THE SEA: THE SCIENCE OF OCEAN PREDICTION
Journal of Marine Research, 75, 403–433, 2017

Forecast Errors, Goodness, and Verification in Ocean Forecasting

by Gary B. Brassington[1,2]

ABSTRACT

Verification is an essential part of the forecast process that provides guidance on the statistical behavior of the system and a framework by which a forecast can be assessed for its "goodness." Much of the framework applicable to ocean forecasting has been developed within the atmospheric community. A review of the available material is given with some commentary on its relevance in the context of ocean forecasting. A statistical theory is presented for errors in an ocean forecast system (both deterministic and ensemble) and for a number of verification metrics. Theoretical results are demonstrated with empirical models and results from an operational ocean forecast system. Some new results are presented comparing the mean absolute error and root mean square error and the inference hypothesis testing of ensemble forecast systems. The progress in ocean verification is discussed, as are advances in technology to analyze international verification databases.

Keywords: forecasting, mean absolute error (MAE), root mean square error (RMSE), verification

1. Introduction

Ocean dynamics comprise a wide range of processes ranging from basin scale thermohaline and wind-driven gyre circulation to small-scale turbulence. In the majority of cases, ocean dynamics are chaotic, meaning the modeling of the dynamics is sensitive to small errors in the estimated initial state and boundary conditions, resulting in rapidly growing errors. A prediction system for a chaotic dynamical system comprises a method to estimate the initial conditions and a discretized model to evolve the state and circulation forward in time. From a stochastic point of view, the errors of a prediction system have a number of properties, as summarized in Figure 1. Let us consider the temperature at a specific location in the ocean to be represented by the time series shown by the solid line in Figure 1a. The temperature is observed by an instrument with a known instrumentation error, which is shown in Figure 1a as the lighter and thicker line following the true state. Based on this time series, we can model the temperature as a normal distribution with a mean temperature and variance, shown in Figure 1b by the blue shaded region. This *climatology* provides a

1. Bureau of Meteorology, Sydney, Australia
2. Corresponding author: ***e-mail: g.brassington@bom.gov.au***

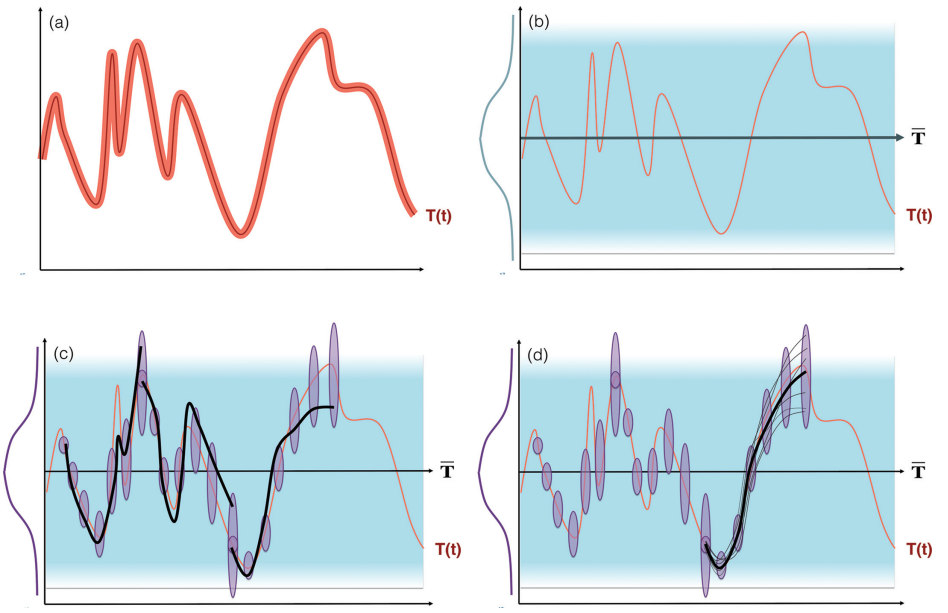

Figure 1. Schematic of ocean forecast uncertainty. (a) A generic time series for an observed variable T; lighter orange shows the observation instrument error; (b) the simplest forecast is to construct the climatological mean and variance as shown by the blue shading and the Gaussian along the y-axis; (c) three deterministic forecasts consisting of a sequential initial condition with a small uncertainty (purple ellipse) and a singlemodel integration with increasing uncertainty (purple ellipse) with time. Forecast uncertainty increases until the values exceed climatological forecast skill; (d) ensemble forecast represents multiple initial conditions followed by multiple model integrations from which the ensemble mean and ensemble variance can be estimated.

baseline stochastic forecast for temperature at that location where the most probable value is the mean and all temperature values within the recorded range have a non-zero probability of occurrence. In the absence of climate change, the climatology provides increasingly robust estimates of the statistical mean and variance with increasing record length. A climatological record is a useful guidance for some purposes, including long-range planning and design. However, for many situations an accurate forecast would lead to positive social and economic benefits. There are effectively two main modeling approaches to forecasting: stochastic and deterministic. Persistence, the simplest stochastic forecast, can perform best in the near-term, while a deterministic forecast performs best on the short- to medium-term, and a combination (ensemble) is best on longer timescales. The state of the art in ocean forecasting has been dominated by deterministic approaches for both global and shelf scale forecasting, owing to the computational expense of these models and maturity of the science and, particularly, the services to derive the benefit. However, multi-model (poor-man's)

ensemble (Spindler et al. 2016) and multi-cycle ensemble systems (Brassington 2013) have recently emerged.

The errors of a well-behaved prediction system are shown in Figure 1c, where the initial conditions have a relatively small error represented by the smallest purple ellipse. Typically, this error is larger than the observation error variance but much smaller than the variance in climatology. As the forecast model evolves forward in time, the error ellipse (uncertainty) increases and eventually can produce forecasts with greater error than a climatological forecast. At or prior to this time, the model state is re-initialized based on newly obtained observations, reducing the uncertainty from which the model is then evolved for another forecast period. A deterministic forecast is a single realization of this stochastic system, as shown by the thick black line in Figure 1c. The rate of error growth varies from forecast to forecast depending on the dynamics and forcing occurring in each instance. Therefore, it is not possible to come up with reliable estimates for the uncertainty other than a statistical estimate based on previous hindcasts. We can improve both the forecast estimate of the state and also provide an estimate of the uncertainty by the use of an ensemble of forecasts (Fig. 1d). In this case, an ensemble of deterministic forecasts is performed, each member with perturbed initial conditions based on the expected uncertainty. For an unbiased and reliable ensemble, the best estimate (lowest error) and uncertainty are given by the ensemble mean and ensemble variance, respectively.

In addition to the forecast error growth varying in time, as depicted by the example outlined above, error growth also varies in spatial location. At any instant in time, the dynamics will comprise discrete regions that are unstable (rapid error growth) and other regions that are relatively stable (slow error growth). The spatial extent, error magnitude, and location all evolve in time. For example, Figures 2b and 2d show the monthly mean of daily ensemble variance for January 2012 and January 2013 in the Tasman Sea from an operational ocean forecast system at the Bureau of Meteorology, which is a four-cycle, time-lagged ensemble system (Brassington 2013). In this example, the large forecast variance corresponds to the position of temperature fronts as estimated by the corresponding ensemble monthly mean of sea surface temperature in Figures 2a and 2c, respectively.

Obtaining benefit from a forecast system rests with the question of whether it is a "good" forecast system. This question can be answered with some precision with the use of validation, verification, and value. Validation and verification are terms used extensively in other fields but may have slightly different meanings. In the context of ocean forecasting, the modeling, predictability, and prediction system are all close analogies with atmospheric forecasting, a chaotic geophysical fluid.

The atmospheric science community has systematically developed a framework and definitions (e.g., Murphy 1993), much of which can be directly reapplied in this context. Validation is concerned with the question of whether the ocean model resolves (i.e., is representative of) the physical processes present in the observations to within some requirement or threshold. Ocean models include numerous assumptions (e.g., incompressibility,

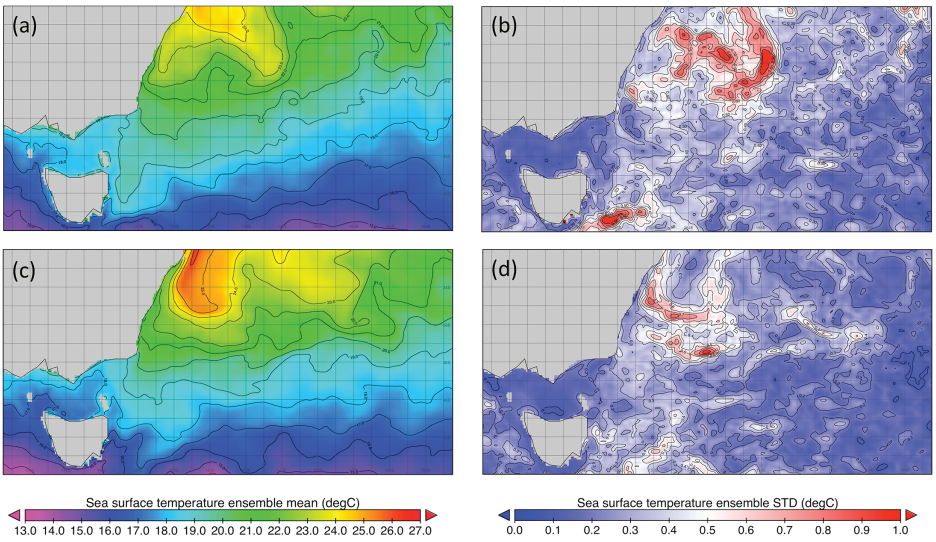

Figure 2. The January four-cycle monthly mean, ensemble mean sea surface temperature for (a) 2012 and (c) 2013, and the corresponding monthly mean, four-cycle ensemble variance for (b) 2012 and (d) 2013. The four-cycle ensemble mean is based on the OceanMAPS operational prediction system at the Bureau of Meteorology (Brassington 2013).

Boussinesq, and hydrostatic), which are convenient in terms of numerical method choices and are valid for a wide range of phenomenon. Other specific configuration design choices can further limit the processes represented or resolved, such as the discretization method, coordinate system, and the time and spatial resolution. An important initial step is to demonstrate that the model configuration is valid for the location and phenomena being represented. Verification is about measuring and monitoring the quality or performance of the forecast system. An extensive range of statistical measures with respect to forecast verification can be found in atmospheric science (e.g., Jolliffe and Stephenson 2012). Finally, forecast value concentrates on whether the forecast system quality is sufficient to meet the user's requirements and adds value in terms of their decisions and outcomes. Necessarily in a maturing science, measures of value tend to be the last to develop.

We first survey how verification is defined within atmospheric science in Section 2. Despite the close analogy, there are several important distinctions that lead to different choices in terms of the metrics to be applied, as well as their modification or adaptation in ocean forecasting, as described in Section 3. In Section 4, we review some of the basic statistical theory that is being applied in ocean forecasting and highlight some important results together with examples. We then provide an overview of the progress and current state of verification in the field of ocean forecasting in Section 5, as well as some of the supporting technology being developed in Section 6, before summarizing.

2. Survey of verification from atmospheric science

There is an extensive literature in atmospheric science on verification, with well-established statistical concepts and experience on their interpretation and pitfalls. There are several books and reports (Stanski et al. 1989; Wilks 2011; Jolliffe and Stephenson 2012), special issues (Ghelli and Ebert 2008) and websites (http://www.cawcr.gov.au/projects/verification/; http://verification.nws.noaa.gov; http://www.eumetcal.org) covering a wide range of topics.

