of ship surface temperatures is investigated with CRPS (Figure 33). The temporal variation of CRPS is very similar to RMSE. The CRPS degrades with forecast lead time, increasing from 0.25°C to 0.5°C, indicating loss of skill. The equivalent CRPS of a deterministic model is the mean absolute error (MAE). The MAE of ensemble mean (red line) degrades much faster (0.32° to 0.7°C) than the CRPS, indicating that the ensemble has better forecast skill. The right panel shows BSS as a function of lead time. The binary event is defined by the temperature observations that fall in the third quartile (Q_3). The forecast skill is assessed relative to the local climatology. Navy-ESPC ensemble is very skillful at all lead times as indicated by BSS > 0.85 (with 1 being the perfect score). There is a small degradation in the skill with lead time.

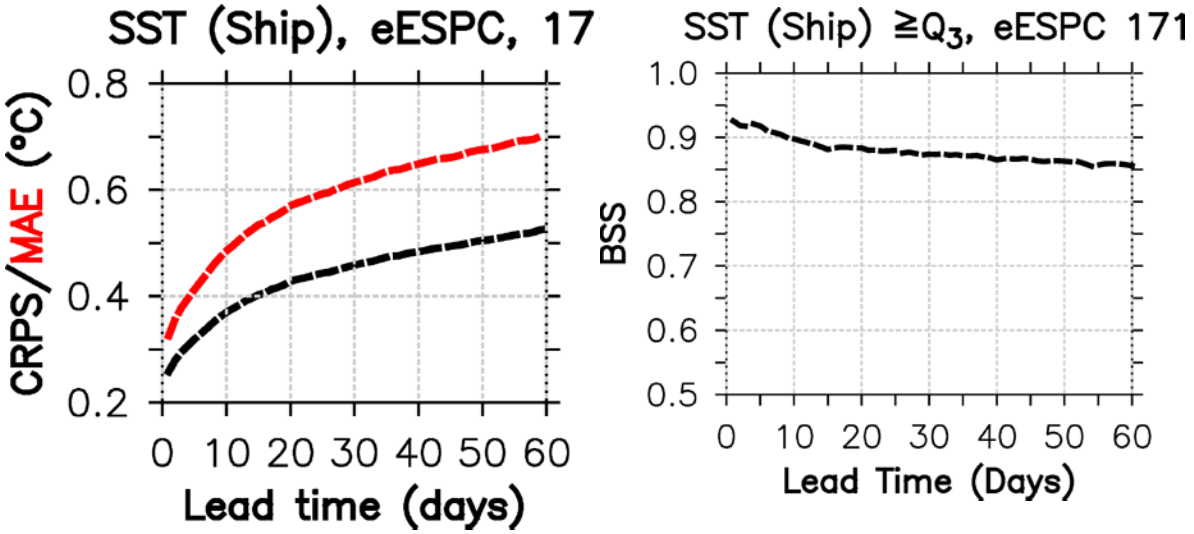

Figure 33: (left) Continuous Ranked Probability Score (CRPS, black), mean absolute error (MAE, red) of the ensemble mean and (right) Brier Skill Score (BSS) for SST greater than the upper quartile as a function of lead time.

Ensemble Sea Surface Height Forecast

The sea surface height (SSH) from the Navy ESPC ensemble is validated here against the ensemble mean of the verifying SSH analysis, as the best available verifying product. An annual mean SSH climatology is generated using the SSH analysis over the period February 1, 2017 to January 31, 2018 in lieu of an SSH climatology (a comparable product does not exist). The analysis is divided into eight different regions as

shown in Figure 34; Globe (GLB, 50°S – 50°N), Gulf Stream (GST), Gulf of Mexico/Intra-America Sea (GOM), Greenland/Iceland/Norwegian Seas (GIN), Kuroshio Extension (KEX), Western Pacific (WPA), Southern California (SCA) and Northern Arabian Sea (NAS). The RMSE of the ensemble mean (black), control member (red), and climatology (green) as a function of forecast length for different regions are shown in Figure 35 along with the anomaly correlations of ensemble mean (black-dash) and control member (red-dash). Since our climatology is the annual mean, anomaly correlation is the measure of SSH anomaly indicating mesoscale eddies. For all regions, the ensemble mean has consistently lower RMSE than control member at all lead times. The separation between these curves widens as the forecast lead time increases, indicating slow error growth in the ensemble mean. Furthermore, the ensemble mean performs better than or equal to climatology over the entire forecast length. The global RMSE of the ensemble mean performs better than climatology out to about 35 days and follows climatology thereafter. For both the ensemble mean and control member, the rate of error growth is shown to be much faster over the first week than in the second week, with a similar trend in the GIN region. The worst forecast skill in the GIN can be attributed to the lack of assimilation of synthetic profiles, as the vertical structure of temperature remains nearly isothermal throughout a year. Since model SSH includes barometric effect, it is likely that high latitude atmosphere pressure systems may have an impact on the SSH. The RMSE of individual members in this region also shows larger spread than any other regions. The RMSE of the control member exceeds climatology in about 10 days, with the ensemble mean approaching climatology. The anomaly correlation drops below 0.6 in less than 10 days. In the GST region, the RMSE of the ensemble mean is significantly smaller than the control member in day 1, with a similar pattern in the KEX region but a small difference. The ensemble mean performs better than climatology out to 30 days, while the control member crosses climatology in 5 days. A similar pattern is also evident in the anomaly correlation, although both drop below 0.6 in less than ~10 days. By comparison, KEX has better performance out to 60 days than the control member with 20 days. The evolution of the RMSE and anomaly correlation in regions GOM, WPA, SCA and NAS is very similar, beating climatology out to 60 days. Among the regions, WPA, SCA and NAS have anomaly correlation above 0.6 over the entire forecast length. In these regions, the RMSE of the ensemble mean is closely following that of the control member, although the ensemble mean has slightly lower RMSE. Overall, the ensemble mean has lower RMSE and is therefore more skillful than control member across all regions and the entire forecast length.

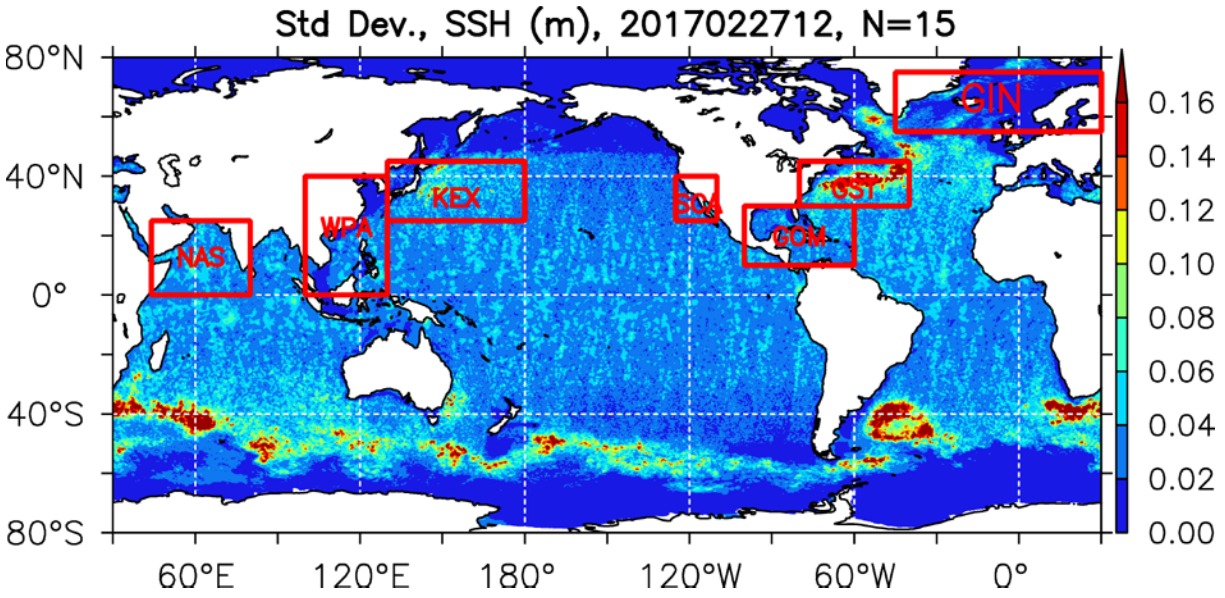

Figure 34: Regions used in the SSH analysis: Globe (GLB, 180°W-180°E, 50°S-50°N), West Pacific marginal seas (WPA, 100°-130°E, 0°-40°N); Kuroshio Extension (KE, 130°E-180°, 25°-45°N); Southern California (SCA, 125°-110°W, 25°-40°N); Gulf of Mexico/Intra-America Seas (GOM, 100°-60°W, 10°-30°N); Gulf Stream (GST, 80°-40°W, 30°-45°N); Greenland/Iceland/Norwegian (GIN, 45°W-30°E, 55°-75°N); Northern Arabian Sea (44°-80°E, 0°-25°N). Shaded regions indicate the standard deviation of the ensemble for February 27, 2017 at analysis time (based on 15 members) showing large spread in the western boundary current regions.

