**ENSO Prediction in Project Minerva: Sensitivity to Atmospheric Horizontal Resolution and Ensemble Size**


* Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia
+ Center for Ocean–Land–Atmosphere Studies, and Department of Atmospheric, Oceanic, and Earth Sciences, College of Science, George Mason University, Fairfax, Virginia
# European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 14 April 2014, in final form 21 November 2014)

**ABSTRACT**

This study examines El Niño–Southern Oscillation (ENSO) prediction in Project Minerva, a recent collaboration between the Center for Ocean–Land–Atmosphere Studies (COLA) and the European Centre for Medium-Range Weather Forecasts (ECMWF). The focus is primarily on the impact of the atmospheric horizontal resolution on ENSO prediction, but the effect from different ensemble sizes is also discussed. Particularly, three sets of 7-month hindcasts performed with ECMWF prediction system are compared, starting from 1 May (1 November) during 1982–2011 (1982–2010): spectral T319 atmospheric resolution with 15 ensembles, spectral T639 with 15 ensembles, and spectral T319 with 51 ensembles. The analysis herein shows that simply increasing either ensemble size from 15 to 51 or atmospheric horizontal resolution from T319 to T639 does not necessarily lead to major improvement in the ENSO prediction skill with current climate models. For deterministic prediction skill metrics, the three sets of predictions do not produce a significant difference in either anomaly correlation or root-mean-square error (RMSE). For probabilistic metrics, the increased atmospheric horizontal resolution generates larger ensemble spread, and thus increases the ratio between the intraensemble spread and RMSE. However, there is little change in the categorical distributions of predicted SST anomalies, and consequently there is little difference among the three sets of hindcasts in terms of probabilistic metrics or prediction reliability.

1. Introduction

In a recent collaborative project between the Center for Ocean–Land–Atmosphere Studies (COLA) and the European Centre for Medium-Range Weather Forecasts (ECMWF), designated as Project Minerva, the ECMWF coupled climate prediction system was used to conduct retrospective seasonal predictions at three different resolutions and two ensemble sizes. Atmospheric spectral horizontal resolutions ranged from T319 (~62 km) to T639 (~31 km) and T1279 (~16 km). The lowest resolution (T319) is higher than those of current operational seasonal forecast models while the highest (T1279) is close to the limit under the hydrostatic assumption and with parameterized convection. Starting from observed 1 May and 1 November initial states for 1980–2013, 7-month ensemble hindcast experiments (some extending to 24 months) were produced for all three resolutions, with either 51 or 15 ensemble members. Project Minerva is an extension of the recently completed multi-institutional Project Athena studying the impacts of high-resolution atmospheric modeling, including cloud resolving, on global climate simulation (Kinter et al. 2013). The current project examines the effects of atmospheric resolution on coupled ensemble seasonal prediction.

In this paper, we evaluate the predictions of the El Niño–Southern Oscillation (ENSO) in Project Minerva. ENSO in the tropical Pacific is the most important climate variation on seasonal-to-interannual time scales, which can have profound effects on weather and climate worldwide. Since its first dynamic prediction conducted around three decades ago (Cane et al. 1986), the ability of dynamical models to predict ENSO has improved significantly (e.g.,

**DOI:** 10.1175/JCLI-D-14-00302.1

© 2015 American Meteorological Society

However, ENSO prediction is still far from perfect in terms of both deterministic (Jin et al. 2008) and probabilistic metrics (Zhu et al. 2013a). A major part of the forecast error is due to inaccuracies and deficiencies in model dynamics and physics that cause systematic errors in model mean state and ENSO evolution. Some of these features are likely sensitive to atmospheric resolution, such as the intertropical convergence zone (ITCZ) and coastal winds. In fact, even though there are only limited investigations examining the effect of atmospheric horizontal resolutions in coupled models, some indeed found beneficial impact of atmospheric horizontal resolutions on simulated ENSO (Guilyardi et al. 2004; Navarra et al. 2008). Guilyardi et al. (2004) used a systematic modular approach to investigate the respective roles of the ocean and atmosphere in setting ENSO characteristics in an ensemble of CGCMs. Their results identified a dominant role of the atmosphere model in determining El Niño characteristics (periodicity and base amplitude) and errors, and a considerable improvement of simulated El Niño power spectra when the atmosphere resolution is significantly increased. Navarra et al. (2008) investigated the effect of atmospheric horizontal resolution on tropical variability using the modified Scale Interaction Experiment (SINTEX) coupled model, and found that higher resolution was generally beneficial even if it did not eliminate all the major systematic errors of the coupled model. Using a slightly older version of SINTEX, Gualdi et al. (2005) also explored the sensitivity of ENSO forecasts to atmospheric horizontal resolution based on relatively limited cases. Their experiments indicated that higher atmospheric model resolution improves the quality of forecasts both for El Niño and La Niña cases, especially during their developing phase. It is noted that the highest resolution in these experiments was T106, clearly lower than the Minerva runs. One focus of this study is to explore the effects of resolving fine structures of atmospheric dynamics and physical processes on ENSO predictions.