The majority of human activity takes place over land and within the atmosphere. Therefore, weather phenomena of all types (e.g., temperature, humidity, precipitation, winds, and fog) shape and influence many human activities and decisions from an individual level (i.e., what to wear, the planning of one's day, and one's emotional state) to group activities and decisions (i.e., building codes, event planning, and emergency response). Many applications require information of a general nature that is served by forecasts of the atmospheric circulation. In these cases, verification using basic statistical metrics such as mean error and mean absolute error are sufficient to quantify the expected performance. As the application becomes less tolerant of particular aspects of the weather, more specialized information is required to minimize the risks and impacts of bad forecast information. For example, forecasts for binary events (e.g., a storm did or did not occur, did or did not arrive) lead to specific verification concepts of Probability of Detection (POD) and Relative Operating Characteristic (ROC) diagrams (see chapter 3 in Jolliffe and Stephenson 2012). In addition, the atmosphere is a chaotic system leading to rapid error growth and forecast uncertainty. Ensemble forecasting is the dominant pragmatic approach used to characterize the probability distribution of this high dimensional problem (chapter 8, Jolliffe and Stephenson 2012). For a comprehensive collection of papers covering atmospheric predictability, ensemble forecasting, and verification, see Palmer and Hagedorn (2006).

The important question, "What is a good forecast?" has a long history in atmospheric science (Winkler and Murphy 1968; Murphy and Winkler 1987; Murphy 1993), initiated well before numerical weather prediction (NWP) systems acquired consistent skill. Many forecast methods were developed preceding NWP, including forecast experience, historical records, and statistical forecasts. Murphy (1993) outlines key concepts relating to the goodness of a forecast in terms of *consistency, quality, and value*. Consistency concerns the degree to which the forecast matches the forecaster's best judgment based on knowledge and experience. Quality is the degree to which the forecasts correspond to what actually happened. Value relates to the degree to which the forecast assists in decisions that will have economic or other benefit. Verification predominantly is concerned with the statistical measures related to quality, though in order for forecasts to have a significant impact, they need to be consistent and have some value or benefit. The important properties of quality cannot be condensed into a single metric. Rather, they can be expressed in a number of statistical properties based on the joint, conditional, and marginal distribution of the forecast and observations (Murphy and Winkler 1987). Murphy (1993) outlined nine aspects that define forecast quality, which are restated with some example metrics in Table 1 (others

Table 1. Concise definitions for aspects of forecast quality and example metrics, where T_M and T_O represent the model forecast and observed state, respectively, and σ_M and σ_O represent the standard deviation of the model and observation, respectively. E represents expectation, | condition, *ens* represents an ensemble forecast.

Aspect	Definition	Example metrics
Bias	Correspondence between mean forecast and mean observation	$E(T_M) - E(T_O)$
Association	Strength of the linear relationship between forecast/observation pairs	$\frac{E[(T_M - E(T_M))(T_O - E(T_O))]}{\sigma_M \sigma_O}$
Skill	Average correspondence between pairs of forecasts and observations	$E(\lvert T_M - T_O \rvert)$
Reliability	Correspondence between conditional mean observation and conditional forecast over all forecasts	$E_{allM}(T_O \lvert T_M = T)$
Resolution	Difference between conditional mean observation and unconditional mean observation averaged over all forecasts	$E_{allM}[(E(T_O \lvert T_M = T) - E_{allM}(T_O))^2]$
Sharpness	Variability of forecasts as described by distribution of forecasts	$E[(T^{ens} - E(T^{ens}))^2]$
Discrimination 1	Correspondence between conditional mean forecast and conditioning observation, averaged over all observations	$E_{allO}(T_M \lvert T_O = T)$
Discrimination 2	Difference between conditional mean forecast and unconditional mean forecast, average over all observations	$E_{allO}(T_M \lvert T_O = T) - E_{allMO}(T_M)$
Uncertainty	Variability of the observations	$E[(T_O - E(T_O))^2]$

can be found in section 2.10 of Jolliffe and Stephenson 2012). Another useful discussion of these aspects is given in Stanski et al. (1989).

3. Applying verification in ocean science

One of the leading distinctions between ocean forecasting and atmospheric forecasting is the maturity and legacy of the latter. Immaturity is both an advantage and disadvantage. For a variety of reasons, including less general public interest for ocean forecast information as well as numerous technical challenges, ocean forecasting as it is today only began to be

defined in the 1980s (Hurlburt 1984). It first became feasible with the launch of satellite altimetry and deployment of an array of autonomous profiling floats (Argo; Gould et al. 2004) in the 1990s. Ocean forecasting has therefore developed over a very short period of time due to a coincidence of a mature science, modeling, and computational technology when the observing system was implemented. Therefore, ocean forecasting has not had a legacy of existing forecast practices to overcome, including the role of humans in the forecast process. There is a significant portion of atmospheric verification concerned with the question of the human forecaster and objectively quantifying the value that is added. Many of these questions and metrics have much less relevance in ocean forecasting. Nonetheless, this has been a disadvantage in ocean forecasting, as the absence of an experienced team of forecasters and established sets of forecast practices has limited the value and impact as the early immature prediction systems suffered from inconsistent skill and model bias. Ocean forecasting development has evolved rapidly to higher levels of reliability in order to deliver value and impact with more limited human forecaster guidance.

a. Physical and dynamical processes and forecast errors

In the Introduction, we noted that the closest analogy to the ocean is the atmosphere. Both the atmosphere and ocean are thin, planetary-scale fluid layers with a comparable system of governing equations (Navier-Stokes on a rotating planet). Many of the dynamical concepts (e.g., conservation of potential vorticity, Rossby waves, barotropic or baroclinic instabilities, and geostrophic turbulence) are common (see Gill 1982). However, there are several unique dynamic features found in the ocean and dynamic processes that have greater significance than the atmosphere. Among of the leading-order differences are the changes in temporal and spatial scales and power spectra that are related to the difference in fluid parameters (e.g., density, viscosity, and specific heat).

The density of seawater is a thousand times greater than air. It requires significant energy transfer to raise and lower the ocean surface, such as the tidal force, wind-driven surge, earth quake-generated Tsunamis; otherwise, the motion of the ocean is bounded within ocean basins and marginal seas. The general circulation of the ocean, featuring basin-scale gyres and western boundary currents (Stommel 1948), have no analog in the atmosphere. The Antarctic Circumpolar Current is the only feature analogous to the atmospheric jet stream. Western boundary current regions are the most energetic, leading to a cascade of eddies and fronts that are chaotic and the most challenging regions to forecast. Being located adjacent to the continental shelf with typically high population centers, they are also important regions to forecast.

The density has an important impact on energy transfer between the ocean and atmosphere. The kinetic energy of an ocean current of 1 ms^{-1} is equivalent to an atmospheric wind of 31 ms^{-1} (113 km hr^{-1}.) The potential energy needed to raise a parcel of ocean by 1 m is equivalent to raising a parcel of atmosphere by 1 km. It also impacts important length scales such as the internal Rossby radius of deformation (Gill 1982), which is an order of magnitude smaller in the ocean leading to a higher number of eddy wavelengths

in the ocean basin. These scales have important implications for the resolution of models (∼10 km Hurlburt 1984 or finer, Hurlburt and Hogan 2000) and the observing system.

Adjacent to the majority of coastlines is a zone referred to as the continental shelf, which extends from the coastline to a shelf break of ∼100–200 m depth before a more rapid decent to abyssal depths of 4000 m or greater. The continental shelf break approximately represents a contour of potential vorticity, a conserved quantity in an earth rotating fluid system. Significant energy is required to transfer vorticity or mass or both between the shelf and deeper ocean, which could be induced by atmospheric winds or ocean currents. In addition, the shallower depths lead to higher amplitude tidal heights and currents, leading to a greater proportion of the water column that is well mixed and a greater influence of bottom drag on the water column. The presence of the coastline boundary also traps atmospheric energy, leading to storm-surge, upwelling, and coastally trapped waves that are less prominent in the atmosphere. The modeling and observing system for this region requires higher temporal sampling.

The specific heat, together with the ocean density, leads to a heat capacity 4000 times greater than air. The majority of the heat transfer from the equator to mid-latitudes is performed by the ocean (Wunsch 2005), which absorbs heat in the tropics driving a complex global scale, density gradient response modified by salinity sources and sinks, and poleward deep convection, referred to as the thermohaline circulation. The timescale of this circulation is multi-decadal; however, its magnitude and volume is perturbed by the annual cycle. In the upper layer, where there is a high concentration of *in situ* and remotely sensed observations, the time and space scales are well observed. However, at abyssal depths, this circulation is poorly observed with unknown contributions to forecast biases.

Away from the polar regions (sea-ice) the only state changes for the ocean occur at the air-sea interface in the form of evaporation leading to a latent heat exchange and loss of mass. Otherwise, there are no internal state changes analogous to that of the atmosphere. There are specialized techniques in atmospheric verification for clouds, convection, and precipitation that are not required in ocean forecasting.

The rapid absorption of electromagnetic energy of seawater limits the penetration of solar radiation to the upper ocean, though the biological impact of the euphotic zone is deeper at ∼200 m. Radiation physics is therefore much simpler than the atmosphere, with the exception being the impact of phytoplankton on the absorption. Present ocean forecasting systems use a climatological distribution of biology based on remotely sensed ocean color to estimate this effect rather than including an active biogeochemical model (e.g., Ohlmann, et al. 1996).

More generally, the air-sea interface exchanges mass, momentum, and heat and is estimated through empirically derived relationships (e.g., Large and Pond 1982; Large et al. 1997) with extensions for high-wind speed conditions (e.g., Powell et al. 2003; Moon et al. 2007). The air–sea exchange is composed of a large number of processes under low-wind conditions (Edson et al. 2007) with additional processes under high-wind conditions (Black et al. 2007). The exchange of momentum and buoyancy fluxes results in fine-scale

turbulence in the upper ocean, which is not directly resolved but parameterized. Unlike the atmospheric boundary layer, which has an ocean boundary that persists, the ocean-mixed layer has a surface boundary condition that has low persistence. The errors of the atmospheric model therefore contribute significantly to the upper ocean and coastal shelf forecast errors.

Other areas of ocean forecasting that require specialized modeling with implications for verification include the sea-ice, coupled ocean-bio-geo-chemistry models, coupled ocean-sediment models, and coupled air-wave-sea models.

b. Observing system sources of forecast errors

Since the 1990s, a global ocean observing system was sequentially implemented to a target threshold and, through community effort (Hall et al. 2010), has been sustained. Nonetheless, with respect to the mesoscale ocean dynamics it remains under-sampled (Brassington et al. 2010). The majority of operational global ocean forecast systems assimilate all available real-time ocean observations (Dombrowsky et al. 2009). The availability of independent observations is restricted to new research platforms and research campaign data, which are limited in temporal and spatial extent and not suited to systematic verification. Verification is therefore limited to the evaluation of system forecasts where the observations are "independent," as not previously assimilated by the prediction system. There remain flaws to this verification when systems are compared. For example, assimilation of sea surface temperature (SST) from drifting buoys has persistence as the drifting buoy is tethered to a drogue in order to follow the ocean currents and therefore sample the same parcel of ocean. Systems that assimilate drifting buoy observation benefit from the observation persistence, unlike systems that do not assimilate. A second example is satellite altimetry. The products that are assimilated require a significant amount of processing, relying on a mixture of observed and model estimated corrections. Systems that assimilate the reference data have an advantage compared with systems that assimilate data processed by a different center (e.g., AVISO (http://www.aviso.altimetry.fr), RADS) (http://rads.tudelft.nl), USGODAE (http://www.usgodae.org)).