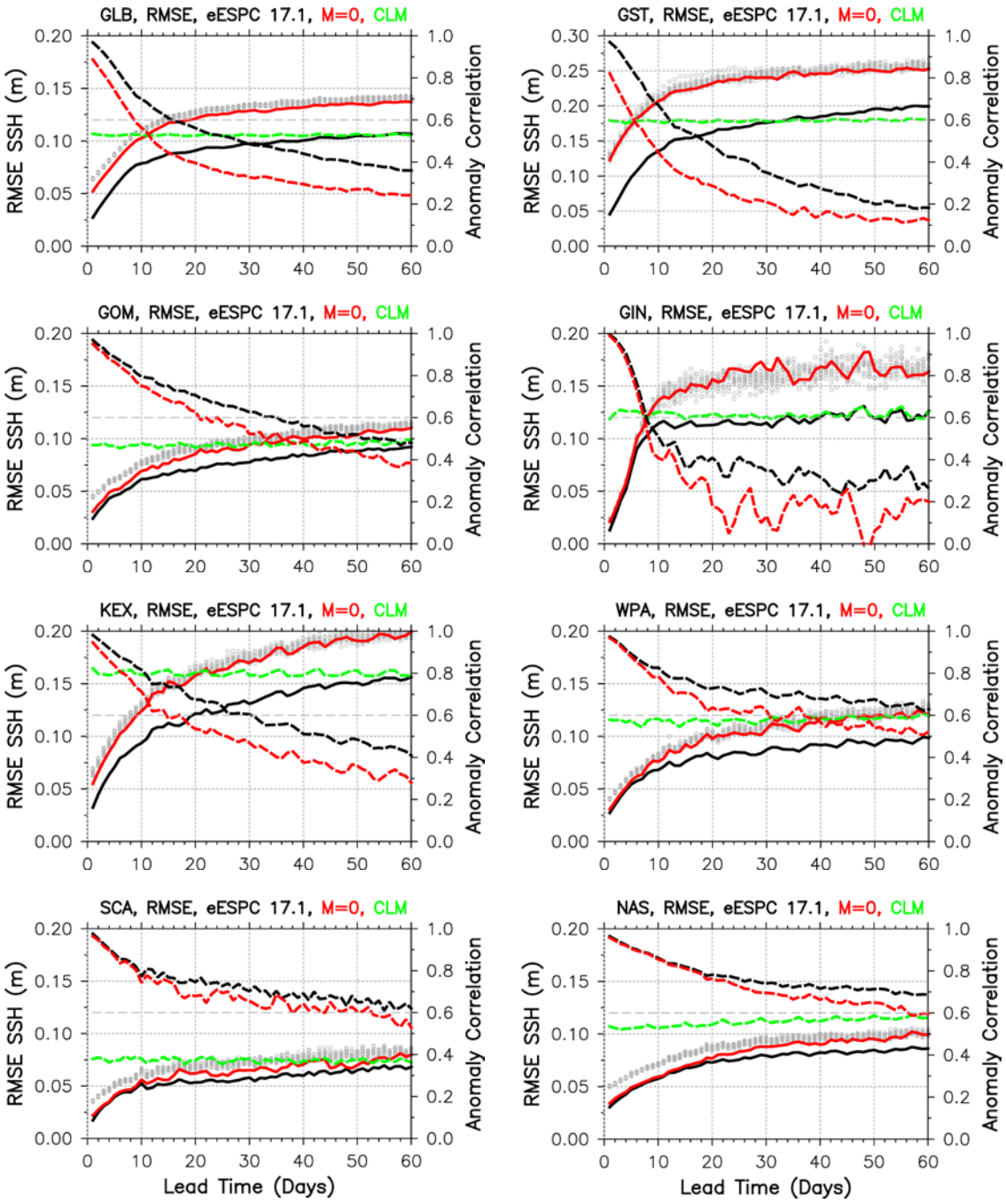

Figure 35: SSH RMSE of ensemble mean (black), control member (red) and climatology (green) using analysis ensemble mean SSH as the truth/observations for different ocean domains. Dashed lines correspond to the anomaly correlation for the ensemble mean (black) and control member (red). The climatology is defined as the annual mean of the analysis SSH over a year (February 1, 2017 to January 31, 2018) since there is no SSH climatology available. Grey symbols are RMSE of 15 ensemble members treated independently.

Figure 36 shows the ensemble spread and mean error or bias for different regions in addition to RMSE discussed above. The ensemble spread measured by the standard deviation around the ensemble mean is considered to be a good predictor of the RMSE of the ensemble mean. For the globe, a small difference between the RMSE and spread indicates good statistical consistency. The spread growth is generally aligned with the RMSE. The ensemble is mostly under-dispersive (spread < RMSE) except for first few days of the forecast, when it is slightly over-dispersive. This slight initial over-dispersion of the ensemble is evident across almost all regions. Except in the GLB and GIN regions where the bias is zero, the ensemble mean shows slightly increasing positive bias with forecast length. The spread in the GST region is over-dispersive during first 25 days and becomes under-dispersive thereafter. In fact, the spread is gradually decreasing as opposed to increasing RMSE. The growth of ensemble spread in most regions is slower than the error growth, indicating that the ensembles do not account for all the error growth. Among the regions, the NAS is severely under-dispersive and shows very little growth over the forecast length. However, the difference between the RMSE and spread becomes small when systematic bias is taken into account. For example, the largest bias occurs in the NAS at ~0.025 m, which is nearly equal to difference between RMSE and spread. When systematic bias is taken into account, the growth of spread is consistent with RMSE across all regions and forecast length and therefore suggests statistical consistency.