Another focus of this paper is on more subtle errors in ENSO predictions. As an example, current ensemble seasonal forecasts at operational centers are generally found to have an apparent “overconfidence” problem, namely that the ensemble perturbations have limited growth relative to the amplitude of mean error (Palmer et al. 2004; Vialard et al. 2005; Saha et al. 2006; Weisheimer et al. 2009; Zhu et al. 2013a). As a result, events (e.g., warm ENSO events) occur more frequently in the ensemble forecasts than the fraction of times such events are observed (e.g., Weigel et al. 2009; Langford and Hendon 2013). This means that the low ensemble spread underestimates the forecast uncertainty and makes it less

---

**Fig. 1.** Distribution of SST mean biases (°C) at 7-month lead time for (a),(d) T319_M15, (b),(e) T639_M51, and (c),(f) the difference between T639_M15 and T319_M15, with ICs in (a)–(c) November during 1982–2010 and (d)–(f) May during 1982–2011.
reliable. The lack of reliability may seriously affect subsequent applications of the forecasts.

Since the relative contributions to such forecast errors from forecast model and initial condition remains unknown, many approaches have been proposed to alleviate the overconfidence problem by sampling different error sources inherent in current forecast systems. For instance, the so-called stochastic physics (Vialard et al. 2005; Molteni et al. 2011) and the multimodel ensemble (MME; Palmer et al. 2004; Weisheimer et al. 2009; Kirtman and Min 2009) approach were applied to represent model uncertainties, which can clearly increase the spread in ENSO forecasts. The traditional lagged ensemble (LE) approach, as used in the National Centers for Environmental Prediction (NCEP) Climate Forecast System (Saha et al. 2014), has the potential of sampling the different phases of high-frequency phenomena such as tropical instability waves (TIW) or the Madden–Julian oscillation (MJO), which may have strong effect on ENSO prediction (Wang et al. 2011). In addition to including the above-mentioned stochastic physics in atmosphere, operational climate predictions at ECMWF also attempts to better represent inherent oceanic perturbations in initial states. This is accomplished by producing an ensemble of ocean data assimilation analyses with five perturbed ocean states (generated by random perturbations driven by sampling uncertainty in winds and in deep ocean initial conditions, and subsampling observation coverage) (Balmaseda et al. 2013). The oceanic ensemble is further augmented by applying SST perturbations, with an associated subsurface temperature signal (Molteni et al. 2011). There are also other procedures of oceanic ensemble generation. For instance, Zhu et al. (2012, 2013a,b) developed a multiocean analysis ensemble (MAE) initialization strategy to better sample the structural uncertainty associated with different ocean data assimilation systems, which have been proven to effectively improve the ensemble spreads and reliability in ENSO prediction (Zhu et al. 2013a).

In this study, we examine the ENSO prediction skills using both deterministic and probabilistic metrics, taking advantage of the combination of multiple resolutions and large ensemble size in Project Minerva. Such a combined analysis is useful because, as a pragmatic matter, a balance has to be struck based on cost and benefit because of the limited computational resource. It should also be noted that higher atmospheric resolution is able to resolve finer atmospheric structures, which would bring extra perturbations on the underlying SST. Such perturbations may be another important uncertainty source for predicting SST.

![Distribution of the STD of SST anomalies (°C) at (a) 3-, (b) 5-, and (c) 7-month lead times with IC in May of 1982–2011. Shown are the STD of observed SST anomalies (contours) and the difference between the STD of the model and observed SST anomalies (shading). The results are shown for (top) T319_M15, (middle) T319_M51, and (bottom) T639_M15.](Figure)
Thus, predictions with a higher atmospheric resolution are expected to have a better SST distribution, which may potentially alleviate the overconfidence problem.