Due to the absorption of electromagnetic radiation, remote sensing of the ocean is limited to observing surface properties including sea surface height, sea surface temperature, and sea surface salinity (exploiting the state dependence of brightness temperature).

Sea surface height varies across a broad range of time scales and processes, including barotropic processes with gravity waves and tides as well as baroclinic density or specific volume changes, or both. The relatively small sea level changes from the baroclinic density anomalies (dynamic range of ± 1 m) provide a description for the distribution of mass analogous to sea level pressure in the atmosphere. Similarly, the distribution of mass is related to the ocean circulation through the geostrophic balance. For global ocean forecast systems, these sea surface height anomalies are the dominant source of information used to constrain the ocean model. However, there are several limitations to altimetry measurements that need to be understood in their use in data assimilation as well as verification. Each

sea surface height anomaly observation is derived as a small residual requiring up to 10 correction terms for the removal of the effects of waves, tides, wet troposphere, and others. Nonetheless, observation errors have been systematically reduced to ∼5 cm (Fu et al. 1994) over the open ocean. In the near real-time altimetry products, some of the corrections use model-estimated products, which can introduce biases into the observations. All altimeters to date are based on nadir instruments on repeat orbits of periods of 9.9 days (Jason-series) or greater. A nadir track provides information of sea surface height gradient only in the along-track direction. Only through multiple satellite altimeters are gradients in the orthogonal direction obtained within a small time window; however, this occurs for at most one point for each ascending and descending orbit per satellite pair. In addition, the nominal resolution of a single altimeter is ∼500 km while two tandem satellites provide an effective resolution of ∼100 km (Fu et al. 2003). It has been demonstrated that four nadir polar-orbiting satellite altimeters are required to constrain an ocean forecast system (Pascual et al. 2006) resolving the mesoscale ocean variability. To this end, cross-calibration of multiple altimeters must be performed, which involves using the Jason-series as the reference altimeter. The time windows and spatial regions need to be sufficiently large to retain an adequate sample size to obtain robust statistics.

Sea surface temperature is observed by multiple satellites with instruments in the infrared bands (which provide higher spatial resolution but are absorbed by clouds) and microwave bands (that are coarser resolution but can observe through clouds but not precipitation). Cloud detection algorithms, quality control, and cross-calibration are undertaken by an international science team (https://www.ghrsst.org/). Cloud clearing algorithms can be problematic for coastal upwelling regions where temperature of the ocean is comparable to that of stratus clouds. Diurnal warming is also an important consideration, which can form a shallow warm layer and mask the underlying foundation temperature. Quality control in this case is based on defining thresholds for the 10-m winds required to remove the vertical temperature gradient in the skin layer (Donlon et al. 2002). The 10-m wind correction is dependent on atmospheric models rather than observed. An alternative strategy adopted by the Bureau of Meteorology is to limit the observations to nighttime equator crossings for polar orbiting satellites, where the threshold on winds is lower and the likelihood of large skin gradients is minimized.

Sea surface salinity remote sensing is based on the dependence of brightness temperature on ocean salinity (Koblinsky et al. 2003). Two missions, Aquarius (Koblinsky et al. 2003) and Soil Moisture Ocean Salinity (SMOS; Font et al. 2004), were launched with relatively large instrument errors due to the weak relationship. The impact of these observations in a multivariate analysis with observations of SST and altimetry limit the region of impact to the tropics (Brassington and Divakaran 2009). However, in this region, the impact is reduced by the lower quantity of data due to interference of cloud or precipitation.

In situ observations of temperature and salinity are composed largely of vertical profiling instruments from fixed moorings, expendable bathythermographs and conductivity temperature depth (CTD) sensors from autonomous profiling floats and gliders. These instruments

require quality control, which includes a set of generic tests and instrument specific tests to remove "bad" observations (Ingleby et al. 2007; Cummings et al. 2010). Temperature sensors are relatively stable; however, salinity instruments can deteriorate over time and require recalibration, and they can sometimes be blacklisted for real-time applications.

Observation errors include measurement errors and representative errors (Daley 1993). Representative errors are related to both grid resolution, which is model-dependent, and the relative power of the sub-grid scales in the observations, which is state-dependent and measurement instrument-dependent (Oke and Sakov 2008). The general formulation to estimate the observation error variance is given by

$$\sigma_O^2 = \sigma_{\text{inst}}^2 + \sigma_{\text{RE}}^2, \tag{3.1}$$

where O, inst, and RE represent observation, instrument, and representative error, respectively. The representative error can be approximated by a function of the model variance (Oke and Sakov 2008). While instrument errors are generally much smaller than model errors, representative errors are larger and comparable to model errors. Representative errors are frequently applied in data assimilation to penalize observations for scales unresolved by the model and increase the weighting of the model background. Similarly, it is important to account for these errors in forecast verification.

4. Theory

An error model for deterministic and ensemble forecasts of generic and observable state variables is presented with assumptions as to their stochastic behavior. This error model is then applied to a variety of common statistic operators for both large and small sample sizes in order to discuss and interpret their properties.

A deterministic forecast model (M) estimates a state variable T (e.g., temperature) as

$$T_M^{\text{biased}} = T_S + \varepsilon_M + \beta, \tag{4.1}$$

where T_S is the true value of the state variable, ε_M is the random error of the model, which we assume is normally distributed $N(\beta, \sigma_M^2)$, and β is the bias.

We assume that the state variable is randomly observed such that

$$T_0 = T_S + \varepsilon_0, \tag{4.2}$$

where ε_0 is the random error of the observations, which we assume is normally distributed, $N(0, \sigma_O^2)$, and unbiased (through calibration). The expectation (average) of the innovation (difference between the model and the set of observations) is given by

$$\begin{aligned} E[T_M^{\text{biased}} - T_O] &= E[\varepsilon_M] - E[\varepsilon_O] + E[\beta] \\ &= \beta \end{aligned} \tag{4.3}$$

We can therefore diagnose the bias using the expectation of a large sample of innovations and retrospectively derive an unbiased modeled estimate, i.e., $T_M = T_M^{\text{biased}} - \beta$. We can then construct the innovation variance as

$$\begin{aligned} E[(T_M - T_O)^2] &= E[\varepsilon_M^2] - 2E[\varepsilon_M \varepsilon_O] + E[\varepsilon_O^2], \\ &= \sigma_M^2 + \sigma_O^2 \end{aligned} \tag{4.4a}$$

where the model and observation errors are independent and uncorrelated. It can be shown that the innovation variance for a biased is given by

$$E[(T_M^{\text{biased}} - T_O)^2] = \sigma_M^2 + \sigma_O^2 + \beta^2 \tag{4.4b}$$

a. Statistical sampling

Verification of ocean forecasts are all based on finite sample sizes. It is therefore important to understand the changes in behavior of the most common metrics. The first metric to consider is the sample mean, which can be defined as

$$\begin{aligned} \bar{T}_M^{\text{biased}} &= \frac{1}{k} \sum_{i=1}^{k} T_M^{\text{biased}} \\ &= T_S + \bar{\varepsilon}_M + \beta \end{aligned}, \tag{4.5}$$

where the overbar is the notation for a sample mean, k is the sample size, and $\bar{\varepsilon}_M$ is the sample mean of the random model error. It is important to note that the sample mean is itself a random variable and can be shown to have the same expectation as the original random variable, but a reduced variance such that, $\bar{\varepsilon}_M$ is a normal distribution, $N\left(\beta, \frac{\sigma_M^2}{k}\right)$. Similarly, the observation error can be expressed as

$$\begin{aligned} \bar{T}_O &= \frac{1}{k} \sum_{i=1}^{k} T_O \\ &= T_S + \bar{\varepsilon}_O \end{aligned} \tag{4.6}$$

where $\bar{\varepsilon}_O$ is a normally distributed $N\left(\beta, \frac{\sigma_O^2}{k}\right)$.

The sample variance (s^2) can be based on a system with a known expectation or a sample mean. Verification of the latter is the more common. In this case, the sample variance must be defined in terms of the unbiased estimator,

$$\begin{aligned} S_{M,k-1}^2 &= \frac{1}{k-1} \sum_{i=1}^{k} (T_M^{\text{biased}} - \bar{T}_M^{\text{biased}}) \\ &= \sigma_M^2 \frac{\chi_{k-1}^2}{k-1} \end{aligned}, \tag{4.7}$$

where χ_{k-1}^2 is a chi-squared distribution, $E[\chi_{k-1}^2] = k - 1$, and $\text{VAR}[\chi_{k-1}^2] = 2(k-1)$.

b. Mean absolute error and root mean square error

An unbiased forecast system does not imply that it is perfect due to the fact that errors of different sign cancel each other in a mean error metric. Two metrics frequently used as a measure of forecast error are the mean absolute error (MAE), a first-order moment, and root mean square error (RMSE), a second-order moment.

The MAE for a biased model can be expressed as

$$\begin{aligned} \mathrm{MAE} &= E[|T_M^{\mathrm{biased}} - T_O|] \\ &= E[|\varepsilon_M + \beta - \varepsilon_O|] \\ &= \sigma_{\mathrm{total}}\sqrt{2/\pi}e^{(-\beta^2/2\sigma^2_{\mathrm{total}})} - \beta erf(-\beta/\sqrt{2}\sigma_{\mathrm{total}}) \end{aligned} \quad , \tag{4.8a}$$

where we note that the difference of two normally distributed errors is another normally distributed error given by $N(\beta, \sigma_M^2) - N(0, \sigma_M^O) \equiv N(\beta, \sigma^2_{\mathrm{total}})$, where $\sigma^2_{\mathrm{total}} = \sigma_M^2 + \sigma_O^2$ and the expectation of the folded normal distribution (Leone et al. 1961) is given by $E[|\varepsilon_X|] = \sigma_X\sqrt{2/\pi}e^{(-\beta^2/2\sigma_X^2)} - \beta erf(-\beta/\sqrt{2}\sigma_X)$, where erf is an error function. For a biased corrected model $\beta = 0$, the folded normal distribution reduces to $E[|\varepsilon_X|] = \sigma_X\sqrt{2/\pi}$ and the MAE is given by

$$\begin{aligned} MAE &= E[|T_M - T_O|] \\ &= E[|\varepsilon_M - \varepsilon_O|] \\ &= \sqrt{\frac{2}{\pi}}\sqrt{\sigma_M^2 + \sigma_O^2} \end{aligned} \tag{4.8b}$$

The RMSE can be expressed as

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{E[(T_M^{\mathrm{biased}} - T_O)^2]} \\ &= \sqrt{\sigma_M^2 + \sigma_O^2 + \beta^2} \end{aligned} \tag{4.9a}$$

where we have substituted Eq. (4.4b). For an unbiased or bias-corrected model, Eq. (4.9a) reduces to

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{E[(T_M - T_O)^2]} \\ &= \sqrt{\sigma_M^2 + \sigma_O^2}, \end{aligned} \tag{4.9b}$$

where $\beta = 0$.

