Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1002/2017MS001204

Monthly ENSO Forecast Skill and Lagged Ensemble Size

L. Trenary[1,2], T. DelSole[1,2], M.K. Tippett[3,4], and K. Pegion[1,2]

[1]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA, USA, [2]Center for Ocean-Land-Atmosphere Studies, Fairfax, VA, USA, [3]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA, [4]Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

Key Points:
- Burst, lagged, and weighted lagged ensemble skill indistinguishable from that of the optimal lagged ensemble
- Current MSE of the NCEP operational ENSO forecasts is close to that estimated for the infinite ensemble
- Parametric method used to find optimal lagged ensemble size for real-time CFSv2 forecast of Niño 3.4 index

Correspondence to:
L. Trenary,
ltrenary@gmu.edu

Citation:
Trenary, L., DelSole, T., Tippett, M. K., & Pegion, K. (2018). Monthly ENSO forecast skill and lagged ensemble size. *Journal of Advances in Modeling Earth Systems*, *10*, 1074–1086. https://doi.org/10.1002/2017MS001204

Received 17 OCT 2017
Accepted 27 MAR 2018
Accepted article online 6 APR 2018
Published online 20 APR 2018

Abstract The mean square error (MSE) of a lagged ensemble of monthly forecasts of the Niño 3.4 index from the Climate Forecast System (CFSv2) is examined with respect to ensemble size and configuration. Although the real-time forecast is initialized 4 times per day, it is possible to infer the MSE for arbitrary initialization frequency and for burst ensembles by fitting error covariances to a parametric model and then extrapolating to arbitrary ensemble size and initialization frequency. Applying this method to real-time forecasts, we find that the MSE consistently reaches a minimum for a lagged ensemble size between one and eight days, when four initializations per day are included. This ensemble size is consistent with the 8–10 day lagged ensemble configuration used operationally. Interestingly, the skill of both ensemble configurations is close to the estimated skill of the infinite ensemble. The skill of the weighted, lagged, and burst ensembles are found to be comparable. Certain unphysical features of the estimated error growth were tracked down to problems with the climatology and data discontinuities.

1. Introduction

Operational centers use different forecast configurations to produce subseasonal-to-seasonal forecasts of the El Niño-Southern Oscillation (ENSO; e.g., Barnston et al., 2017). Some differences in configuration are related to the method used to generate ensemble members. In the burst configuration, ensemble members are initialized on the same start date. Alternatively, a lagged ensemble is formed by pooling forecasts initialized on different start dates that verify on the same target. Computationally, the lagged ensemble requires lower peak computational resources than the burst for the same ensemble size because fewer ensemble members are initialized for each start time. On the other hand, since a lagged ensemble is formed by pooling forecasts with different leads, ensemble averaging can reduce forecast skill if the lagged ensemble contains sufficiently disparate leads—something not seen in a burst configuration. Directly determining the ensemble size that yields the most skillful forecast for a lagged ensemble forecast system is a challenging process requiring extensive testing and computationally expensive model runs.

The role of ensemble size in prediction skill has been discussed in numerous papers. In general, skill improves only up to some upper limit that depends on the system's inherent predictability (Kumar & Hoerling, 2000). The optimal lagged ensemble can be computed analytically only for highly simplified models (e.g., first-order autoregressive processes; Kharin et al., 2001; Trenary et al., 2017). However, the problem of estimating the skill of an infinite ensemble forecast, or estimating the optimal lagged ensemble for realistic forecast models, has received relatively little attention. Recently, Trenary et al. (2017) developed a methodology capable of identifying the lagged ensemble configuration that minimizes the mean square forecast error without requiring the calculation of all possible candidate configurations. The method is based on fitting error covariances to a parametric model and then extrapolating to arbitrary ensemble size and initialization frequency. Remarkably, the method can estimate the skill of a burst ensemble even if the original data does not contain burst ensembles, simply by taking the limit of an infinitely small interval between initialization times. While the method was developed for subseasonal forecasts of the MJO, the methodology is general enough that it can be readily adapted and applied to forecasts targeting different time scales and variables.

The purpose of this paper is to evaluate the relative benefit of a burst versus lagged ensemble configuration in the context of monthly ENSO forecasts of the Climate Forecast System version 2 (CFSv2). Since operational centers use a fixed forecast configuration throughout the year (same number of ensemble members,

regardless of season or forecast amplitude), we seek the forecast configuration that performs best in an average sense. Accordingly, we pool all months together to estimate MSE.