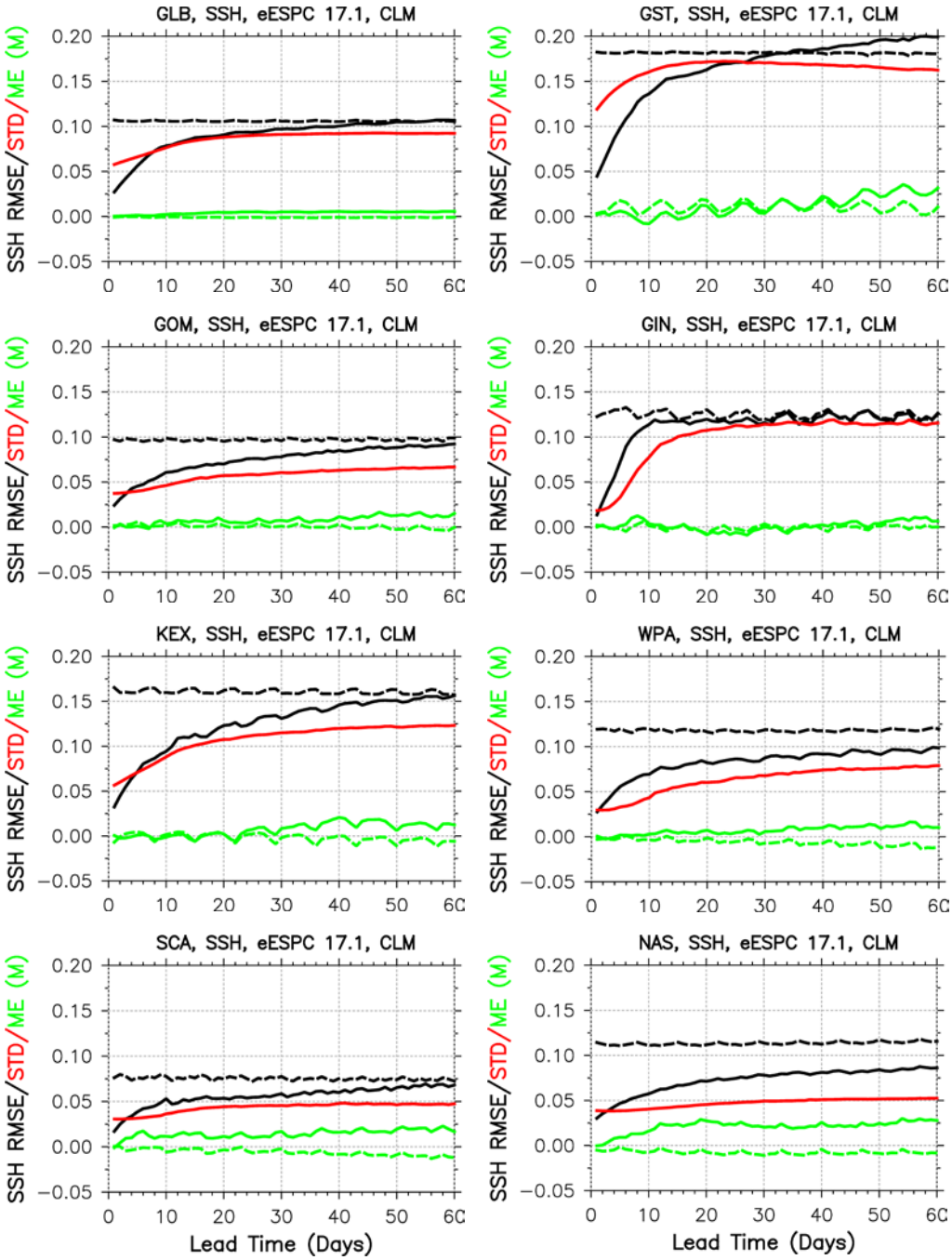

Figure 36: SSH RMSE (black) and ME (green) of the ensemble mean using the verifying ensemble mean analysis as truth/observations, and standard deviation (red) for different analysis regions. Dashed lines correspond to RMSE (black) and ME (green) using climatology. The climatology is defined as the annual mean of the analysis SSH over a year (February 1, 2017 to January 31, 2018) since there is no SSH climatology available.

Ensemble Acoustic Parameter Forecast

The profile matchup dataset used in the ensemble mean subsurface temperature assessment was processed here to evaluate the ensemble mean prediction of sound speed profile parameters sonic layer

depth (SLD), in-layer gradient (ILG), and below-layer gradient (BLG). The matchups for 50°S to 50°N were processed on standard levels (both observed and model profiles were interpolated to those levels) before the sound speed profile was calculated and the acoustic parameters assessed. Outliers identified as matchups with unreliable ensemble variances were removed from the comparisons (SLD ensemble standard deviation > 150 m, ILG standard deviation > 0.3 m/s/100ft; BLG standard deviation > 2.5 m/s/100ft). The observations were also matched with the Acoustic Parameter Climatology (APC; Helber et al. (2015)), an observation-based climatology of monthly mean and variance of SLD and BLG.

Figure 37 and Figure 38 show the ME and RMSE of SLD forecasts as a function of the forecast lead time for observed SLD shallower than 50 m and observed SLD between 50 m and 250 m, respectively, for 50°S - 50°N. For both shallow and deep SLDs, the RMSE is lower than the APC climatology at all forecast lead times. For shallow SLDs, the RMSE is a minimum at the 24 h forecast and increases somewhat at longer lead times, while the ME grows consistently from -0.7 to 9.7 m at 60 d, all lower than the climatological bias of about 12 m. For the deeper SLDs, the ensemble mean forecast RMSE is lower than the climatology at all lead times, but shows a noticeable negative bias at the initial time that decreases with forecast time and is only smaller than the climatological bias after about 45 d.

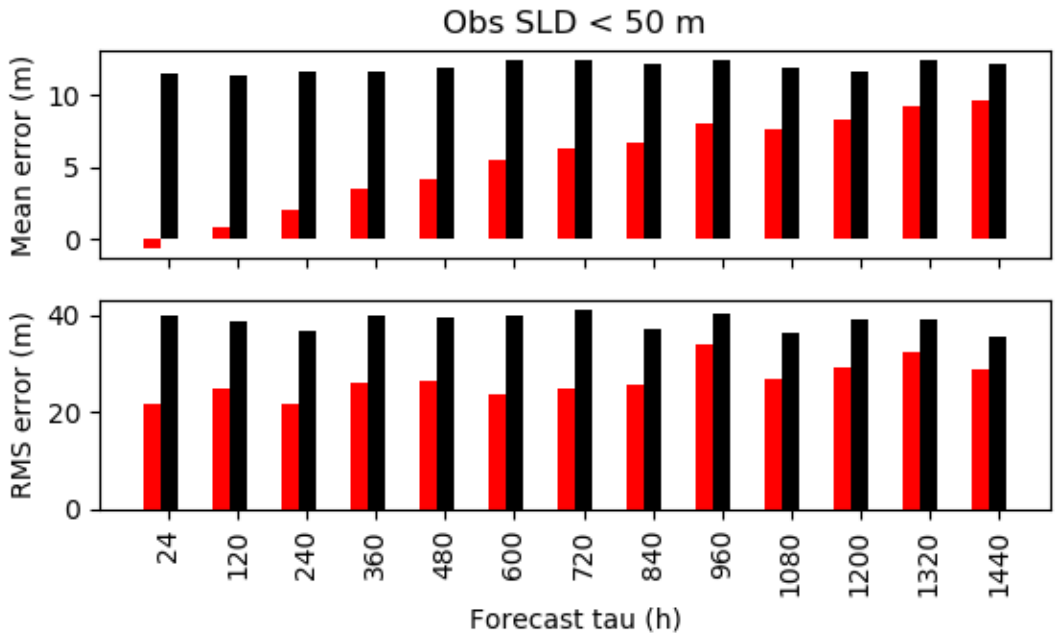

Figure 37: ME (top) and RMSE (bottom) of the ensemble mean forecast (red) and the APC climatology (black) for observed SLD less than 50 m as a function of forecast lead time.

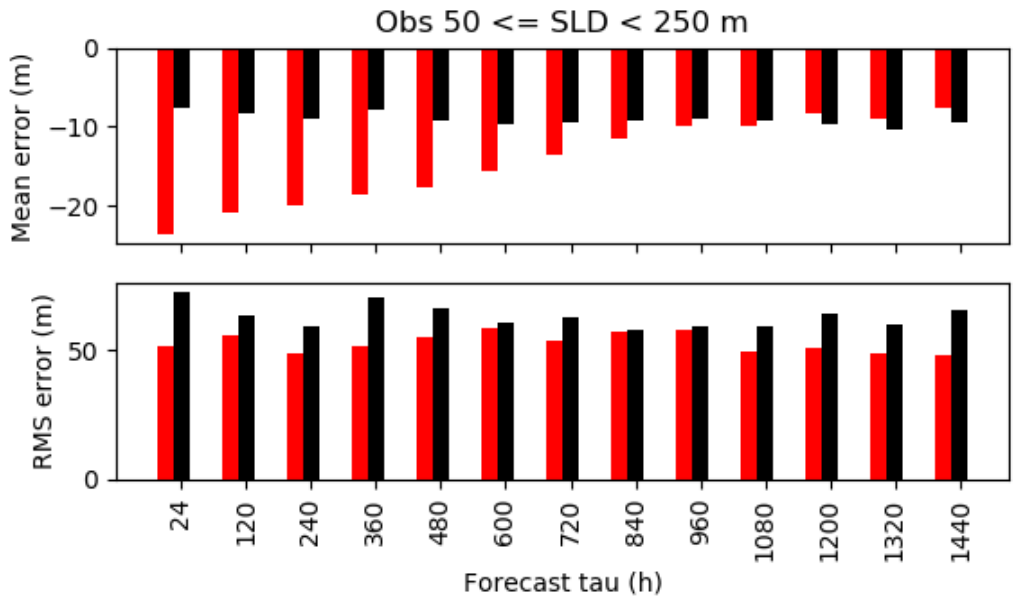

Figure 38: As in Figure 37 for SLD from 50 m to 250 m.

Figure 39 and Figure 40 show the ME and RMSE of ILG forecasts as a function of the forecast lead time for observed SLD shallower than 50 m and observed SLD between 50 m and 250 m, respectively, for 50°S - 50°N. For both shallow and deep SLDs, the ILG ensemble mean forecast RMSE is lower than the APC climatology at all forecast lead times, but the bias is significantly negative at all lead times, compared to the climatology small positive bias.