Using an empirical statistical model, we will highlight the difference in properties and discuss in what context MAE and RMSE might be better applied in ocean forecasting. The ocean state T_S is assumed to be a normally distributed random variable with unit variance $N(0, 1)$. The forecast error is defined by Eq. (4.1), where ε_M is modelled by $N(0, 0.2)$

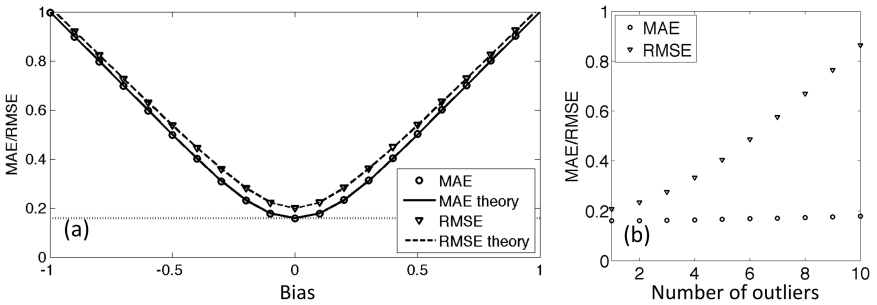

Figure 3. (a) The empirical estimates of mean absolute error (MAE; circle) and root mean square error (RMSE; triangle) for a range of biases are shown. This is compared with the theory Eq. (4.8b) (dashed) and Eq. (4.9b) (solid) lines, respectively. The dotted line corresponds to $\sqrt{2/\pi}\sqrt{\sigma_M^2 + \sigma_O^2}$, the value in Eq. (4.9a), corresponding to zero bias. (b) Empirical estimates of MAE (circle) and RMSE (triangle) relative to an increasing number of outliers.

such that $\sigma_M^2 < 1$, i.e., less than the state variance. The MAE and RMSE are estimated using a sample size of 10,000 and a range of biases $-1 \leq \beta \leq 1$, as shown in Figure 3a. Both MAE and RMSE provide a symmetric behavior as they are both positive metrics. For low bias, the RMSE provides an estimate of the model variance while MAE provides a more optimistic measure. Both metrics increase approximately linearly and converge with increasing bias magnitude. It is therefore useful to diagnose and remove any gross biases. It is noted that MAE is less sensitive than RMSE to statistical outliers (Jolliffe and Stephenson 2012, section 5.3.2). Using the same statistical model for the forecasts, the observations are augmented to include an increasing number of outliers modeled by random model with a higher likelihood of, large-magnitude errors, $\varepsilon_{\text{outlier}} = \text{sgn}(\varepsilon_1)|2 + 2\varepsilon_2|$, where both ε_1 and ε_2 are $N(0, 1)$. As shown in Figure 3b, while the MAE grows slowly with a small but increasing number of outliers, the RMSE is unstable and grows rapidly.

For real-time verification applications where automatic quality control will permit a small number of outliers, the MAE metric should be adopted in preference to RMSE due to the stability to "bad observations." It is important to note that results from RMSE and MAE can be compared provided an account is made of the scaling factor $\sqrt{2/\pi}$. While the stability or insensitivity of MAE to outliers is an advantage when the outliers reside in the observations, it is a disadvantage when the outliers occur in the model forecasts. If the ability of the forecasts to estimate rare extreme conditions is important, then there is value in repeating the forecast verification *posteriori*, using delayed mode, quality-controlled observations (i.e., free of fictitious outliers) based around the RMSE metric. When the random errors are stationary, the influence of bias is similar for both MAE and RMSE, in that metrics show little change for biases that are small relative to the error variance but grow to a linear increase as the bias becomes large.

The sample mean absolute error based on Eq. (4.8a) and Eq. (4.8b) is given by

$$\mathrm{MAE} = \overline{|T_M - T_O|}, \tag{4.10}$$

where the overbar is the notation for a sample mean. The expectation and variance is shown in (G. B. Brassington, unpubl. data) to be given by

$$E[\overline{|T_M - T_O|}] = \sqrt{\sigma_M^2 + \sigma_O^2}\sqrt{\frac{2}{\pi}}, \text{ and} \tag{4.11a}$$

$$\mathrm{VAR}[\overline{|T_M - T_O|}] = (\sigma_M^2 + \sigma_O^2)\frac{1}{k}\left(1 - \frac{2}{\pi}\right), \tag{4.11b}$$

where the expectation and variance is based on a folded normal distribution (Leone et al. 1961) for $N(0, \sigma_M^2 + \sigma_O^2)$ is modified by the sample size.

The sample root mean square error is given by

$$\mathrm{RMSE} = \sqrt{\overline{(T_M - T_O)^2}}, \tag{4.12}$$

where the overbar is the notation for a sample mean. The expectation and variance is shown in (G. B. Brassington, unpubl. data) to be given by

$$E\left[\sqrt{\overline{(T_M - T_O)^2}}\right] = \sqrt{\sigma_M^2 - \sigma_O^2}\sqrt{\frac{2}{k}}\frac{\Gamma((k+1)/2)}{\Gamma(k/2)}, \text{ and} \tag{4.13a}$$

$$\mathrm{VAR}\left[\sqrt{\overline{(T_M - T_O)^2}}\right] = (\sigma_M^2 - \sigma_O^2)\left(1 - \frac{2}{k}\left(\frac{\Gamma((k+1)/2)}{\Gamma(k/2)}\right)^2\right), \tag{4.13b}$$

where the RMSE can be modeled by a weighted chi-distribution, Γ is the gamma function, and k is the sample size.

Using the same empirical statistical model for the forecast error and observation error, as described above, we now compare the theoretical results for the expectation and variance of the sample MAE and RMSE. For each sample size (degrees of freedom) of [10, 15, 20, 30, 40, 60, 80], a random sample of the error pairs is obtained to compute the MAE and RMSE metrics. This is repeated 5000 times for each sample size in order to estimate the mean and standard deviation of the metrics. The values for the error in the mean value from the asymptotic expectation values (Eq. 4.8b and Eq. 4.9b, respectively) are shown in Figure 4 for MAE (circle) and RMSE (square) relative to the sample size. These empirical values compare well with the theory in Eq. (4.11a) and Eq. (4.13a) for MAE (blue) and RMSE (red), respectively. The empirical values for the standard deviation for both MAE and RMSE also compare very well with the theory in Eq. (4.11b) and Eq. (4.13b), respectively, as shown as dotted lines in Figure 4.

The theory corroborated by the empirical data shows that the RMSE has larger error for smaller sample sizes but converges to be indistinguishable to MAE for the parameters of

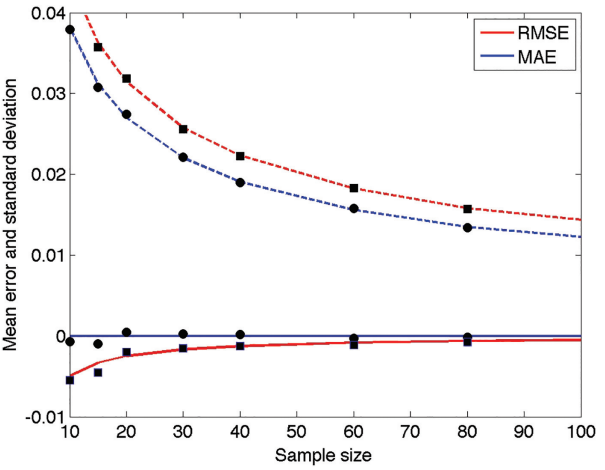

Figure 4. Sample mean error and standard deviation for the mean absolute error (MAE) and root mean square error (RMSE) relative to sample size. The analytical theory for the error in the expectation of the sample for MAE (blue) (Eq. 4.11a) and RMSE (red) (Eq. 4.13a) is shown as a solid line relative to the asymptotic value for an infinite sample (Eq. 4.8b) and (Eq. 4.9b), respectively. The standard deviation of the sampled MAE and RMSE are shown as a dotted line. The equivalent sampled empirical model is shown for MAE (circle) and RMSE (square).

this experiment, at sample sizes greater than 100. The theory also shows a persistently lower uncertainty (standard deviation) for MAE than with RMSE. These results provide further support for the conclusion that MAE should be used for measuring errors in ocean forecasting. Root mean square error should be used as indicated above, but with the additional caveat of sample sizes in excess of 100. A debate on the relative merits of MAE and RMSE is contemporary with empirical arguments put forward favoring MAE over RMSE (see Willmott and Matsuura 2005) and other papers countering against some of their claims (see Chai and Draxler 2014). Here we provide the theoretical underpinning to both schemes and examine the properties of both schemes. In an unbiased or de-biased system, both metrics are estimating the same quantity within a scale factor $\sqrt{\frac{2}{\pi}}$. The distinction between the two metrics relates to their convergence, uncertainty, and sensitivity to outliers. In general, MAE has the more favorable behavior. However, whether the sensitivity to outliers is "good" or "bad" is application-specific.

This statistical theory is based upon a number of statistical properties, including the independence of the modeled and observed random errors, the stationarity of the normal distributions, and the sample size (or degrees of freedom). Of these, the most difficult to comply with in ocean forecasting is independence. It is common practice to treat verification of the forecasts against new observations as independent. However, there is a potential flaw in this assumption with respect to systematic errors. Only the combined bias can be removed through the mean error. Any part of the bias that is shared by the two systems will not

be detectable other that through calibration. If an observing platform is biased, there is no account made within data assimilation as it is applied today, and such biases will be included in the analyses. If some fraction of these biases project onto large spatial scale modes, then they will have persistence and be reinforced by the sequential assimilation of observations from the same platforms. Metrics based on model-observation differences such as MAE and RMSE will not be impacted by a shared bias. Comparison metrics such as anomaly correlation will be impacted.

c. Skill scores and summary metrics

During the early development of verification, there was a pursuit of "the score" (Murphy 1988) that summarized all important aspects of a forecast system from which the notion of "best" can be definitively quantified. The S1 skill score has a long history in NWP as a baseline score upon which system performance is published, presented, and differentiated; however, there are known limitations. For example, the S1 skill score depends only on gradients and cannot address the problem of systematic errors. Secondly, there is a dependence on the spatial resolution and the spectrum resolved (WCRP-JWG 2016). As outlined in Section 2, the atmospheric verification community abandoned the notion of a single score and moved to a framework based on the joint, marginal, and conditional distributions of a model and observing system, outlining the various aspects of a "good" forecast and the metrics that can be used in that assessment (Table 1).