The paper is arranged as follows. The CGCM, the experimental design and datasets are described in the next section. The results are presented in section 3. A summary and discussion are given in section 4.

2. Model, hindcast experiments, and datasets

Project Minerva is an extension of Project Athena (Kinter et al. 2013), but in a coupled framework by using a state-of-the-art coupled operational long-range prediction system. The prediction system used is similar to ECMWF Seasonal Forecast System 4 (simply System 4 hereinafter; Molteni et al. 2011), in terms of ocean model, coupling, initialization, and ensemble perturbation generation methods. The major difference lies in the atmospheric component. In System 4, the ECMWF Integrated Forecast System (IFS) cycle 34r4 with a horizontal resolution of T255 was used, while ECMWF IFS cycle 38r1 with three different horizontal resolutions (T319, T639, and T1279) was applied in Project Minerva. The ocean component uses Nucleus for European Modelling of the

![Figure 3](image-url)
Ocean (NEMO; Madec 2008), version 3.0. Its grid configuration adopts the ORCA1 grid (http://www.noc.soton.ac.uk/nemo/), which has a horizontal resolution of approximately 1° (meridionally refined to 1/3° near the equator), and 42 levels in the vertical, 18 of which are in the upper 200 m. IFS and NEMO are coupled every 3 h. More details about the prediction system, particularly the ensemble perturbation generation, can be found in Molteni et al. (2011).

In Project Minerva, a set of hindcasts was conducted with three different atmospheric horizontal resolutions, and all hindcasts use the same ocean initial conditions (Balmaseda et al. 2013). For the T319 (T639) resolution, 7-month hindcasts with 51 (15) ensemble members started from 1 May and 1 November observed initial conditions (ICs) during 1980–2011 (1980–2010 for the 1 November starts), and an extension to 2013 is being conducted. The T1279 hindcast, however, only went back to 2000 because of the limit of computing resources, and are in progress back to 1990. In addition, some of the hindcasts were also extended from 7 months to 2 years, such as the May starts in the T319 hindcasts and November starts in the T639 hindcasts, both having 15 ensemble members. These 2-yr hindcasts can be used to explore the possibility of extended ENSO predictions, which, however, is beyond the scope of this study.

In this study, our analyses focus on the 7-month hindcasts with T319 and T639 resolutions, which have the largest hindcast samples. The two experiments will be
referred to as T319_M51 and T639_M15, by also identifying their different ensemble sizes. To highlight the difference in atmospheric horizontal resolution, a third group of hindcasts (referred to as T319_M15) is constructed, which is a subset of T319_M51 by using its first 15 ensemble members only. By comparing T319_M51 and T319_M15, we can explore whether detectable gains occur in ENSO prediction as the ensemble size increases from 15 to 51. The comparison will also answer the question of how robust the T639_M15 results are, considering its ensemble size of 15. Finally the T1279 hindcasts (referred to as T1279_M15) will also be briefly examined.

The predicted SST anomalies (SSTAs) in all experiments are derived by subtracting a time- and lead-dependent climatology from the total SSTs. In this paper, predictions for the first month will be defined as 1-month lead, following Molteni et al. (2011). The observation-based monthly SST analysis used for validation is from the optimum interpolation analysis, version 2 (OIv2) SST dataset (Reynolds et al. 2002), which has a resolution of 1.0° × 1.0°. Because OIv2 SST is available since late 1981, predictions since 1982 are examined.

### 3. Results

Before showing SST predictive skills, the global SST mean biases in the hindcasts are examined. Figures 1a, 1b, 1d, and 1e present the global biases relative to the OIv2 SST climatology in hindcasts T319_M15 and T639_M15 starting from both November and May at the 7-month lead time. The longest lead time is 7 months, and it is expected to have the largest biases. For both November and May ICs, the global SSTs in three hindcasts are generally too cold, except some warm biases in the southeastern tropical Atlantic Ocean, and parts of the subpolar North Atlantic and the Southern Ocean. There are cold and warm biases in the western boundary current regions, which are likely related to the poor representation of the paths of the Gulf Stream and the Kuroshio caused by low ocean model resolution. Over the tropical Pacific (the focus of this study), the cold bias in the cold tongue region is larger in November starts, reaching approximately −5°C, but is significant in the hindcasts for both ICs. There is also a slight warm bias in the southeastern tropical Pacific Ocean. Generally, the above SST bias pattern is common in the ECMWF coupled model (Magnusson et al. 2013). When ensemble
size increases from 15 to 51 (not shown), the above biases are essentially unchanged. This result is not surprising, since 30 (29) hindcast cases for May (November) ICs with 15 ensemble members already construct a large enough sample for deriving a reliable SST climatology. On the other hand, when atmospheric resolution increases from T319 (Figs. 1a,d) to T639 (Figs. 1c,f), even though the SST bias pattern looks unchanged, there are some regions with reduced biases. To highlight this difference, Figs. 1c and 1f show the difference between T319_M15 and T639_M15. It can be seen that the cold biases in the subtropical oceans and the Pacific cold tongue region are decreased by up to 0.4°C. The warm bias in the southeastern tropical Pacific Ocean is also slightly reduced.