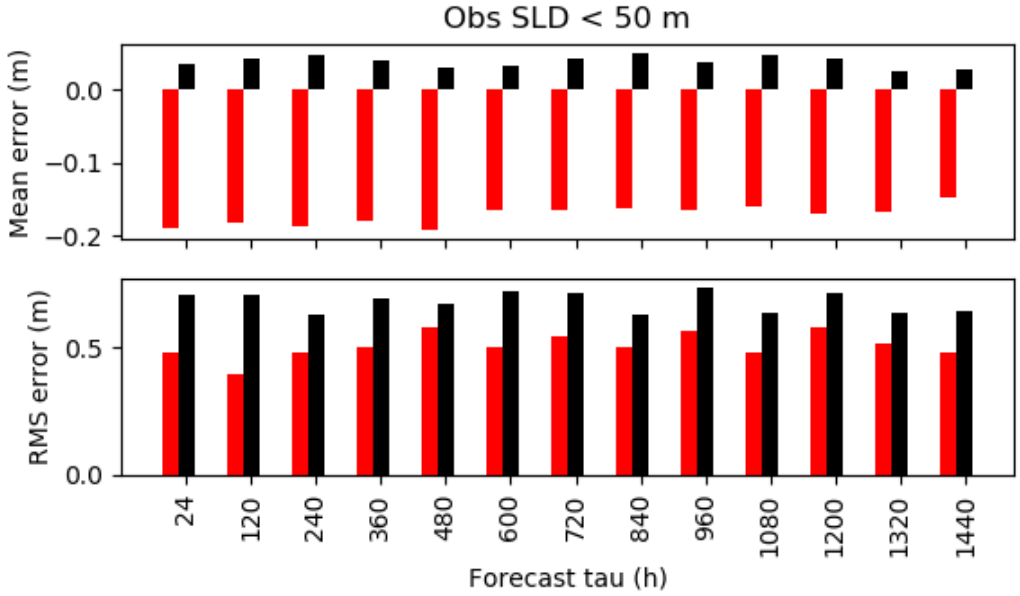

Figure 39: ME (top) and RMSE (bottom) of the ensemble mean forecast ILG (red) and the APC climatology ILG (black) for observed SLD less than 50 m as a function of forecast lead time.

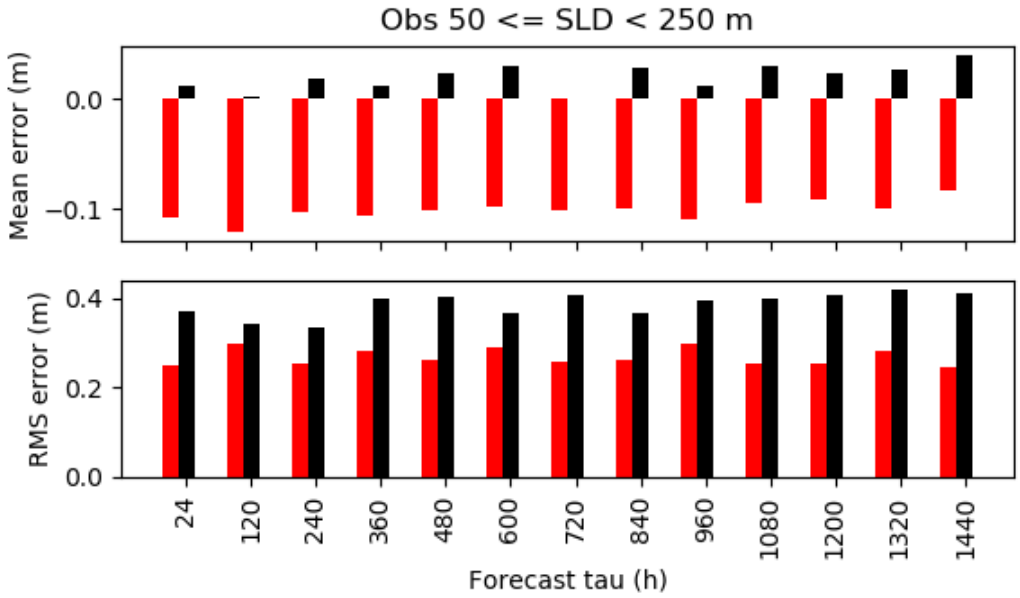

Figure 40: As in Figure 39, ILG ME and RSME for observed SLD from 50 m to 250 m.

The ensemble mean forecast of the BLG (Figure 41) shows slightly lower RMSE compared to the APC climatology out to 55 d forecast time. The bias (ME) is more negative than the consistent climatology bias of approximately -0.1 m/s/100ft at the initial time, and increases with forecast lead to about +0.19 m/s/100ft at 35 d.

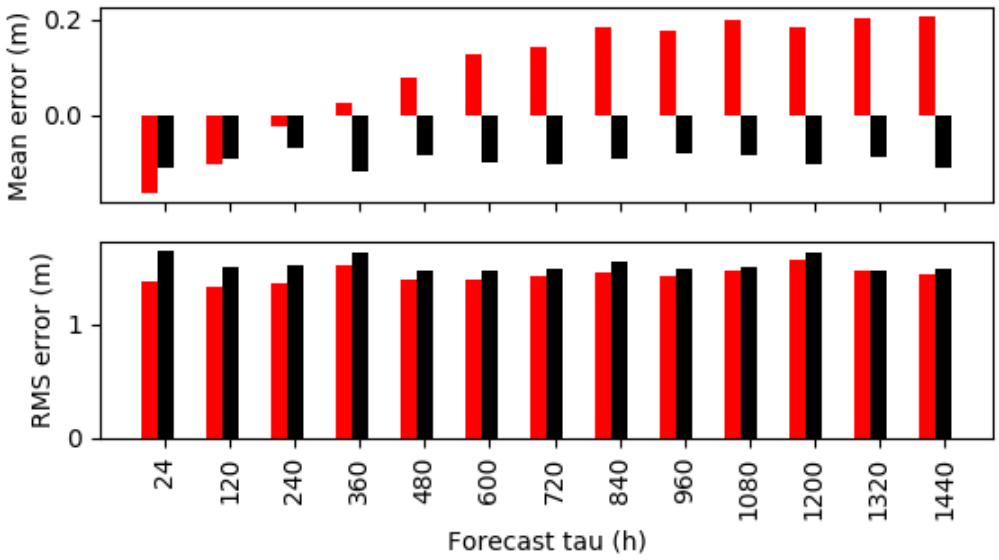

Figure 41: Below-layer gradient (BLG) forecast ME (top) and RMSE (bottom) of the ensemble mean forecast (red) and the APC climatology (black) as a function of forecast lead time.

Ensemble Acoustic Parameter Forecast

The ocean ensemble forecast spread skill for SLD is evaluated using the same matchup dataset. Matchups are binned by the absolute error of the ensemble mean forecast, and the bin medians and quartile values of error and ensemble variance are calculated. The bin medians (Figure 42) show a positive relationship between the forecast error and the ensemble spread, indicating that the ensemble spread forecast has some skill as a predictor of the ensemble mean forecast error, but the ensemble is overall under-spread, with spreads smaller than the forecast errors, and the limited spread of the ensemble makes the APC climatology a more skillful predictor of the forecast error. The slope of the best linear fit of the bin centers is an indication of the under- or over- spread of the ensemble, and the results show a significant under-spread in the range 2.2 - 3 for the shallow SLD and 2.3 - 6 for the deep SLD. The reference comparison shown here is for the APC climatology variance as a predictor of the error of the ESPC prediction. Using the overlap of the quartile ranges as a measure of the ability to distinguish small and large forecast errors, both the ensemble spread and the APC variance have limited skill and can likely only distinguish the smallest and largest errors.