There remains value in a summary score provided its limitations are understood and the differentiation in score is not used for decisions on system design and investment beyond those limits. We will follow Tonani et al. (2009), who define a skill score based on the relative change in RMSE of the forecast with a reference truth and persistence. The RMSE for the forecast and persistence forecasts are defined as

$$\mathrm{RMSE}_M^{FC}(\Delta t_f) = \left\langle \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(T_{M,n}^{AN}(\Delta t_f) - T_{M,n}^{FC}(\Delta t_f)\right)^2} \right\rangle \tag{4.14a}$$

$$\mathrm{RMSE}_M^{PFC}(\Delta t_f) = \left\langle \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(T_{M,n}^{AN}(\Delta t_f) - T_{M,n}^{AN}(\Delta t_f = \Delta t_0)\right)^2} \right\rangle \tag{4.14b}$$

where FC represents a model forecast and AN represents the analysis after the model has been initialized with the data assimilation increment, $\Delta t+ = f\Delta t$ where Δt is the forecast time interval and $f \in [0, \ldots, F]$, Δt_0 is the last time interval before real-time, $F\Delta t_f$ is the forecast period, and N is the number of forecasts included in the sample mean. An additional operator $\langle\ \rangle$ can be included to compress the model fields into horizontal or vertical averaging, or both. Following Murphy et al. (1988), a skill score can be defined as

$$SS(\Delta t_f) = \left(1 - \frac{\mathrm{RMSE}_{FC}(\Delta t_f)}{\mathrm{RMSE}_{AN}(\Delta t_f)}\right) \times 100 \tag{4.15}$$

A similar skill score can be defined from observations as the reference truth. As observations are not made in the same location, it is necessary for the averaging operator to be performed as part of the sample mean.

$$\mathrm{RMSE}^{FC}_{M,O}(\Delta t_f) = \left\langle \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\left(T^{AN}_{O,n,k}(\Delta t_f) - T^{FC}_{M,n,k}(\Delta t_f)\right)^2} \right\rangle \tag{4.16a}$$

$$\mathrm{RMSE}^{PFC}_{M,O}(\Delta t_f) = \left\langle \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\left(T^{AN}_{O,n,k}(\Delta t_f) - T^{AN}_{M,n,k}(\Delta t_f = \Delta t_0)\right)^2} \right\rangle, \tag{4.16b}$$

where K is the number of observations. The observations can be grouped into bins or larger spatial regions. Another metric that is popular in ocean forecasting is the anomaly cross-correlation (aCC), which can be defined as

$$aCC^{FC}_{M}(\Delta t_f) = \left\langle \frac{\frac{1}{N}\sum_{n=1}^{N}\left(T^{AN}_{M,n}(\Delta t_f) - \overline{T^{AN}_{M}(\Delta t_f)}\right)\frac{1}{N}\sum_{n=1}^{N}\left(T^{FC}_{M,n}(\Delta t_f) - \overline{T^{FC}_{M}(\Delta t_f)}\right)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(T^{AN}_{M,n}(\Delta t_f) - \overline{T^{AN}_{M}(\Delta t_f)}\right)^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(T^{FC}_{M,n}(\Delta t_f) - \overline{T^{FC}_{M}(\Delta t_f)}\right)^2}} \right\rangle.$$

This provides a measure of the correlation between the analysis anomalies and the forecast anomalies. Similarly, an analogous anomaly correlation can be derived based on the observations. One of the advantages in using RMSE is the geometric relationship with the aCC, which can be summarized in a single diagram (Taylor 2001).

d. Ensemble statistics

Ensemble forecasting, though not yet mainstream in ocean forecasting, is the dominant approach available to characterize forecast error as well as estimate the probability density. These methods increase forecast reliability, delivering societal and economic benefit in particular for applications for which there are higher risks or costs. Ensemble forecasting is essential in long-range forecasting e.g., seasonal forecasting, and has become mainstream in medium-range numerical weather prediction. For an overview of the verification methods established in these communities, refer to Weigel (2012).

Extending the earlier estimation to an ensemble, we have

$$T_{M,\gamma} = T_S + \varepsilon_{M,\gamma} \tag{4.17}$$

where we assume that the model bias has been removed as before, $\varepsilon_{M,\gamma} \sim N(0, \sigma^2_{M,\gamma})$ and $\gamma \in [1, \ldots, \Gamma]$ represents the ensemble members, and Γ is the ensemble size. The ensemble mean is then

$$\begin{aligned}\bar{T}_{M,\gamma} &= \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} T_{M,\gamma}, \\ &= T_S + \bar{\varepsilon}_{M,\Gamma}\end{aligned} \tag{4.18}$$

where the overbar represents the ensemble averaging operator. For a "well-behaved" ensemble, $|\bar{\varepsilon}_{M,\Gamma}| \leq |\varepsilon_{M,\gamma}|$, where the ensemble members are independent forecast samples. Applying the sample mean theory $\bar{\varepsilon}_{M,\Gamma}$ and assuming the ensemble is homogeneous, we get $\bar{\varepsilon}_{M,\Gamma} \sim N\left(0, \frac{\sigma_M^2}{\Gamma}\right)$.

Extending Eq. (4.4a) applied to the ensemble mean innovations, we get

$$\begin{aligned}E[(\bar{T}_{M,\Gamma} - T_O)^2] &= E[\bar{\varepsilon}_M^2] + \sigma_O^2 \\ &= \frac{\sigma_M^2}{\Gamma} + \sigma_O^2 \\ &= \sigma_{M,\Gamma}^2 + \sigma_O^2\end{aligned} \tag{4.19}$$

where the model and observation errors are assumed independent.

The ensemble sample variance is given by

$$\begin{aligned}S_{\Gamma-1}^2 &= \frac{1}{\Gamma-1}\sum_{\gamma=1}^{\Gamma}(T_{M,\gamma} - \bar{T}_{M,\Gamma})^2 \\ &= \frac{1}{\Gamma-1}\sum_{\gamma=1}^{\Gamma}(\varepsilon_{M,\gamma} - \bar{\varepsilon}_{M,\Gamma})^2\end{aligned}, \tag{4.20}$$

where the ensemble variance is an unbiased estimator. For a well-behaved ensemble $|\bar{\varepsilon}_M| \to 0$ as Γ increases, which from Eq. (4.18) implies $\bar{T}_{M,\Gamma} \to T_S$ and the ensemble variance converges to the model error variance, $\mathrm{VAR}[T_{M,\gamma}] \to \sigma_M^2$.

For verification of an ensemble, (G. B. Brassington, unpubl. data) proposed constructing the normalized error distribution as

$$\varepsilon_Z = \frac{\bar{T}_{M,\gamma} - T_O}{\sqrt{S_{\Gamma-1}^2[T_{M,\gamma}] + \sigma_O^2}} \tag{4.21}$$

where the normalizing standard deviation is amended to include the observation error variance to match the expected variance of the ensemble mean error, Eq. (4.19). Therefore, our hypothesis is that, for a well-behaved ensemble, where Eq. (4.20) converges to the model error variance, ε_Z will represent a unit-normal distribution $N(0, 1)$.

To demonstrate the importance of this modified inference model, we will use the Bureau of Meteorology's operational ocean forecast system, Ocean Model, Analysis, and Prediction System version 2 (OceanMAPSv2; Brassington et al. 2012). OceanMAPSv2 is a four-cycle system where four independent cycles are performed on four consecutive days, with

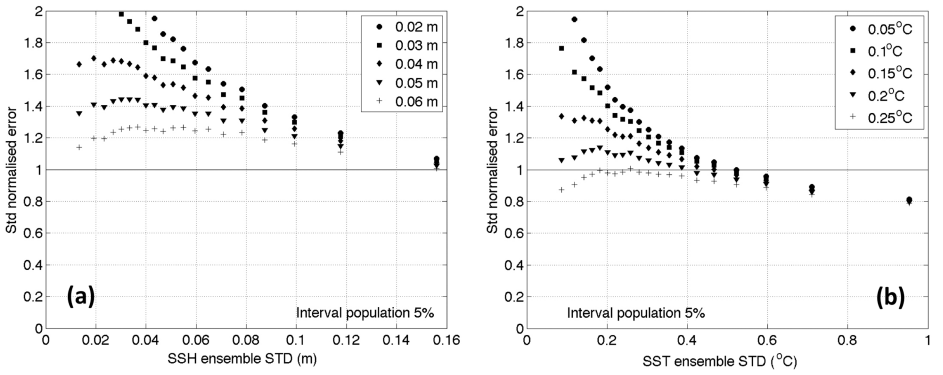

Figure 5. Standard deviation (STD) of the normalized error relative to equal intervals of the ensemble STD. (a) The normalized error is based on the ensemble mean of sea surface height with observation error STD equal to 0.02–0.06 m. (b) The normalized error is based on the ensemble mean of sea surface temperature (SST) with observation error STD equal to 0.05°C to 0.25°C. The normalized error is based on OceanMAPSv2 in 2014 for the Tasman Sea. Ensemble STD intervals are based on equal 5% populations.

each cycle time-lagged and repeated every fourth day. This four-cycle system forms an ensemble that was shown by Brassington (2013) to provide consistent reductions in RMSE and increases in the aCC, defined in a similar way to that in Section 4c. The ensemble mean and variance of the analysis are verified based on Eq. (4.21) using all of the observations of satellite altimetry and satellite SST in the Tasman Sea (150E–160E, 43S–32S) for all of 2014. The sample size of sea surface height anomaly (SSHA) and SST is 148,461 and 203,896, respectively. The Tasman Sea is the region with the highest kinetic energy in the Australian region due to the presence of the East Australian Current. The observation error Eq. (3.1) is composed of instrument error and representation error. The latter is not unique, as it is model and state variance-dependent and needs to be estimated. The total sample population is ordered with respect to the magnitude of the ensemble variance and partitioned into equal intervals of 5%. For each partition, the normalized error distribution in Eq. (4.21) is prepared for a range of observation error variance for SSHA (0.02–0.06 m) and SST (0.05°C to 0.25°C). The standard deviation for each subpopulation is shown in Figure 5a and 5b for SSHA and SST respectively.

For the populations in which the ensemble variance is large, the standard deviation of the population is insensitive to the magnitude of the observation variance. However, as the ensemble variance decreases the error, standard deviation becomes increasingly sensitive to the magnitude of the observation error variance. If the observation error variance is underestimated, the standard deviation of the normalized variance grows without limit, as the variance of the ensemble error converges to the true observation error variance. In the Tasman Sea, we can therefore diagnose that the observation error variance for SSHA is approximately $(0.06\ \text{m})^2$ and for SST it is $(0.25°\text{C})^2$, which are consistent with previous

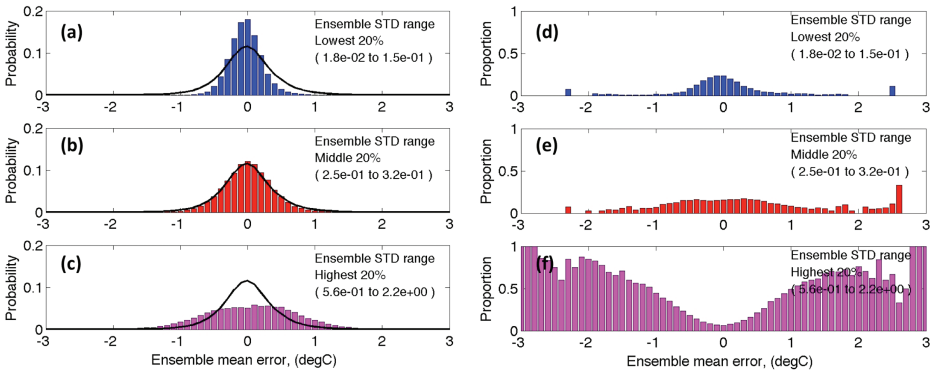

Figure 6. Frequency distribution of sea surface temperature (SST) normalized error for the (a) lowest, (b) middle, and (c) highest 20% ensemble standard deviation (STD) values. The normalized errors are based on the OceanMAPSv2 in the Tasman Sea for 2014. The frequency distribution for the entire population is shown for reference in black. The total SST ensemble mean errors within the subpopulation as a proportion of the total distribution are shown in (d) lowest, (e) middle, and (f) highest 20% ensemble STD values. The ensemble mean errors are based on the OceanMAPSv2 in the Tasman Sea for 2014 against satellite SST.

efforts to estimate them (Oke and Sakov 2008). Once a realistic value is assigned to the observation error variance, the correct estimate of the normalized error variance can be obtained, which is ∼1.2 for SSHA and ∼1.04 for SST, indicating that the ensemble variance is underestimated by 20% and 4%, respectively.