Figure 2 presents the standard deviation (STD) of SSTA, which is used to measure interannual variability of SSTA. Comparing with the STD of observations (contours), the difference between the STD of model and observations is much smaller, suggesting that the model predicted well the overall observed STD pattern. However, at 3-month lead time, the model underestimates the STD in the far eastern tropical Pacific and overestimates the STD between 150° and 100°W. As the lead time increases, the STD decreases in the whole central-eastern tropical Pacific in comparison with their respective observations, which are consistently smaller than observations in the region at 7-month lead time. Further, comparing the T319_M15, T319_M51, and T639_M15 hindcasts, they present almost identical STD distributions at all three lead times.

As representative metrics of the overall hindcast performance, Figs. 3a and 3b show the anomaly correlation and root-mean-square error (RMSE) between the observed and predicted Niño-3.4 (5°S–5°N, 170°–120°W) SSTA time series as a function of lead month. In terms of anomaly correlation, the differences between the three hindcasts are insignificant but all show good prediction skills. These forecasts can in fact be categorized among the best in comparison with other ENSO prediction systems (Jin et al. 2008; Zhu et al. 2012). Specifically, for the November hindcast, correlation skills remain above 0.85 at 1–6-month leads, but drop dramatically by more than 0.1 at a 7-month lead (corresponding to May, a spring month), suggesting that the model has encountered the spring barrier. For the May experiments, all three hindcasts have correlation skills above 0.8 at almost all lead times. When examined in detail, at short lead times model skills drop more quickly than in the November cases, which may also be a result of...
of the spring barrier and the fact that summer is the season with the largest error growth rate; however, when approaching the ENSO peak phase (the winter season), the prediction skill clearly starts to recover. In terms of RMSE, results are generally insensitive to increasing ensemble size from 15 to 51. On the other hand, there is a nonnegligible difference in RMSEs when atmospheric resolution increases from T319 to T639. In contrast to expectations from previous results, however, higher resolution seems to degrade the RMSE skill, especially for the November start in which there is RMSE difference of \(0.05-0.1^\circ\text{C}\) at long lead times (roughly equivalent to 10%-20% of RMSEs themselves).

Figure 3c presents the temporal evolution of intra-ensemble spreads for the predicted Niño-3.4 index. As expected, in all hindcasts the spreads grow as a function of lead time, as perturbations grow. More interestingly, while T319 hindcasts with either 15 or 51 members have spreads with similar magnitudes for both November and May ICs, T639 hindcasts have clearly larger spreads (especially at long lead times). Even though the physics behind this change is not clearly understood, it is possibly related to the different levels of weather noise. Specifically, in the T639 hindcasts the weather noise may be more energetic with the higher atmospheric resolution (Lindzen and Fox-Rabinovitz 1989) and, accordingly, the underlying SST is more strongly disturbed, thus generating larger spreads.

Figure 3d presents the temporal evolution of the ensemble spread versus ensemble-mean RMSE (SvR) ratio. As is well known, ensemble forecasting requires reliable probabilistic features (i.e., forecast probabilities that match the observed relative frequencies), and one of its necessary conditions is that the ensemble spread should be comparable to RMSE (Johnson and Bowler 2009). From Fig. 3d, it can be seen that the ensemble spread of Niño-3.4 index is generally smaller than the RMSE (especially at short lead times), which happens in almost all single-model ensembles (Palmer et al. 2004; Vialard et al. 2005; Saha et al. 2006; Weisheimer et al. 2009; Zhu et al. 2013a, among others). This suggests that the hindcasts also tend to be “overconfident,” even though the new ECMWF seasonal forecast system (System 4) exhibited increased reliability (Molteni et al. 2011). In particular, the SvR ratio is around 0.2-0.6 at 1-3-month lead, in contrast to CFSv2 in which the ratio is about 0.5-0.7 at the same lead time (Xue et al. 2013; Zhu et al. 2013a). In connection with our previous study (Zhu et al. 2013a), the problem of limited perturbations at short lead times could be alleviated by combining with other ensemble generation methods, such as lagged ensemble (Saha et al. 2006; Xue et al. 2013) or MAE...
On the other hand, as a result of increased ensemble spreads in the T639 hindcasts at long lead times (Fig. 3c), its SvR ratio is increased accordingly, with ratios even larger than 1.0 at 5–7-month (7 month) lead times in the hindcasts of November (May) initial condition.