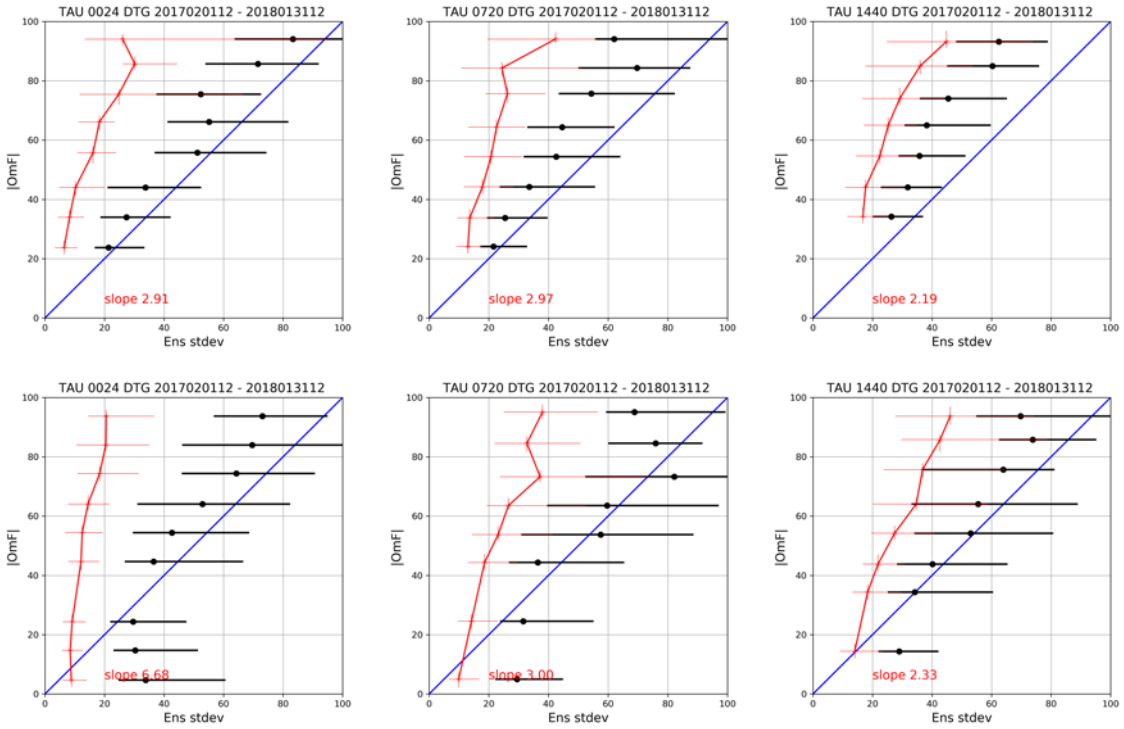

Figure 42: The relationship between bin-median absolute forecast SLD error (|OmF| of the ensemble mean) and the ensemble forecast spread (ensemble standard deviation), for (top) observed SLD < 50 m and (bottom) 50 m < SLD < 250 m, at 24 h (1 d), 720 h (30 d), and 1440 h (60 d) (left to right). Matchups are binned by the absolute error. The bars indicate the 25% and 75% quartile ranges in the bins. The EPSC ensemble results (red) are compared to the APC climatology variances as a predictor of the ESPC ensemble mean error (black dots and bars). The slope of the linear fit to the bin means is indicated, for (left to right) forecast times 24 h (1 d), 720 h (30 d), and 1440 h (60 d).

5.2.3. Sea Ice

Figure 43 shows the IIEE and BSS for both the Arctic and Antarctic as a function of forecast day averaged for the entire 2017 test period. For IIEE, Navy EPSC out-performs climatology out to 32 days in the Arctic and out to 40 days in the Antarctic. It outperforms a persistence forecast in the Antarctic for the entire time period. In the Arctic, Navy-ESPC outperforms persistence past 5 days, and is comparable to (slightly worse than) persistence during the first five days. In terms of BSS, the Navy-ESPC outperforms both persistence and climatology in both the Arctic and Antarctic for the entire 60 day forecast.

To illustrate how forecast skill varies as a function of season for the 2017 time period, Table 11 shows different measures of forecast skill for forecasts starting in Northern Hemisphere Spring (MAM), Summer (JJA), Fall (SON), and Winter (DJF), as well as for the entire year (All). The first two rows show the first forecast day in which the Navy-ESPC BSS becomes less than zero (that is, the Navy-ESPC forecast becomes less skillful than a climatological forecast). For the Arctic, Navy-ESPC retains skill above climatology for the entire 60-day forecast period. For the Antarctic, Navy-ESPC becomes less skillful than climatology after 46 days for both SON and DJF. Rows 3 and 4 of Table 11 indicate the day when the IIEE of the Navy-ESPC forecast becomes greater than the IEE from a forecast using climatology. There is considerable seasonal variation in this metric, ranging from a low of 11 days during JJA to a high of 46 days for MAM in the Arctic. In the Antarctic, the Navy-ESPC becomes less skillful than climatology at 28 days for JJA and SON, but retains skill above climatology for the entire 60-day forecast in MAM.

In addition to the IIEE and BSS, it is of interest to see if the forecasts in general are predicting too much or too little ice (of course both can occur in the same season in different regions). The last six rows of Table 11 show the over-forecast and under-forecast ice edge areas (in 106 km^2) as well as the area extent error (AEE, over-forecast minus under-forecast) for the Antarctic and Arctic, as a function of season, at a lead time of 30 days. The largest signal in the results is the tendency of the Navy-ESPC to over-predict sea-ice extent in the Antarctic in JJA, SON, and DJF. Indeed, during these seasons, the over-prediction of sea ice is 3 to 5 times larger than the under-prediction of sea ice. In the Arctic, the over-prediction and under-prediction areas are generally more balanced, although in JJA the under-prediction area is about 2.5 times as large as over-prediction area.

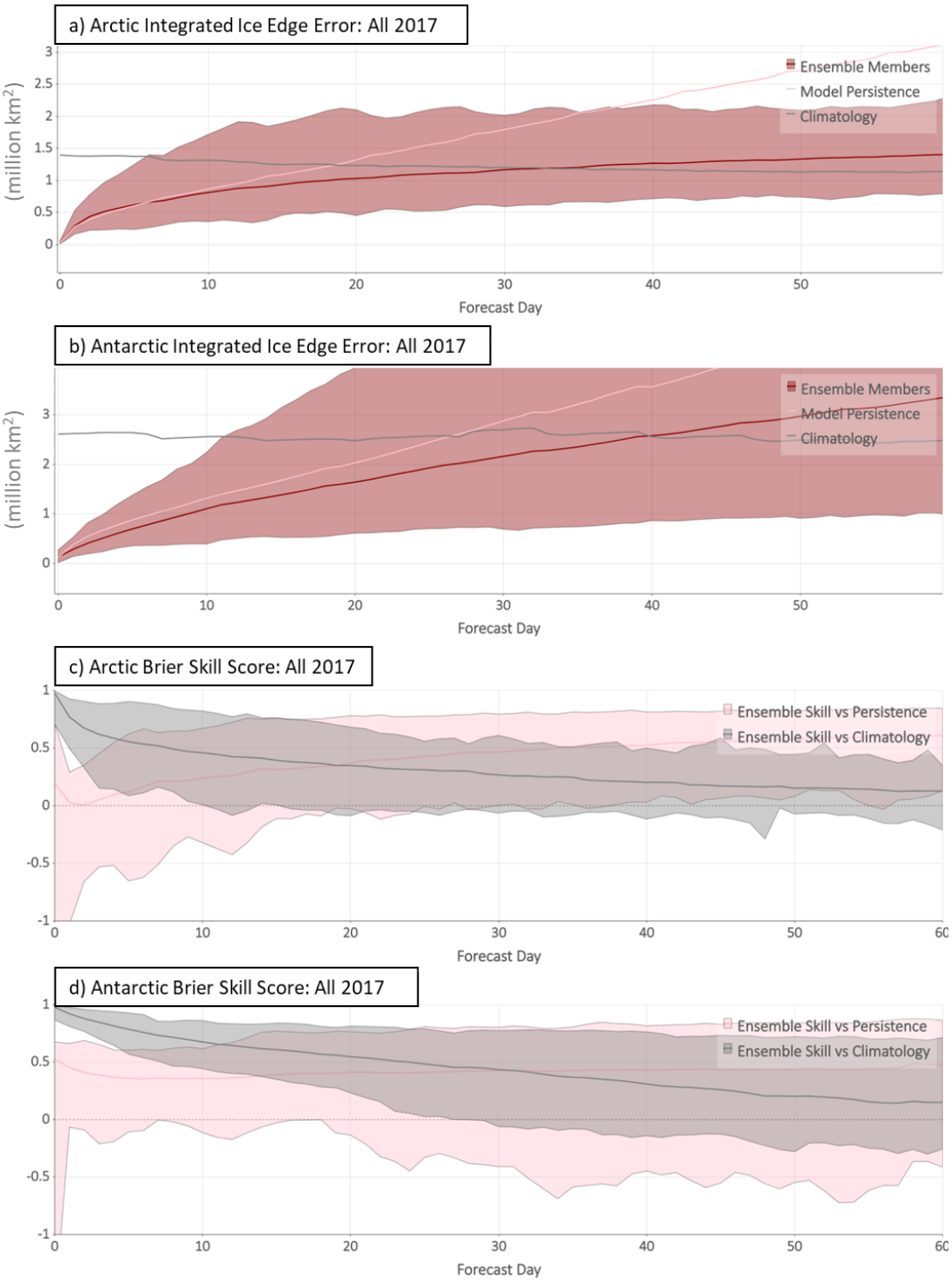

Figure 43: Panels a and b: Integrated Ice Edge Error (106 km2) for the Arctic and Antarctic, respectively, for Navy-ESPC ensemble mean (maroon), persisted analyses (pink) and 2007-2017 climatology (gray). Range of skill of individual ensemble members indicated by shaded region. Panels c and d: Brier Skill Score for ice concentration less than 15% for the Arctic and Antarctic, respectively, for Navy-ESPC ensemble as compared persistence (pink curve) and 2007-2017 climatology (gray curve). BSS range for individual ensemble members indicated by pink and gray areas.

Table 11: First two rows, day on which the Brier Skill Score (BSS) fall below zero for the Navy-ESPC ensemble as compared to climatology. Second two rows, day on which the Integrated Ice Edge Error (IEEE) of the Navy-ESPC ensemble becomes larger than using a climatological forecast. Last six rows, the total region of over-forecasted ice, total region of under-forecasted ice, and total ice areal extent error (AEEE) in 106 km2.