For a well-behaved ensemble, we expect the magnitude of errors to have a relationship with the magnitude of the ensemble STD. The frequency distribution of the ensemble errors for the OceanMAPSv2 system in the Tasman Sea in 2014 is shown as a histogram for three subpopulations based on the magnitude of ensemble STD. The lowest 20%, the middle 20%, and the highest 20% for SST are shown in Figures 6a and 6c, respectively. For reference, the frequency distribution of all ensemble mean errors is shown as a solid line. The four-cycle ensemble shows a consistent relationship in which the lowest 20% corresponds to a relatively higher proportion of small-magnitude errors and a significantly lower proportion of medium to large magnitude errors. The middle 20% population has a proportion of small and large errors that is comparable to the full distribution. The highest 20% show a comparatively low proportion of small errors and a significantly larger proportion of large errors.

To illustrate the relationship more clearly, the subpopulation of ensemble mean errors is re-expressed as a proportion of the total population shown in Figures 6d, 6e, and 6f for SST. The proportion of errors associated with the highest magnitude ensemble STD values explains 50%–100% of the largest magnitude errors in the Tasman Sea. For this subpopulation, the underlying relationship is a monotonically increasing proportion relative to an increase in the magnitude of error. There are, however, clear departures from this relationship that may be explained by a small number of "bad" observations that have larger errors but are able to pass the real-time quality control. The subpopulation (of ensemble

STD) for the lowest 20% (by magnitude) does show an increased proportion of small error values; however, this increase is a relatively modest 20%–30%. The ineffectiveness of the ensemble STD to identify regions of low error is attributed to the loss of precision to detect small ensemble errors due to the observation error variance, Eq. (4.19). As the magnitude of the ensemble STD is reduced (by the selection of the subpopulation), the spread of the ensemble mean innovations asymptotes to the observation error variance. Nonetheless, the largest proportion of small innovations is still associated with the subpopulation of smallest ensemble STD.

5. Progress in verification of ocean forecast systems

Verification is an essential task both in the development and routine operating tasks of an ocean forecast system. The application of verification methods in ocean forecasting has been led by the United States Navy (the first to implement such systems) and Europe with successive MERSEA, MyOcean and Copernicus programs to develop a comprehensive public service. This has formed the basis for international science and expert teams to provide coordination and drive best practice. We outline some of the key stages in these three efforts.

a. U.S. Navy

The U.S. Navy can trace the first acknowledgement of the importance of ocean forecasting to a workshop in 1976 (Burnett et al. 2014). Burnett et al recount that a mandate to develop operational ocean forecasting was initiated in 1986, based on naval research setting out the requirements (e.g., Hurlburt 1984) to deliver the first-generation systems by 1992. This date coincides with the launch of the TOPEX-Poseidon mission initiating the first reference platform (together with Geosat and ERS) providing estimates of the mesoscale sea-level signal to a precision $\sim$5 cm (Fu et al. 1994). Since that time, the U.S. Navy has been at the forefront of the science of operational oceanography, leading developments in ocean analysis (Fox et al. 2002; Cummings 2005) and modeling systems (Wallcraft et al. 2003; Barron et al. 2006; Chassignet et al. 2007). Validating ocean model representation of known physical oceanographic features (i.e., consistency as defined in Section 2) is an ongoing effort that has guided model requirements and development (Chassignet et al. 2000; Hogan and Hurlburt 2000; Hurlburt and Hogan 2000). Implementation of global operational forecast systems are accompanied by a range of common verification metrics (e.g., Smedstad et al. 2003) including RMSE, anomaly correlation, and skill scores together with more detailed regional evaluations (Metzger et al 2010), where observing systems are available.

b. MERSEA and Copernicus Frameworks

Europe has been a driving force in the broader public development of ocean forecasting, providing policies for open data sharing and frameworks for operational ocean services

(Dahlin et al., 2003). A series of European programs including MERSEA, MyOcean, and now Copernicus have established a sophisticated ocean forecast capability from multiple institutions. The first comprehensive framework for model validation and verification was proposed in MERSEA (Crosnier and Provost 2006, 2007). A comprehensive evaluation program was developed leveraging the available observation campaigns in the North Atlantic Ocean and Mediterranean Sea. A series of classes were developed to perform the comparisons: CLASS1 2D and 3D model fields are defined on standard grids to simplify comparison; CLASS2 is defined as sections and fixed vertical profiles that correspond to observed locations; CLASS3 represents integrated quantities such as volume and heat transports, and CLASS4 is defined as model errors based on interpolated model fields to observation time and space position for reference data (MERSEA IP, 2006). CLASSES 1 and 3 provide data for model evaluation or *consistency*, while CLASSES 2 and 4 provide data to assess *quality* and performance. An example of the performance assessment for the Mediterranean is based on CLASS4 comparisons, using a similar set of metrics (Tonani et al. 2009).

c. GODAE, GODAE OceanView and JCOMM ETOOFS

Motivated by the success of the global ocean observing system to implement a range of *in situ* networks, the community outlined the case for the sustainment of these networks as tied to delivery of societal benefits, which included their integration with ocean modeling and forecast services (Smith and Koblinsky, 2001). The concept for a demonstration experiment was first raised by Smith (2000) and has progressed into an international science team (now GODAE OceanView, https://www.godae-oceanview.org) over the subsequent period (Bell et al. 2009, 2015). An important activity within this science team has been the inter-comparison and verification of the global ocean forecast systems. Early developments leveraged significantly from the MERSEA framework (Hernandez et al. 2009, 2015). Metrics that have matured into accepted practice are being documented as Guidance by the JCOMM Expert Team for Operational Ocean Forecast Systems (ETOOFS; see http://jcomm.info). For pragmatic considerations of data sharing, CLASS4 emerged as the priority being a more compact data type and lending itself to the standard verification metrics (Oke et al. 2012; Divakaran et al. 2015; Ryan et al. 2015). As communication and storage capacity permits, other classes will become feasible and will permit the development of multi-model ensemble products.

6. Technology

A common approach to presenting verification statistics is to routinely generate graphics for a set of predefined statistics. For example, a set of verification metrics is published online for global and regional systems within the European Copernicus service (http://marine.copernicus.eu/services-portfolio/scientific-quality/). Each of these graphics

is predefined and routinely updated after each forecast cycle. They are based on global and basin scale summary statistics to monitor day-to-day system performance variations.

A more flexible approach makes use of some of the recent technologies to provide data discovery to the archive via the internet. Ocean forecasting has promoted open data to the observing system and live access to gridded model products through technologies such as OPeNDAP (*www.opendap.org*)/Thredds (www.unidata.ucar.edu/software/thredds/*current/tds)*. With these foundations, verification can exploit this technology to provide access to a wider range of users interested in the conditional performance (e.g., performance in at a specific location and time) of a system rather than summary statistics of the total system or annual or longer time periods. An example is the CLASS4 web-analyzer developed at the Bureau of Meteorology based on open layers (*openlayers.org)*. The web-analyzer provides access to the GOV CLASS4 data archive outlined in Section 5 (http://130.56.244.252/monitoring/index.php?pg=class4). There are four main stages to an interactive verification process: (1) population selection, (2) population analysis, (3) statistical analysis, and (4) graphical presentation, where population represents a subsample of the model and observation pairs. The population selection provides a date slider and a menu options for the period (1 day, 7 day, 30 day, 6 months, 1 year), as shown in Figure 7a. A region can be defined on the map by drawing a polygon with a mouse. The polygon can be edited and saved for future recall from a menu. When the period and polygon are selected, a population analysis is performed on the metadata to define the sample size and the distribution is time and within the polygon, as shown in Figure 7b. The statistic menus provide options to perform a statistical analysis that is computed on the host server. Selections include the CLASS4 *dataset* (Argo, Jason2, and SST), the statistical operation (Mean, STD, RMSE, MAE), and the forecast parameter (Best estimate, Forecasts day 1–5), as shown in Figure 7c. The CLASS4 analyzer is designed for interactive and ad hoc analysis for a variety of applications such as forecast events, system investigations, and discussions as well as planning for exercises and campaigns. For analyses on large sample sizes, the analysis can be performed in batch mode, which computes statistics offline with job management and notification. The data can then be represented with a range of graphical presentation options.

There is a growing range of analysis software and tools available to perform verification analyses which range from commercial packages to institutional supported software as outlined in (Jolliffe and Stephenson 2012, appendix).

7. Summary

Verification within the field of atmospheric and climate science has matured, as evidenced by the range of specialized scientific literature. A majority of the high-level frameworks and metrics are transferable into the ocean sciences, and progress has been made in this regard within national agencies and ocean forecast communities. There are a number of areas of complexity within the atmosphere, such as cloud physics and precipitation, that lead to specialist topics that have no analog in ocean forecasting. There is also verification devoted to

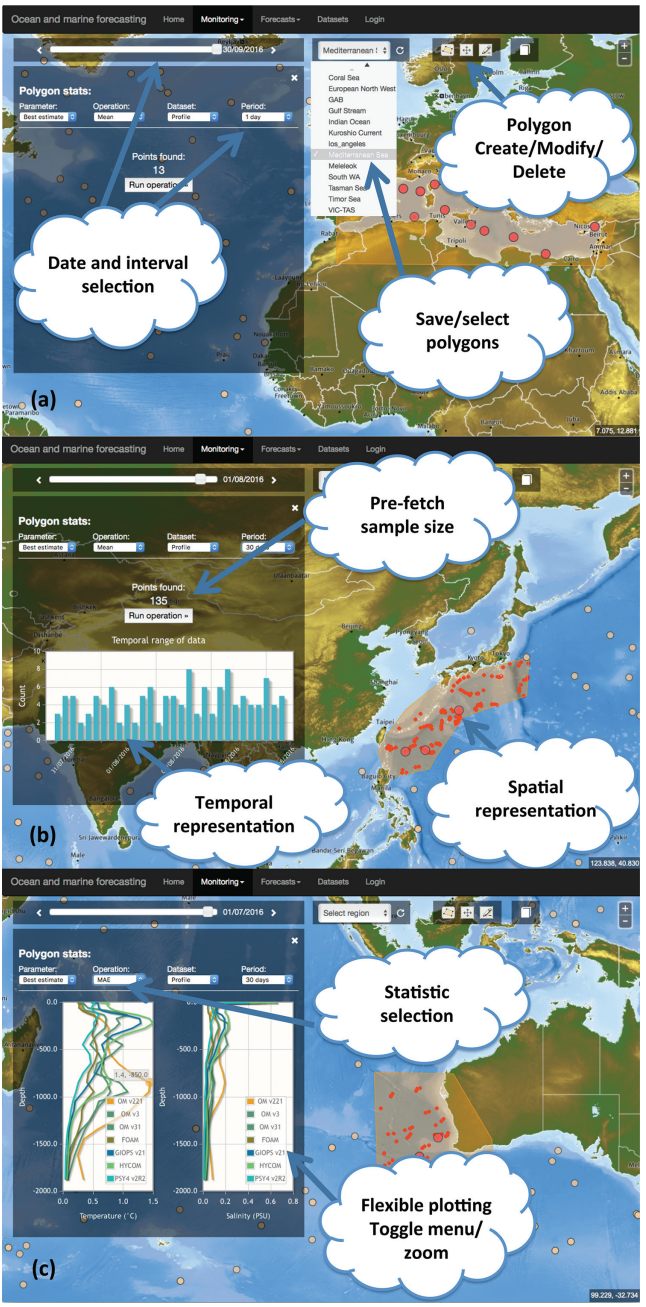

Figure 7. A verification web-tool for conditional statistics on CLASS4 Argo data from an operational global ocean forecasting system (Australia, UK MetOffice, NOAA, Mercator, Environment Canada). (a) Data selection: reference data type (Argo, Altimetry, Buoy SST), start time slider, time interval menu selection, polygon area selection of pre-computed or user-drawn/modified; (b) pre-fetch population information, sample size, temporal frequency distribution, and spatial distribution within the polygon; and (c) statistic selection (Mean, STD, RMSE, MAE), flexible plotting zoom/selection of lines and exporting.