Although the Niño-3.4 index is good for representing the canonical El Niño (Trenberth 1997), it is not appropriate for denoting a newly identified El Niño type, in which the SST warming occurs primarily in the central, as opposed to eastern equatorial Pacific Ocean. This type of El Niño has been referred to by various names in literature, including date line El Niño (Larkin and Harrison 2005), El Niño Modoki (Ashok et al. 2007; Weng et al. 2007), warm pool (WP) El Niño (Kug et al. 2009), and central Pacific (CP) El Niño (Yeh et al. 2009; Kao and Yu 2009; Yu and Kim 2010). In this study, the prediction of this El Niño type in Project Minerva is also briefly evaluated by examining the El Niño Modoki index (EMI) (Ashok et al. 2007; Weng et al. 2007). EMI is defined as $(SSTA)_C - 0.5(SSTA)_E - 0.5(SSTA)_W$, where $(SSTA)_C$, $(SSTA)_E$, and $(SSTA)_W$ stand for the area-mean SSTA over the central (subscript $C$; 10°S–10°N, 165°E–140°W), eastern (subscript $E$; 15°S–5°N, 110°–70°W), and western (subscript $W$; 10°S–20°N, 125°–145°E) Pacific, respectively. In terms of the anomaly correlation metric, the models clearly show lower prediction skills for the EMI index than the Niño-3.4 index in the November hindcasts (see Fig. 3a versus Fig. 4a). This result is consistent with previous studies showing that the WP El Niño is less predictable than the canonical El Niño (Hu et al. 2012). Furthermore, in contrast to the Niño-3.4 index, the EMI index exhibits little difference between November and May hindcasts at lead times of shorter than 5 months, which suggests that the spring prediction barrier is not as severe for the WP El Niño as for the canonical El Niño. In addition, similar to the Niño-3.4 index, EMI prediction skills also do not show significant differences when atmospheric resolution increases from T319 to T639 or ensemble size increases from 15 to 51. On the other hand, RMSEs and intraensemble spreads are both clearly smaller than those for the predicted Niño-3.4 index (see Figs. 3b,c versus Figs. 4b,c), which should be related to the fact that the WP El Niño usually presents weaker SST anomalies than the canonical El Niño (Hu et al. 2012). When atmospheric resolution increases from T319 to T639, the RMSE skill of EMI index is slightly (but insignificantly) improved in contrast to that of the Niño-3.4 index, and their ensemble spreads are also slightly larger but with a smaller magnitude compared to the Niño-3.4 index. The above two improvements together contribute to a larger SvR ratio in the predicted
EMI index when atmospheric resolution increases, even though the overconfidence problem consistently appears at all lead times.

Figure 5 (Fig. 6) shows horizontal distributions of the SST prediction skill in terms of anomaly correlations (RMSEs) at lead times of 3, 5, and 7 months for May ICs. Generally, for the correlation metric (Fig. 5), all T319_M15, T319_M51, and T639_M15 present almost identical skill distributions at the three lead times. In all hindcasts, high correlations are mostly located in the central and eastern tropical Pacific, with correlations over a large region greater than 0.7 for all three lead times. As the lead time increases from 3 to 5 months, the correlation drops markedly in the above skillful regions. For example, at the 3-month lead (Fig. 5a), many regions have anomaly correlations greater than 0.8, with some areas above 0.9; in contrast, at the 5-month lead (Fig. 5b), the correlation skill is mostly lower than 0.8. As the lead time increases to 7 months (Fig. 5c), the correlation skill recovers strongly in the central equatorial Pacific, where the correlation skill returns to >0.8. This rebound suggests that the model, as with most ENSO prediction systems (Jin et al. 2008; Zhu et al. 2012), exhibits higher skill in predicting the ENSO mature phase than the developing and decaying phases (the 7-month lead time for May ICs is corresponding to the calendar November), which is also seen from Fig. 3a. For the RMSE metric (Fig. 6), T319_M15, T319_M51, and T639_M15 also present similar skill distributions, except that T639_M15 seems to have slightly smaller RMSEs in the western equatorial Pacific at the 5- and 7-month lead times, which was more evident in the normalized RMSE (i.e., the RMSE divided by STD of observations; not shown). In all hindcasts, RMSEs grow continuously as a function of lead time in the eastern equatorial Pacific (see also Fig. 3b). This is different from the correlation.