	MAM	JJA	SON	DJF	All
BSS < 0 Arctic	>60	>60	>60	>60	>60
BSS < 0 Antarctic	>60	>60	46	46	>60
IIEE > Climo Arctic	46	11	35	37	32
IIEE > Climo Antarctic	>60	28	28	39	39
Arctic Over Forecast at 30 d	0.655	0.433	0.383	0.65	0.508
Arctic Under Forecast at 30 d	0.42	1.106	0.553	0.35	0.661
Arctic AEE at 30 d	0.235	-0.673	-0.17	0.3	-0.153
Antarctic Over Forecast at 30 d	0.42	1.541	2.094	3.092	1.555
Antarctic Under Forecast at 30 d	0.809	0.515	0.432	0.804	0.61
Antarctic AEE at 30 d	-0.389	1.026	1.662	2.288	0.94

6. Summary

The Navy-ESPC system represents new capabilities for the Navy in predicting the atmosphere, ocean, and sea ice with a coupled system out to lead times past the typical operational periods of 7 days. This represents the first time atmospheric forecasts past 16 days and ocean/sea ice forecast past 7 days will be available. In addition, this is the first global ensemble system for the Navy that provides ocean forecasts. In this report, we analyzed one full year of the 16 member Navy-ESPC ensemble. As this is a brand new system, there is no apples-to-apples comparison to a current operational system (in contrast to the typical VTR documenting existing system upgrades). Our main baseline for this report is therefore climatological forecasts. In addition, we compared the Navy-ESPC ensemble to the current NAVGEM-ET ensemble and forecasts from other centers on the S2S data base.

For operations, we recommend a 16 member ensemble out to at least 45 day forecast. A 60 day forecast would be beneficial for some metrics if computational time is available. Below is summary of lead times that support a 45 to 60 day ensemble forecast:

- SST diagnostics compared to ship observations have lower RMSE values than climatology out to 60 days, while analyzing one member only has RMSE values lower than climatology out to 20 days (Figure 29).
- Global ocean temperatures averaged between 8 and 500 meters in depth have RMSE values lower than climatology out to about 30 days in the forecast (Figure 23).
- Sonic Layer Depth (SLD), In-Layer Gradient (ILG), and Below-Layer Gradient (BLG) have RMSE lower than the APC acoustic parameter climatology to at least 55 days (Figure 37 - Figure 41).
- Pan Arctic and Antarctic sea ice extent diagnostics show skill over climatology from 11 to 60 days depending on season and metric (Table 11, Figure 43). The National Ice Center (NIC) is currently utilizing 45 day forecasts from the Navy-ESPC in near-real-time testing.

Though the atmospheric metrics did not show skill over climatology up to 45 days, skill was very comparable to other leading centers (Figure 10, Figure 11, Figure 13, Figure 14, Figure 15), and the below is a summary of the forecast lead times.

- The MJO diagnostics are comparable to other leading centers (and better than NOAA CFSv2), and the anomaly correlation drops below 0.6 at around 20-30 days into the forecast (Figure 11).
- Forecasts for the AO, AAO, NAO, and PNA are also comparable to other leading centers and have ensemble mean anomaly correlations greater than 0.6 in the 8 to 12 day time frame depending on diagnostics (Figure 15).

It is very encouraging to see that the Navy-ESPC provides value above climatology on the multi-week time scale for sea-ice extent. There is also anecdotal evidence supporting the utility of the Navy-ESPC sea-ice forecasts on multi-week timescales. As an extension of the Navy's participation in the NOAA Subseasonal Experiment (SubX), NRL has been producing Navy-ESPC 45-day forecasts 4 times per week in real time and making these forecasts available to the National Ice Center (NIC). Over the last two years, NIC personnel have used the Navy-ESPC forecasts for real-time planning of resupply missions and DoD exercises in both the Arctic and Antarctic and have provided feedback that they find the forecasts useful for these purposes. While current results are promising, we note that the comparison between GOFS 3.1 (with CICE v4) and GOFS 3.5 (with CICE v5) show enhanced performance with GOFS 3.5, especially during the winter freeze-up period. We anticipate that upgrading to CICE v5 will also substantially improve Navy-ESPC sea ice forecasts on both short and extended lead times.

A limitation of the Navy-ESPC system is that it is under-dispersive in the initial conditions and forecast compared to the NAVGEM-ET ensemble, as shown by many spread metrics (Figure 16, Figure 17) and the comparison of AO, NAO, AAO, and PNA using ensemble mean versus deterministic forecasts (Figure 10, Figure 14). This limitation in the Navy-ESPC model is currently being address for future transitions and is described in the next section.

We are pleased with the performance of the Navy-ESPC system and recommend operational implementation. At the same time, we look forward to increased performance with the next Navy-ESPC upgrade which will include many performance upgrades as described below.

7. Future Work

Ensemble Spread: Increase Spread in Initial Conditions and Forecast

This report identified that deficiencies in ensemble design are top priorities for improvement as we move forward toward the next implementation of the Navy-ESPC coupled system, and there is current work to address this issue.

The methods of Analysis Correction Additive Inflation (ACAI) and Relaxation to Prior Perturbation (RTPP) are currently being tested in the stand-alone NAVGEM system. ACAI addresses the need for improved stochastic parameterization by introducing additional forecast tendency terms based on short-term forecast errors and biases. This tendency term combines a seasonal bias correction term and a stochastic representation of the model error by drawing from an archive of past analysis corrections. RTTP addresses a problem of overconfidence of atmospheric data assimilation in the mid-latitudes. Specifically, a pure perturbed observation ensemble removes almost all of the uncertainty in the location of the mid-latitude baroclinic wave. RTTP addresses this problem by generating the initial perturbation as a weighted sum of the forecast and analysis perturbation.

Examples of improvements to the spread skill metric in the atmosphere-only ensemble prototype system are shown in Figure 44. As shown in the VTR, the baseline perturbed observations ensemble generation is significantly under spread compared to the target of ensemble variance/MSD. Adding RTTP, ACAI, and the stochastic kinetic energy backscatter (SKEB) scheme to the baseline perturbed observations NAVGEM system significantly improves the spread at the initial time and through the forecast (Figure 15) though more testing is needed to calibrate the parameters of the system to achieve the desired spread.

In addition to the work on ACAI and RTTP in NAVGEM, there are other methods that will be tested to improve ensemble dispersion. Much of this work will be accomplished through the NRL 6.2 project "Accounting for Model Uncertainty in the Coupled System". The objective of this project is to develop, implement and evaluate a global coupled forecast system (Navy-ESPC) that will provide reliable extended-range probabilistic forecasts by accounting for model uncertainty in the individual components of the system. Under this project we will apply stochastic perturbations to each model component of the global system, as well as add stochastic perturbations directly to the coupled flux interfaces to account for the effects of disparate space and time scales between the different components of the coupled system. We will also test variations to model parameters such as water clarity estimates and ice strength. One of the first tests will be to turn on the stochastic kinetic energy backscatter scheme (which is currently in the stand-alone NAVGEM system) in the Navy-ESPC ensemble to examine its impact on ensemble performance.

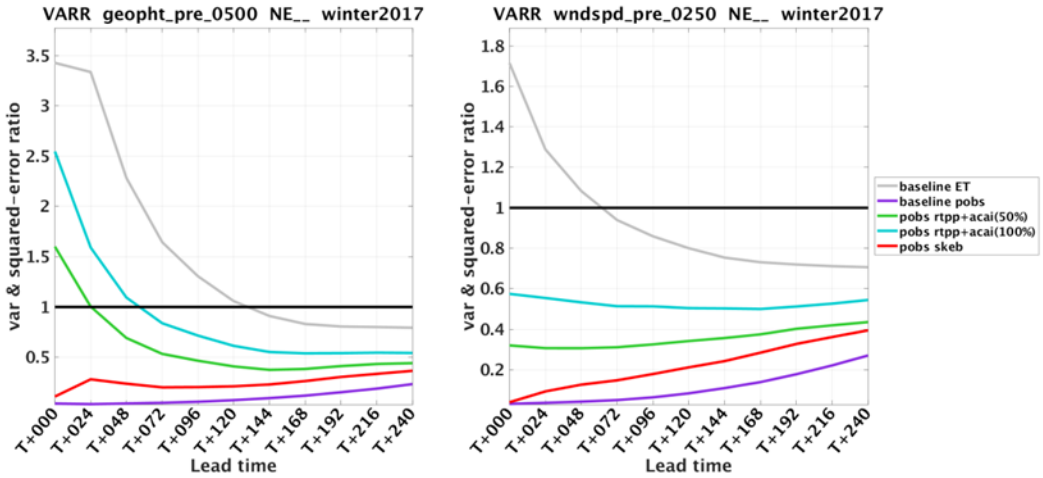

Figure 44: NAVGEM ensemble testing using ACAI and RTPP. NAVGEM stand-alone runs of (red) perturbed observations plus SKEB, (blue) perturbed observations plus RTPP and ACAI using 100% of the stochastic perturbation, (green) perturbed observations plus RTPP and ACAI using 50% of the stochastic perturbation, (purple) perturbed observations, and (grey) NAVGEM-ET ensemble. The ensembles consisted of 5 perturbed members plus a control member, and the time period of testing is Dec. 15 2016 – Jan. 14, 2017. The ensemble target line is defined at ensemble variance/ model-verify-bias. The VARR and squared error ration of (left) 500 hPa geopotential heights and (right) 250 hPa geopotential heights are displayed.