quantifying the added value of an experienced forecaster. General ocean forecasting products are served automatically with limited forecaster guidance. There are exceptions where the risk–benefit is higher, such as in search and rescue, hazardous chemical response, and defense operations, where there is a role for an ocean forecaster. In such cases, techniques are available to assess the added value. On the other hand, a number of important aspects to ocean forecasting have limited analogy in atmospheric physics, namely, western boundary currents, a turbulent interfacial forcing (i.e., the mixed layer is strongly influenced by a chaotic atmosphere, while the ABL is weakly influenced by a chaotic ocean), and continental shelf regions, where tidal amplitudes and shallow water physics dominate. Ocean physics also has its areas of complexity in the role of biogeochemical concentrations in radiation physics, sediment discharge and re-suspension from rivers and on the continental shelf and sea-ice physics.

The ocean physics in terms of the absorption of electromagnetic radiation restricts remote sensing to surface and skin properties. It is fortunate that the vertically integrated mass of the ocean has an expression at the surface, which, since the 1990s, has been observed from satellite altimetry. This provides global maps of sea surface height anomaly, the analog to sea level pressure, from which the majority of the ocean model state is constrained through data assimilation. At present, satellite altimeters use nadir instruments, which provide along-track gradients and repeat orbits of 9.9 days and greater. Satellite sea surface temperature is the most frequently observed variable by multiple satellite with multiple instruments, largely measured in infrared and microwave frequencies. The cross-calibration and quality control of these platforms is a specialist area requiring support from an active science team (www.ghrsst.org). *In situ* networks are challenging to operate in this corrosive and biofouling environment; however, significant success has been achieved in autonomous platforms both profiling floats (Argo and Gliders) and surface observers (drifting buoys). Nonetheless, the spatial resolution observed by these platforms is on the order of degrees and temporal resolution of days. For ocean verification, the representative error from this sparse observing network is an important consideration, as we have shown.

For an ocean forecast system that resolves the mesoscale with spatial and temporal scales of 0.1 degree and 1 hour, a significant proportion of the spectrum is unconstrained. Combining these high wavenumber errors with a chaotic system leads to forecast errors persisting and growing, resulting in forecast uncertainty. Ensemble forecasting, although in its infancy in ocean forecasting, is likely to feature prominently in the future as the best available approach to estimate the probability density. Preliminary poor-man's (an ensemble based on different forecast systems) and multi-cycle ensemble systems (Brassington, 2013), which are computationally efficient, are already showing promise. We have extended the statistical theory to include verification for ensemble forecast systems and added hypothesis testing as a robust metric for quantifying the reliability of ensemble variance matching ensemble mean errors. We also demonstrate the importance of including observing system variance.

The focus of this chapter has concentrated on global ocean forecasting and the verification fundamentals that provide the initial underpinning to all systems in this field. In particular, we have examined the statistical theory behind some of these metrics, notably RMSE and MAE, which are frequently used. Analytical expressions based on (G. B. Brassington, unpubl. data) for both the expectation and variance of the sample RMSE and MAE facilitate interpretation of their behavior. A debate on the relative merits of RMSE and MAE is contemporary with empirical arguments put forward that favor MAE over RMSE (see Willmott and Matsuura 2005) and other papers countering against some of their claims (see Chai and Draxler 2014). In an unbiased or de-biased system, both metrics are estimating the same quantity within a scale factor. The distinction between the two metrics relates to their convergence, uncertainty, and sensitivity to outliers. In general, MAE has the more favorable behavior. However, interpretation of whether the sensitivity to outliers is "good" or "bad" is application-specific. In ocean forecasting, where automatic quality control can lead to a proportion of "bad" observations in the verifying data, MAE should be used in near real-time. However, when delayed mode data is available, then RMSE may be preferred if the behavior of the system at the extremes is important. It is worth noting that MAE could be applied in this case using conditional sampling.

Ocean forecasting has emerged at the same time as the internet, smart device technologies, and data mining. Large data archives of ocean forecasts and verification products can be exploited by these technologies to add significant value to the forecasts for specialist users or downstream service providers. Verification has largely targeted a standard set of predefined metrics that can be repeated and frequently compared. In Section 6 we provide a glimpse of how new technologies can provide a flexible environment to extend verification as a product in its own right, with users able to control the definition of metrics.

Acknowledgments. I would like to acknowledge the Intercomparison and Validation task team under GODAE OceanView and the JCOMM Expert Team for Operational Ocean Forecast Systems who have been leading the international coordination of verification in this field. I would like to thank Dr Beth Ebert and two anonymous referees for their constructive comments.

REFERENCES

Barron, C. N., A. B. Kara, P. J. Martin, R. C. Rhodes, and L. F. Smedstad. 2006. Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). Ocean Model., *11*(3), 347–375. doi: 10.1016/j.ocemod.2005.01.004

Bell, M. J., M. Lefebvre, P.-Y. Le Traon, N. Smith, and K. Wilmer-Becker. 2009. GODAE: The global ocean data assimilation experiment. Oceanography, *22*, 14–21. doi: 10.5670/oceanog.2009.62

Bell, M. J., A. Schiller, P.-Y. Le Traon, N. R. Smith, E. Dombrowsky, and K. Wilmer-Becker. 2015. An introduction to GODAE OceanView. J. Oper. Oceanogr., *8*(suppl 1), s2–s11. doi: 10.1080/1755876X.2015.1022041

Black, P. G., E. A. D'Asaro, W. M. Drennan, and J. R. French. 2007. Air-sea exchange in hurricanes: Synthesis of observations from the Coupled Boundary Layer Air-Sea Transfer Experiment. Bull. Am. Meteorol. Soc., *88*(3), 357–374. doi: 10.1175/BAMS-88-3-357

Brassington, G. B. 2013. Multicycle ensemble forecasting of sea surface temperature, Geophys. Res. Lett., *40*, 6191–6195. doi: 10.1002/2013GL057752

Brassington, G. B. 2017a. Mean absolute error and root mean square error: Which is the better metric of model performance? Mon. Weather Rev., Submitted.

Brassington, G. B. 2017b. Operational four-cycle ensemble ocean forecasting: verifying the ensemble variance, Q. J. R. Meteorol. Soc., Submitted.

Brassington, G. B., and P. Divakaran. 2009. The theoretical impact of remotely sensed sea surface salinity observations in a multi-variate assimilation system. Ocean Model., *27*(1), 70–81. doi: 10.1016/j.ocemod.2008.12.005

Brassington, G. B., J. Freeman, X. Huang, T. Pugh, P. R. Oke, P. A. Sandery, A. Taylor, et al. 2012. Ocean Model, Analysis and Prediction System (OceanMAPS): Version 2. CAWCR Technical Report No. 052. Melbourne, Victoria: Centre for Australian Weather and Climate Research, 110 p.

Brassington, G. B., A. Hines, E. Dombrowsky, S. Ishizaki, F. Bub, and M. Ignaszewski. 2010. Short to medium-range ocean forecasts: Delivery and observational requirements, *in* Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Vol. 1, J. Hall, D. E. Harrison, and D. Stammer, eds. ESA Publication [Hamburg]: [European Space Agency] WPP-306. doi: 10.5270/OceanObs09

Burnett, W., S. Harper, R. Preller, G. Jacobs, and K. LaCroix. 2014. Overview of operational ocean forecasting in the U.S. Navy: Past, present, and future. Oceanography, *27*, 24–31. doi: 10.5670/oceanog.2014.65

Chai, T., and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev., *7*, 1247–1250. doi: 10.5194/gmd-7-1247-2014

Chassignet, E. P., H. Arango, D. Dietrich, T. Ezer, M. Ghil, D. B. Haidvogel, C.-C. Ma, et al. 2000. DAMEE-NAB: The base experiments. Dyn. Atmos. Oceans, *32*, 155–183. doi: 10.1016/S0377-0265(00)00046-4

Chassignet, E. P., H. E. Hurlburt, O. M. Smedstad, G. R. Halliwell, P. J. Hogan, A. J. Wallcraft, R. Baraille, et al. 2007. The HYCOM (hybrid coordinate ocean model) data assimilative system. J. Mar. Syst., *65*, 60–83. doi: 10.1016/j.jmarsys.2005.09.016

Crosnier, L., and C. Le Provost. 2006. Internal metrics definition for operational forecast systems intercomparison: Example in the North Atlantic and Mediterranean Sea, *in* Ocean Weather Forecasting, J. Verron and E. P. Chassignet, eds. Rotterdam: Springer, 455–465. doi: 10.1007/1-4020-4028-8

Crosnier, L., and C. Le Provost. 2007. Inter-comparing five forecast operational systems in the North Atlantic and Mediterranean basins: The MERSEA-strand1 methodology. J. Mar. Syst., *65*(1), 354–375. doi: 10.1016/j.jmarsys.2005.01.003

Cummings, J. A. 2005. Operational multivariate ocean data assimilation. Q. J. R. Meteorol. Soc., *131*(613), 3583–3604. doi: 10.1256/qj.05.105

Cummings, J., G. Brassington, R. Keeley, M. Martin, and T. Carval. 2010. GODAE ocean data quality control intercomparison project, *in* Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Vol. 1, J. Hall, D. E. Harrison and D. Stammer, eds. ESA Publication [Hamburg]: [European Space Agency] WPP-306. doi: 10.5270/OceanObs09

Dahlin, H., N. C. Flemming, K. Nittis, and S. E. Petersson, eds. 2003. Building the European Capacity in Operational Oceanography: Proceedings 3rd EuroGOOS Conference, Vol. 69. New York: Elsevier Science. (Elsevier Oceanography Series).

Daley, R. 1993. Estimating observation error statistics for atmospheric data assimilation. Ann. Geophysicae, *11*, 634–647.

Divakaran, P., G. B. Brassington, A. G. Ryan, C. Regnier, T. Spindler, A. Mehra, F. Hernandez, et al. 2015. GODAE OceanView Class-4 inter-comparison for the Australian Region. J. Oper. Oceanogr., *8*, 115–128. doi: 10.1080/1755876X.2015.1022333

Dombrowsky, E., L. Bertino, G. B. Brassington, E. P. Chassignet, F. Davidson, H. E. Hurlburt, M. Kamachi, et al. 2009. GODAE systems in operation. Oceanography, *22*(3), 80–95. doi: 10.5670/oceanog.2009.68

Donlon, C. J., P. J. Minnett, C. Gentemann, T. J. Nightingale, I. J. Barton, B. Ward, and M. J. Murray. 2002. Toward improved validation of satellite sea surface skin temperature measurements for climate research. J. Clim., *15*(4), 353–369.