Fig. 9. Diagnostics of predicted SSTAs at 5-month lead time with IC in November of 1982–2010 in (left) T319_M15, and the difference (middle) between T319_M15 and T319_M51 and (right) between T319_M15 and T639_M15. Distributions are shown for (top to bottom) anomaly correlation, RMSE (°C), ensemble spread (°C), and ensemble SvR ratios. The same color bar is used for RMSE and ensemble spread.
skills (Figs. 3a and 6), which present a clear skill rebound at month 7. The difference reflects a common phenomenon appearing in current ENSO prediction models, that is, the ENSO peak phases can be relatively well predicted, but their amplitudes are usually incorrectly predicted (Jin et al. 2008; Zhu et al. 2012).

Figure 7 presents the spatial distribution of ensemble spread of predicted SSTAs at lead times of 3, 5, and 7 months for the different hindcast sets for May ICs. Generally, the ensemble spread distribution is similar to the SSTA variation distribution (Fig. 2), with the largest spread appearing in the mideastern equatorial Pacific. The ensemble spread also grows as function of lead times, implying that the potential predictability generally decays as lead time increases. When comparing T319_M15 and T319_M51, it shows the increase of ensemble size from 15 to 51 does not have much effect on the ensemble spread. In contrast, the comparison between T319_M15 and T639_M15 indicates that higher atmospheric horizontal resolution significantly increases the ensemble spread in the mideastern equatorial Pacific at all three lead times (see also Fig. 3c). Figure 8 exhibits horizontal distributions of SvR ratio. As is common in almost all single-model ensembles (Palmer et al. 2004; Vialard et al. 2005; Saha et al. 2006; Weisheimer et al. 2009; Zhu et al. 2013a, among others), the ensemble spread is generally smaller than the RMSE, particularly at shorter lead times (e.g., 3 and 5 months). In the ECMWF system, the smallest SvR ratio appears in the western equatorial Pacific, which is different from the NCEP Climate Forecast System, version 2 (CFSv2) (Zhu et al. 2013a) where the smallest ratio exists in the eastern equatorial Pacific. This difference suggests that, even though the overconfidence problem is common in single-model predictions, their detailed spatial distributions have strong system dependence. In addition, the SvR ratio is too low in the west-central Pacific even at 5-month lead time, suggesting that the system is not reliable in forecasting central Pacific events during summer and fall. Comparing our three hindcast sets there is little difference when different ensemble sizes are used (15 versus 51), but clear differences with different atmospheric horizontal resolutions (T319 versus T639). In particular, when the atmospheric horizontal resolution increases from T319 to T639, the SvR ratio is greater than 1 over a sizable region in the eastern basin, especially at the 7-month lead time. The ratio increase in T639_M15 is mainly due to the increase in ensemble spreads (Fig. 7), but in the western Pacific it is also due to the decrease in RMSE (Fig. 6).

While the above horizontal distributions are all based on hindcasts starting from May ICs, similar conclusions can be derived from the November starts. Figure 9, as an example, presents horizontal distributions of anomaly correlations, RMSEs, ensemble spreads, and SvR ratios of predicted SSTAs in the three hindcasts at 5-month lead time. For the ensemble spreads and SvR ratios, increasing ensemble size from 15 to 51 does not contribute to discernible differences, but there are clear increases in ensemble spreads and SvR ratios when atmospheric horizontal resolution increases from T319 to T639, particularly in the Niño-3.4 region.