Products for Tropical Cyclones

The Navy-ESPC global coupled forecast model will provide the Navy, for the first time, with operational ensemble forecasts of atmosphere, ocean, and sea-ice conditions out to 45 days. This completely new capability for the Navy requires a paradigm shift in how forecast information is provided. Effective use of multi-week forecasts will require development of new types of forecast products as well as outreach to potential users to learn what information would be valuable and how this information may be effectively conveyed to decision-makers. To address these issues, we have created a "Production, Promotion, and Dissemination" plan for Navy-ESPC ensemble products. Our strategy is to start with outreach to potential users to advertise Navy-ESPC capabilities and to learn about decision-maker needs and dissemination pathways. Work under a 2019 B&P proposal and other outreach efforts will gather the information needed for the development of a longer-term plan and associated proposals, including hosting a workshop with Navy decision makers in fall 2019 or spring 2020. In addition, a new NRL 6.2 project on extended-range TC prediction will start FY2020. A major goal of this project is to blend statistical and dynamical forecasts to produce useful TC activity forecasts on a weekly basis out to 7 weeks. NRL is already providing 45-day Navy-ESPC forecasts four times per week to the National Ice Center (NIC), the Advanced Climate Analysis and Forecast (ACAF) system, and the NOAA SubX project.

Wave Watch 3: inclusion of wave model

The initial implementation of the Navy-ESPC does not include ocean surface waves. Wave Watch 3 (WW3) has recently been incorporated into the Navy-ESPC at the ensemble resolution (1/4°), and is showing comparable performance to the control WW3 forecasts, but this was not incorporated in time for the VTR runs. For FOC, we plan to incorporate WW3 into the Navy-ESPC system at 1/4° for the ensemble component and 1/8° for the deterministic component. This will be the first Navy global wave model with resolution higher than 1/4°, surface currents in forcing, and ensembles that take into account uncertainty associated with errors in ice edge position. The inclusion of WW3 into Navy-ESPC will allow, for the first time, prediction of anomalous ocean surface wave conditions out to 45 days. Feedback from WW3 to the other components of the system will enhance extended-range forecast skill in all Navy-ESPC component models. Accounting for ice-edge uncertainty will improve prediction of environmental conditions in the marginal ice zone.

CICE6: upgrade sea ice model

The current Navy-ESPC system includes CICE v4. Comparison with GOFS 3.5, which includes CICE v5.1.2, indicates that the more advanced version of CICE provides more accurate sea ice prediction, especially during winter freeze up. This is due to the inclusion of important processes for sea ice thermodynamics, such as melt ponds, snow cover, and landfast sea ice, which are currently missing in CICEv4. We plan to incorporate CICE v6 into the Navy-ESPC system for FOC for the deterministic and ensemble resolutions, as this will include the improvements of CICE v5.1.2 as well as other advantages, including the fact that the model will not have to be recompiled to be run at different resolutions.

HYCOM upgrades: Tides, layers, and physics

The current version of Navy-ESPC has 41 vertical hybrid layers in HYCOM and does not account for tides in the ensemble configuration. For FOC, we anticipate increased fidelity and performance through the addition of tides to the ensemble configuration and increased vertical levels (50-60). Higher order (more accurate) advection, proper treatment of rivers and evaporation and precipitation, more accurate

vertical mixing and flux calculations will also be incorporated. We expect these improvements to result in increased fidelity in ocean structure, particularly in regions of strong vertical gradients, which will lead to improved sound speed profiles and representation of the acoustical environment. The addition of tidal forcing provides a good representation of internal waves that is important to the submarine community.

NAVGEM upgrades: higher horizontal resolution and inclusion of high altitude physics

The current NAVGEM version in Navy-ESPC IOC has a top at 72 km, with rudimentary middle-atmospheric physics. For FOC, we plan to incorporate NAVGEM-HA (High-Altitude), which has a model top at 100 km and sophisticated middle-atmosphere physics. This capability will allow for operational forecasts (for the first time, for any lead time) of the middle atmosphere, including the mesosphere and lower thermosphere. This will provide accurate lower boundary conditions for the ionospheric forecasts, protect assets traveling in and through the middle atmosphere, and improve S2S tropospheric weather prediction as the stratosphere has been shown as a source of long-lead tropospheric predictability.

In addition, NAVGEM model physics will require additional updating for the coupled system, not only with respect to processes currently treated, but in order to provide additional capabilities, such as coupling with ocean waves and the representation of aerosol physics.

8. References

Bennett, A. F., 2002: *Inverse Modeling of the Ocean and Atmosphere.* Cambridge University Press.

Bleck, R., 2002: An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. *Ocean Model.*, **4,** 55-88.

Bloom, S. C., L. L. Takacs, A. M. d. Silva, and D. Ledvina, 1996: Data Assimilation Using Incremental Analysis Updates. *Monthly Weather Review*, **124,** 1256-1271.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78,** 1-3.

Carnes, M., R. W. Helber, C. N. Barron, and J. M. Dastugue, 2010: Validation Test Report for GDEM4NRL/MR/7330--10-9271.

Chassignet, E. P., L. T. Smith, G. R. Halliwell, and R. Bleck, 2003: North Atlantic Simulations with the Hybrid Coordinate Ocean Model (HYCOM): Impact of the vertical coordinate choice, reference pressure, and thermobaricity. *J Phys Oceanogr*, **33,** 2504-2526.

Cleary, R. J., 2012: Verification of cloud analyses used to support of overhead imagery collection (Master's Thesis). *Naval Postgraduate School*.

Clough, S. A., and Coauthors, 2005: Atmospheric radiative transfer modeling: a summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transf.*, **91,** 233-244.

Cummings, J. A., 2005: Operational multivariate ocean data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **131,** 3583-3604.

Cummings, J. A., and O. M. Smedstad, 2013: Variational Data Assimilation for the Global Ocean. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, S. K. Park, and L. Xu, Eds., Springer Berlin Heidelberg, 303-343.

Eckermann, S., 2009: Hybrid sigma-p Coordinate Choices for a Global Model. *Monthly Weather Review*, **137,** 224-245.

Eckermann, S. D., J. P. McCormack, J. Ma, T. F. Hogan, and K. A. Zawdie, 2014: Stratospheric Analysis and Forecast Errors Using Hybrid and Sigma Coordinates. *Monthly Weather Review*, **142,** 476-485.

Emanuel, K. A., 1991: A SCHEME FOR REPRESENTING CUMULUS CONVECTION IN LARGE-SCALE MODELS. *Journal of the Atmospheric Sciences*, **48,** 2313-2335.

Emanuel, K. A., and M. Zivkovic-Rothman, 1999: Development and evaluation of a convection scheme for use in climate models. *Journal of the Atmospheric Sciences*, **56,** 1766-1782.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air-sea fluxes: Updates and verification for the COARE algorithm. *Journal of Climate*, **16,** 571-591.

Goessling, H. F., S. Tietsche, J. J. Day, E. Hawkins, and T. Jung, 2016: Predictability of the Arctic sea ice edge. *Geophysical Research Letters*, **43,** 1642-1650.

Gregory, D., R. Kershaw, and P. M. Inness, 1997: Parametrization of momentum transport by convection. II: Tests in single-column and general circulation models. *Quarterly Journal of the Royal Meteorological Society*, **123,** 1153-1183.

Halliwell, G. R., 2004: Evaluation of vertical coordinate and vertical mixing algorithms in the HYbrid-Coordinate Ocean Model (HYCOM). *Ocean Model.*, **7,** 285-322.

Han, J., and H. L. Pan, 2011: Revision of Convection and Vertical Diffusion Schemes in the NCEP Global Forecast System. *Weather Forecast*, **26,** 520-533.

Hebert, D. A., and Coauthors, 2015: Short-term sea ice forecasting: An assessment of ice concentration and ice drift forecasts using the US Navy's Arctic Cap Nowcast/Forecast System. *J. Geophys. Res.-Oceans*, **120,** 8327-8345.