Edson, J., T. Crawford, J. Crescenti, T. Farrar, N. Frew, G. Gerbi, A. Plueddemann, et al. 2007. The coupled boundary layers and air–sea transfer experiment in low winds. Bull. Am. Meteorol. Soc., *88*(3), 341–356. doi: 10.7916/D8GF0TCQ

Font, J., G. Lagerloef, D. LeVine, A. Camps, and O. Z Zanife. 2004. The determination of surface salinity with the European SMOS space mission. IEEE Trans. Geosci. Remote. Sens., *42*(10), 2196–2205. doi: 10.1109/TGRS.2004.834649

Fox, D. N., W. J. Teague, C. N. Barron, M. R. Carnes, and C. M. Lee. 2002. The Modular Ocean Data Assimilation System (MODAS). J. Atmos. Ocean. Technol., *19*(2), 240–252.

Fu, L. L., E. J. Christensen, C. A. Yamarone, M. Lefebvre, Y. Menard, M. Dorrer, and P. Escudier.1994. TOPEX/Poseidon mission overview. J. Geophys. Res. Oceans, *99*(C12), 24369–24381. doi: 10.1029/94JC01761

Fu, L. L., D. Stammer, R. R. Leben, and D. B. Chelton. 2003. Improved spatial resolution of ocean surface topography from the T/P-Jason-1 altimeter mission. Eos Trans. AGU, *84*(26), 241–248. doi: 10.1029/2003EO260002

Ghelli, A., and E. Ebert. 2008. Special issue on forecast verification. Meteorol. Appl., *15*(1), 1. doi: 10.1002/met.69

Gill, A. E. 1982. Atmosphere-Ocean Dynamics. New York: Academic Press, 662 pp.

Gould, J., D. Roemmich, S. Wijffels, H. Freeland, M. Ignaszewsky, X. Jianping, S. Pouliquen, et al. 2004. Argo profiling floats bring new era of in situ ocean observations. Eos Trans. AGU, *85*(19), 179–191. doi: 10.1029/2004EO190002

Hall, J., D. E. Harrison, and D. Stammer, eds. 2010. Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Venice, Italy, 21–25 September 2009, ESA Publication [Hamburg]: [European Space Agency] WPP-306. doi: 10.5270/OceanObs09

Hernandez, F., L. Bertino, G. B. Brassington, E. Chassignet, J. Cummings, F. Davidson, M. Drévillon, et al. 2009. Validation and intercomparison studies within GODAE. Oceanography, *22*(3), 128–143. doi: 10.5670/oceanog.2009.71

Hernandez, F., E. Blockley, G. B. Brassington, F. Davidson, P. Divakaran, M. Drévillon, S. Ishizaki, et al. 2015. Performance evaluations, near real-time assessment of operational oceanography forecast products. J. Oper. Oceanogr., *8*, s221–s238. doi: 10.1080/1755876X.2015.1050282

Hogan, P. J., and H. E. Hurlburt. 2000. Impact of upper ocean-topographical coupling and isopycnal outcropping in Japan/East Sea models with 1/8° to 1/64° resolution. J. Phys. Oceanogr., *30*(10), 2535–2561.

Hurlburt, H. E. 1984. The potential for ocean prediction and the role of altimeter data. Mar. Geod., *8*, 17–66. doi: 10.1080/15210608409379497

Hurlburt, H. E., and P. J. Hogan. 2000. Impact of 1/8 to 1/64 resolution on Gulf Stream model–data comparisons in basin-scale subtropical Atlantic Ocean models. Dyn. Atmos. Oceans, *32*(3), 283–329. doi: 10.1016/S0377-0265(00)00050-6

Ingleby, B., and M. Huddleston. 2007. Quality control of ocean temperature and salinity profiles—Historical and real-time data. J. Mar. Syst., *65*(1), 158–175. doi: 10.1016/j.jmarsys.2005.11.019

Jolliffe, I. T., and D. B. Stephenson. 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd ed. New York: John Wiley and Sons, 274 pp.

Koblinsky, C. J., P. Hildebrand, D. LeVine, F. Pellerano, Y. Chao, W. Wilson, S. Yueh, and G. Lagerloef. 2003. Sea surface salinity from space: Science goals and measurement approach. Radio Sci., *38*(4), 8064. doi: 10.1029/2001RS002584

Large, W. G., and S. Pond. 1982. Sensible and latent heat flux measurements over the ocean. J. Phys. Oceanogr., *12*(5), 464–482. doi: 10.1175/1520-0485(1982)012<0464:SALHFM>2.0.CO;2

Large, W. G., G. Danabasoglu, S. C. Doney, and J. C. McWilliams. 1997. Sensitivity to surface forcing and boundary layer mixing in a global ocean model: Annual-mean climatology. J. Phys. Oceanogr., *27*(11), 2418–2447. doi: 10.1175/1520-0485(1997)027<2418:STSFAB>2.0.CO;2

Leone, F. C., L. S. Nelson, and R. B. Nottingham. 1961. The folded normal distribution. Technometrics, *3*(4), 543–550. doi: 10.2307/1266560

MERSEA IP (Marine EnviRonment and Security for the European Area – Integrated Project). 2006. List of internal metrics for the MERSEA-GODAE Global Ocean: Specification for implementation. MERSEA-WP05-MERCA-STR-0015-01C. Mercator Ocean, France.

Metzger, E. J., H. E. Hurlburt, X. Xu, J. F. Shriver, A. L. Gordon, J. Sprintall, R. D. Susanto, et al. 2010. Simulated and observed circulation in the Indonesian Seas: 1/12 global HYCOM and the INSTANT observations. Dyn. Atmos. Oceans, *50*(2), 275–300. doi: 10.1016/j.dynatmoce.2010.04.002

Moon, I. J., I. Ginis, T. Hara, and B. Thomas. 2007. A physics-based parameterization of air-sea momentum flux at high wind speeds and its impact on hurricane intensity predictions. Mon. Weather Rev., *135*(8), 2869–2878. doi: 10.1175/MWR3432.1

Murphy, A. H. 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev., *116*, 2417–2424. doi: 10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A. H. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecast., *8*, 281–293.

Murphy, A. H., and R. L. Winkler. 1987. A general framework for forecast verification. Mon. Weather Rev., *115*(7), 1330-1338. doi: 10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

Ohlmann, J. C., D. A. Siegel, and C. Gautier. 1996. Ocean mixed layer radiant heating and solar penetration: A global analysis. J. Clim., *9*(10), 2265–2280. doi: 10.1175/1520-0442(1996)009<2265:OMLRHA>2.0.CO;2

Oke, P. R., and P. Sakov. 2008. Representation error of oceanic observations for data assimilation. J. Atmos. Ocean Technol., *25*, 1004–1017. doi: 10.1175/2007JTECHO558.1

Oke, P. R., G. B. Brassington, J. Cummings, M. Martin, and F. Hernandez. 2012. GODAE inter-comparisons in the Tasman and Coral Seas. J. Oper. Oceanogr., *5*(2), 11–24. doi: 10.1080/1755876X.2012.11020135

Palmer, T., and R. Hagedorn, eds. 2006. Predictability of Weather and Climate. Cambridge: Cambridge University Press. 702 p.

Pascual, A., Y. Faugère, G. Larnicol, and P.-Y. Le Traon. 2006. Improved description of the ocean mesoscale variability by combining four satellite altimeters. Geophys. Res. Lett., *33*(2), L02611. doi: 10.1029/2005GL024633

Powell, M. D., P. J. Vickery, and T. A. Reinhold. 2003. Reduced drag coefficient for high wind speeds in tropical cyclones. Nature, *422*(6929), 279–283. doi: 10.1038/nature01481

Ryan, A. G., C. Regnier, P. Divakaran, T. Spindler, A. Mehra, G. C. Smith, F. Davidson, et al. 2015. GODAE OceanView Class 4 forecast verification framework: Global ocean inter-comparison. J. Oper. Oceanogr., *8*(suppl 1), S98–S111. doi: 10.1080/1755876X.2015.1022330

Smedstad, O. M., H. E. Hurlburt, E. J. Metzger, R. C. Rhodes, J. F. Shriver, A. J. Wallcraft, and A. Birol Kara. 2003. An operational eddy resolving 1/16 global ocean nowcast/forecast system. J. Mar. Syst., *40*, 341–361. doi: 10.1016/S0924-7963(03)00024-1

Smith, N. R., and C. J. Koblinsky, eds. 2001. The ocean observing system for the 21st Century: A consensus statement, *in* Observing the Oceans in the 21st Century. Melbourne, Australia: GODAE Project Office, Bureau of Meteorology, 1–25.

Smith, N. R. 2000. The Global Ocean Data Assimilation Experiment. Adv. Space Res., *25*, 1089–1098. doi: 10.1016/S0273-1177(99)00868-6

Spindler, T., A. Mehra, and D. Wilson-Diaz. 2016. Applying multi-model superensemble methods to global ocean operational systems [Abstract], *in* GODAE OceanView IV/TT Workshop. Montreal, September 20–22, 2016.

Stanski, H. R., L. J. Wilson, and W. R. Burrows. 1989. Survey of Common Verification Methods in Meteorology. Geneva: World Meteorological Organization, 114 p.

Stommel, H. 1948. The westward intensification of wind-driven ocean currents. Eos Trans. AGU, *29*(2), 202–206. doi: 10.1029/TR029i002p00202

Taylor, K. E. 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Oceans, *106*(D7), 7183–7192. doi:10.1029/2000JD900719

Tonani, M., N. Pinardi, C. Fratianni, J. Pistoia, S. Dobricic, S. Pensieri, et al. 2009. Mediterranean Forecasting System: Forecast and analysis assessment through skill scores. Ocean Sci., *5*, 649–660.

Wallcraft, A. J., A. B. Kara, H. E. Hurlburt, and P. A. Rochford. 2003. The NRL Layered Global Ocean Model (NLOM) with an Embedded Mixed Layer Submodel: Formulation and tuning. J. Atmos. Ocean. Technol., *20*(11), 1601–1615. doi: 10.1175/1520-0426(2003)020<1601:TNLGOM>2.0.CO;2

WCRP-JWG. 2016. Joint Working Group on Forecast Verification Research. Bureau of Meteorology, Last accessed 26 July 2017, Publisher, Bureau of Meteorology, Melbourne. http://www.cawcr.gov.au/projects/verification/

Weigel, A. P. 2012. Ensemble forecasts, *in* Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd ed., I. T. Jolliffe and D. B. Stephenson, eds. New York: John Wiley and Sons, 141–166.

Wilks, D. S. 2011. Statistical Methods in the Atmospheric Sciences, Vol. 100. New York: Academic Press, 704 p.

Willmott, C. J., and K. Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res., *30*(1), 79–82. doi: 10.3354/cr030079

Winkler, R. L., and A. H. Murphy. 1968. "Good" probability assessors. J. Appl. Meteorol., *7*(5), 751–758.

Wunsch, C. 2005. The total meridional heat flux and its oceanic and atmospheric partition. J. Climate, *18*(21), 4374. doi: 10.1175/JCLI3539.1

Received: *9 May 2016*; revised: *28 February 2017*.

Editor's note: Contributions to *The Sea: The Science of Ocean Prediction* are being published separately in special issues of *Journal of Marine Research* and will be made available in a forthcoming supplement as Volume 17 of the series.