In addition, if we further look at the hindcasts with T1279, it can be found the ENSO prediction difference (Fig. 10) among T319_M15, T319_M51, and T639_M15 hindcasts present almost identical skill distributions in terms of both anomaly correlations and RMSEs, except that T639_M15 has slightly smaller RMSEs in the western equatorial Pacific. For the ensemble spreads and SvR ratios, increasing ensemble size from 15 to 51 does not contribute to discernible differences, but there are clear increases in ensemble spreads and SvR ratios when atmospheric horizontal resolution increases from T319 to T639, particularly in the Niño-3.4 region.

In addition, if we further look at the hindcasts with T1279, it can be found the ENSO prediction difference (Fig. 10) among T319_M15, T639_M15, and T1279_M15 is significantly lower than that resulting from the difference in prediction periods (1982–2011 in Fig. 5 versus 2000–11 in Fig. 10).

The above analysis indicates that, while most metrics are insensitive to changes in atmospheric horizontal
resolution, larger SSTA ensemble spreads were found in
the higher-resolution ensemble (T639_M15). However,
a more important question is whether the probabilistic
distributions have been improved accordingly as a result
of the increased ensemble spreads. Such probabilistic
distributions are vital for our skill evaluations by prob-
abilistic metrics, in which the prediction target is the
probability of an event occurrence ($P_f$). To explore
the question we examine the frequency distributions of
$P_f$ based on predicted SSTAs over the Niño-3.4 region,
where ensemble spreads exhibit significant differences
with atmospheric horizontal resolution. The predicted
SSTAs are divided into three categories: warm (larger
than 0.43°C), cold (less than −0.43°C), and neutral
(falling in between), and frequency distributions are
examined for each category. 0.43 is chosen because 43% of
a standard deviation is the tercile threshold for normally
distributed data, and the standard deviation of
Niño-3.4 index during the hindcast period is close to 1°C.
However, if a 0.5°C threshold is chosen, results do not
change. In addition, to accumulate enough samples for
deriving reliable frequency distributions, hindcasts at all
1–7 lead months are used, providing a bulk measure over
the full integration. Figure 11 shows the frequency dis-
tributions of $P_f$ with eight total probability bins, for
November (Figs. 11a–c) and May (Figs. 11d–f) starts. It
is clear that the derived distributions in all three hind-
casts are remarkably similar. For instance, comparing
T319_M15 (red curves) with T319_M51, and T639_M15,
respectively.

![Figure 11](image-url)

**Fig. 11.** Predicted SST (over the Niño-3.4 region) frequency distributions for the (a),(d) upper (warm), (b),(e) middle (neutral), and (c),(f) lower (cold) categories, based on hindcasts at the leading 1–7 months with ICs in (a)–(c) November during 1982–2010 and (d)–(f) May during 1982–2011. The probabilities are binned as 0.125-wide intervals (total of eight probability bins). Red, green, and blue curves correspond to hindcasts from T319_M15, T319_M51, and T639_M15, respectively.
between T319_M15 (red curves) and T319_M51 (green curves), it can be seen that an ensemble size of 15 is sufficient to realistically capture the frequency distributions of the predicted SSTAs over the Niño-3.4 region.

To directly compare the probabilistic skill in three hindcasts, the Brier skill scores (BSS) are calculated for the above three Niño-3.4 index-based ENSO categories, i.e., warm (Niño-3.4 SSTAs larger than 0.43°C), cold (Niño-3.4 SSTAs less than −0.43°C), and neutral (Niño-3.4 SSTAs falling in between). The BSS is a measure of the relative benefit of the forecasts with respect to using the naive climatological probabilities, in which three categories have even chances of occurrence. It is defined as $\text{BSS} = 1 - \text{BS}/\text{BSc}$, where BS is the Brier score, defined as the sum over all forecasts of the quadratic distance in probability space between the forecast probability and an observational step function that takes the value one (zero) if the event does (does not) verify, and BSc is the Brier score of the climatological forecast. It was found from the BSSs (Fig. 12) that, while all hindcasts generally show better skills than the climatologic forecast for three categories, the warm and cold categories seem to be better predicted than the neutral one. What is more interesting, consistent with Fig. 11, is that increasing either ensemble size from 15 to 51 or atmospheric horizontal resolution from T319 to T639 does not change much difference in the BSSs. In fact, higher resolution even occasionally degrades the probabilistic skill, for instance, in the predictions of the cold category from May initial states (solid curves in Fig. 12c). Generally, it can be concluded that there is little difference among the three sets of hindcasts in terms of probabilistic metrics.