Helber, R. W., C. N. Barron, and J. M. Dastugue, 2015: The Acoustic Parameter Climatology. *Navy Journal of Underwater Acoustics*, **JUA_2015_001_L**.

Hogan, T., and Coauthors, 2014: The Navy Global Environmental Model. *Oceanography*, **27,** 116-125.

Hogan, T. F., 2007: Land surface modeling in the Navy Operational Global Atmospheric Prediction System. *AMS 22nd Conference on Weather Analysis and Forecasting/ 189th Conference on Numerical Weather Prediction*.

Hunke, E. C., and W. Lipscomb, 2008: CICE: The Los Alamos sea ice model, documentation and software user's manual, version 4.0. *Techincal Report*, **LA-CC-06-012, Los Alamos National Laboratory, Los Alamos, NM**.

Janiga, M. A., C. J. Schreck, J. A. Ridout, M. Flatau, N. P. Barton, E. J. Metzger, and C. A. Reynolds, 2018: Subseasonal Forecasts of Convectively Coupled Equatorial Waves and the MJO: Activity and Predictive Skill. *Monthly Weather Review*, **146,** 2337-2360.

Jiang, X., and Coauthors, 2015: Vertical structure and physical processes of the Madden-Julian oscillation: Exploring key model physics in climate simulations. *Journal of Geophysical Research-Atmospheres*, **120,** 4718-4748.

Kain, J. S., and J. M. Fritsch, 1990: A ONE-DIMENSIONAL ENTRAINING DETRAINING PLUME MODEL AND ITS APPLICATION IN CONVECTIVE PARAMETERIZATION. *Journal of the Atmospheric Sciences*, **47,** 2784-2802.

Kain, J. S., and J. M. Fritsch, 1993: Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme. *The Representation of Cumulus Convection in Numerical Models*, K. A. Emanuel, and D. J. Raymond, Eds., American Meteorological Society, 165-170.

Kara, A. B., H. E. Hurlburt, and A. J. Wallcraft, 2005: Stability-dependent exchange coefficients for air-sea fluxes. *J Atmos Ocean Tech*, **22,** 1080-1094.

Lim, Y., S. W. Son, and D. Kim, 2018: MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models. *Journal of Climate*, **31,** 4075-4094.

Loeb, N. G., and Coauthors, 2018: Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product. *Journal of Climate*, **31,** 895-918.

Louis, J. F., M. Tiedtke, and J. F. Geleyn, 1982: A short history of the operational PBL parameterization at ECMWF. *ECMWF Workshop on Planetary Boundary Parameterizations***,** 59-79.

McCormack, J. P., S. D. Eckermann, D. E. Siskind, and T. J. McGee, 2006: CHEM2D-OPP: A new linearized gas-phase ozone photochemistry parameterization for high-altitude NWP and climate models. *Atmos. Chem. Phys.*, **6,** 4943-4972.

McLay, J., and T. Whitcomb, 2017: NAVGEM Ensemble Prediction System (EPS): Upgrade to introduce SST variation. *Validation Test Report. Naval Research Laboratory, Monterey, CA.*

McLay, J., J. Shafer, and W. T., 2016a: NAVGEM Ensemble Prediction System (EPS): Upgrades to model, ensemble transform, and resolution. . *Validation Test Report. Naval Research Laboratory, Monterey, CA.*

McLay, J., T. Whitcomb, J. Shafer, and R. C., 2016b: NAVGEM 42d Intra-Seasonal Ensemble Forecast System (EFS). *Validation Test Report. Naval Research Laboratory, Monterey, CA.*

Meier, W. N., F. Fetterer, J. S. Stewart, and S. Helfrich, 2015: How do sea-ice concentrations from operational data compare with passive microwave estimates? Implications for improved model evaluations and forecasting. *Annals of Glaciology*, **56,** 332-340.

Metzger, E. J., and Coauthors, 2010: Simulated and observed circulation in the Indonesian Seas: 1/12 degrees global HYCOM and the INSTANT observations. *Dyn. Atmos. Oceans*, **50,** 275-300.

Metzger, E. J., and Coauthors, 2014: US Navy Operational Global Ocean and Arctic Ice Prediction Systems. *Oceanography*, **27,** 32-43.

Moorthi, S., H.-L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. *NWS Technical Procedures Bulletin*, **484,** 14.

Peng, M. S., J. A. Ridout, and T. F. Hogan, 2004: Recent modifications of the emanuel convective scheme in the Navy Operational Global Atmospheric Prediction System. *Monthly Weather Review*, **132,** 1254-1268.

Ridout, J. A., and C. A. Reynolds, 1998: Western Pacific warm pool region sensitivity to convective triggering by boundary layer thermals in the NOGAPS atmospheric GCM. *Journal of Climate*, **11,** 1553-1573.

Ridout, J. A., Y. Jin, and C. S. Liou, 2005: A cloud-base quasi-balance constraint for parameterized convection: Application to the Kain-Fritsch cumulus scheme. *Monthly Weather Review*, **133,** 3315-3334.

Ritchie, H., 1987: SEMI-LAGRANGIAN ADVECTION ON A GAUSSIAN GRID. *Monthly Weather Review*, **115,** 608-619.

——, 1988: APPLICATION OF THE SEMI-LAGRANGIAN METHOD TO A SPECTRAL MODEL OF THE SHALLOW-WATER EQUATIONS. *Monthly Weather Review*, **116,** 1587-1598.

——, 1991: APPLICATION OF THE SEMI-LAGRANGIAN METHOD TO A MULTILEVEL SPECTRAL PRIMITIVE-EQUATIONS MODEL. *Quarterly Journal of the Royal Meteorological Society*, **117,** 91-106.

Ritchie, H., C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud, 1995: IMPLEMENTATION OF THE SEMI-LAGRANGIAN METHOD IN A HIGH-RESOLUTION VERSION OF THE ECMWF FORECAST MODEL. *Monthly Weather Review*, **123,** 489-514.

Roadcap, J. R., J. van den Bosch, and W. Pereira, 2015: Analysis of Air Force Weather Agency cloud data for assessing single and multiple scattering through clouds at optical wavelengths. *J Appl Remote Sens*, **9,** 16.

Rosmond, T., and L. Xu, 2006: Development of NAVDAS-AR: non-linear formulation and outer loop tests. *Tellus Ser. A-Dyn. Meteorol. Oceanol.*, **58,** 45-58.

Slingo, J. M., 1987: THE DEVELOPMENT AND VERIFICATION OF A CLOUD PREDICTION SCHEME FOR THE ECMWF MODEL. *Quarterly Journal of the Royal Meteorological Society*, **113,** 899-927.

Stengel, M., C. Schlundt, S. Stapelberg, O. Sus, S. Eliasson, U. Willen, and J. F. Meirink, 2018: Comparing ERA-Interim clouds with satellite observations using a simplified satellite simulator. *Atmos. Chem. Phys.*, **18,** 17601-17614.

Suselj, K., J. Teixeira, and D. Chung, 2013: A Unified Model for Moist Convective Boundary Layers Based on a Stochastic Eddy-Diffusivity/Mass-Flux Parameterization. *Journal of the Atmospheric Sciences*, **70,** 1929-1953.

Suselj, K., T. F. Hogan, and J. Teixeira, 2014: Implementation of a Stochastic Eddy-Diffusivity/Mass-Flux Parameterization into the Navy Global Environmental Model. *Weather Forecast*, **29,** 1374-1390.

Teixeira, J., and T. F. Hogan, 2002: Boundary layer clouds in a global atmospheric model: Simple cloud cover parameterizations. *Journal of Climate*, **15,** 1261-1276.

Theurich, G., and Coauthors, 2016: The Earth System Prediction Suite: Toward a Coordinated U.S. Modeling Capability. *Bulletin of the American Meteorological Society*, **97,** 1229-1247.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, **98,** 163-173.

Wayand, N. E., C. M. Bitz, and E. Blanchard-Wrigglesworth, 2019: A Year-Round Subseasonal-to-Seasonal Sea Ice Prediction Portal. *Geophysical Research Letters*, **46,** 3298-3307.

Webster, S., A. R. Brown, D. R. Cameron, and C. P. Jones, 2003: Improvements to the representation of orography in the Met Office Unified Model. *Quarterly Journal of the Royal Meteorological Society*, **129,** 1989-2010.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, **132,** 1917-1932.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences.* Elsevier Academic Press.

Xu, L., T. Rosmond, and R. Daley, 2005: Development of NAVDAS-AR: formulation and initial tests of the linear problem. *Tellus Ser. A-Dyn. Meteorol. Oceanol.*, **57,** 546-559.

Zhao, Q. Y., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Monthly Weather Review*, **125,** 1931-1953.