We further use the reliability diagram (Wilks 2006; Zhu et al. 2013a), which compares the forecast probabilities against their corresponding frequencies of observed occurrence, to quantitatively examine the ENSO forecasting reliability in Project Minerva. If a forecast system is perfectly reliable in probability forecasting, its forecast probabilities of an event occurrence should be equal to the observed relative frequency, and, thus, it would display a 1:1 diagonal line in the diagram. Similar for frequency distributions (Fig. 11), several measures have been taken to accumulate a large enough number of samples for the reliability diagram. For one, contingency tables are calculated for each grid point in the Niño-3.4 area, and all forecasts (T319_M15, T319_M51, and T639_M15) during the lead 1–7 months are used. The predicted SSTAs are also divided into three categories based on the 0.43°C threshold. Furthermore, as a stricter requirement of sample size for reliability diagram, a slightly lower number of probability bins are chosen: 0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%. Such a lower number of probability bins results in an identical reliability diagram (Fig. 13) derived for all T319_M15, T319_M51, and T639_M15 hindcasts, which, however, is consistent with Fig. 11 showing similar frequency distributions among them with finer probability bins. As shown in Fig. 13, for all...
warm, cold and neutral categories in both November and May starts, hindcasts generally produce ENSO forecasts with the overconfidence problem, even though the estimation is still composed of certain uncertainties. The overconfidence bias is common to most reliability diagrams derived from current prediction systems (Saha et al. 2006; Zhu et al. 2013a). As to the sharpness (three inset histograms in Fig. 13), all three SSTA categories exhibit high confidence, except neutral categories with a May start, which has intermediate confidence.

4. Conclusions and discussion

This study examined ENSO prediction in Project Minerva, and the effect of both the atmospheric horizontal resolution and ensemble size on ENSO prediction is explored. Three sets of hindcasts conducted in Project Minerva are compared, starting from 1 May (1 November) during 1982–2011 (1982–2010): spectral T319 atmospheric resolution with 15 ensembles, spectral T639 with 15 ensembles, and spectral T319 with 51 ensembles. Our analysis shows that simply increasing either ensemble size from 15 to 51 or atmospheric horizontal resolution from T319 to T639 does not necessarily lead to major improvement in the ENSO prediction skill in this model. For deterministic prediction skill metrics, the three sets of predictions do not produce a significant difference in either anomaly correlation or root-mean-square error (RMSE). For probabilistic metrics, the increased atmospheric horizontal resolution generates a larger ensemble spread and thus increases the ratio between the intraensemble spread and RMSE. However, there is little change in the categorical distributions of predicted SST anomalies. Consequently, there is little difference among the three sets of hindcasts in terms of probabilistic metrics or prediction reliability.

Our results seem inconsistent with some previous investigations (Guilyardi et al. 2004; Gualdi et al. 2005; Navarra et al. 2008), where a beneficial impact of improving atmospheric horizontal resolutions on ENSO simulation/prediction was identified. Our analysis suggests that ENSO prediction skills in Project Minerva are generally insensitive to increases in atmospheric horizontal resolution beyond T319, and ensemble size beyond 15. However, we do agree with the explanation of improved results found in these studies; that is, in their experiments higher atmospheric resolution is much closer to their oceanic grid resolution, by which the atmosphere might be able to better capture the ocean modeled SST structures by avoiding too many interpolations, and the ocean–atmosphere coupling was consequently improved (Guilyardi et al. 2004; Gualdi et al. 2005). Actually, this idea may also explain our results. In our experiment, T319 in atmosphere is already close to the oceanic resolution in the tropical region (globally approximate 1°, but meridionally refined to 1/3° near the equator), and the higher-resolution atmosphere (T639 and T1279) may require finer-scale SST signals that the present low-resolution ocean model is unable to produce. Therefore, our results, together with previous experiments (Guilyardi et al. 2004; Gualdi et al. 2005; Navarra et al. 2008), may suggest a matching requirement between the atmospheric and oceanic model resolutions in coupling systems, which has not been emphasized enough previously when they were designed. On the other hand, it could also be possible that with current parameterizations the atmosphere does not improve for resolution higher than 50 km, and that the existing parameterizations will need to be reformulated to properly take advantage of those higher resolutions as in Project Minerva.

Acknowledgments. Funding for this study on the COLA part is provided by grants from NSF (ATM-0830068), NOAA (NA09OAR4310058), and NASA.
(NNX09AN50G). Computing resources provided by Yellowstone are also gratefully acknowledged.

REFERENCES