Following from Trenary et al. (2017), in section 2, we show the forecast MSE can be directly estimated from a quantity known as the cross-lead error covariance matrix. This relation allows us to estimate the MSE for arbitrary initialization frequency as well as for burst ensembles. A description of the real-time CFSv2 forecast and hindcast data sets used in this study is presented in section 3. In section 3.2.1, we describe how the CFSv2 forecast climatology is smoothed to remove discontinuities that are apparent when the forecast climatology is expressed in terms of lead time and target (Tippett et al., 2018). This smoothing of the forecast climatology removes some unrealistic dependence on lead time from the forecast errors. In section 4.1, the cross-lead error covariance matrix is estimated using real-time CFSv2 forecasts of the Niño 3.4 index. Next, a parametric model of the lagged error covariance is fit using CFSv2 forecasts initialized once per day. Given the relatively short record of real-time forecasts, the robustness of our estimate of the cross-lead error covariance matrix is a concern. In this section, we establish the robustness of our estimate by showing that the parametric model obtained from real-time forecasts is consistent with a model fit to hindcast data. We then demonstrate the ability of the parametric model to accurately predict the structure of the cross-lead error covariance and the associated MSE of forecasts initialized four times daily. With the accuracy of the parametric model established, we show the reduction of forecast errors that can be achieved through lagged ensembles. In addition, the error of an optimally weighted lagged ensemble is compared to that of an equally weighted lagged ensemble. We then adapt our methodology to estimate the relative impact on forecast skill of using a burst ensemble versus a lagged ensemble. These results show that the skill of CFSv2 ENSO forecast with an infinite ensemble are not far from that of the current operational configuration. The paper then concludes with a summary of our findings.

2. Methodology

The forecast error $\epsilon(v, v-\tau)$ is defined as the difference between a forecast $f(v, v-\tau)$ and the observed anomaly $o(v)$ and is written as:

$$\epsilon(v, v-\tau) = f(v, v-\tau) - o(v), \tag{1}$$

where v is the verification time, $v-\tau$ is the initial condition time, and τ is lead time.

Trenary et al. (2017) show that the MSE of a lagged ensemble forecast can be derived directly from a quantity called the cross-lead error covariance matrix. In particular, the MSE of a lagged ensemble of size L for lead time τ is

$$MSE(L, \tau) = \frac{1}{L^2} \sum_{m=0}^{L-1} \sum_{n=0}^{L-1} C(\tau + m\Delta, \tau + n\Delta), \tag{2}$$

where Δ is the time interval between initialization times, and $C(i, j)$ is the cross-lead error covariance matrix

$$C(i, j) = \langle \epsilon(v, v-i\Delta) \epsilon(v, v-j\Delta) \rangle. \tag{3}$$

Note that when C is estimated from hindcast or forecast data, i and j typically are integers, but later we will parameterize $C(i, j)$ so that it can be evaluated even for fractional values of i and j. The diagonal element $C(i, i)$ gives the mean square error of a single forecast at lead i.

Trenary et al. (2017) proposed a parameterization of the cross-lead error covariance matrix. Following the tradition started by (Lorenz, 1982), this parameterization assumed that MSE growth could be captured by a sigmoid function. However, there is very little known about the structure of the error statistics between lagged ensemble members. Empirically, the covariance drops discontinuously from the diagonal to the off-diagonal elements and then decays more gradually thereafter. Because the covariances are very noisy, it is hard to justify a parameterization more complicated than simple exponential or linear decay. Trenary et al. (2017) chose an exponential decay with coefficients that depend linearly on lead time. However, fitting this function to ENSO data analyzed in this paper lead to an identifiability problem because the decay of the cross-lead error covariances was so slow that very slow exponential decay was indistinguishable from linear decay. Accordingly, we modified the parameterization so that the decay of the cross-lead covariances was simply a linear function of the difference in lead times (i.e., no exponential decay). Furthermore, the starting

point of the linear decay was the sigmoid function. The result of all these changes lead to the following parameterization for the cross-lead error covariances:

$$\widehat{C}_{\text{off-diag}}(i,j)=\frac{\epsilon_o}{1+e^{-\kappa(\min[i,j]-\tau_o)}}(a|i-j|+b), \tag{4}$$

where min $[i,j]$ denotes the minimum of i and j, the caret ($\widehat{\ }$) denotes a parameterized quantity, and $|i-j|$ denotes the absolute value of the difference between i and j (where the time step Δ is absorbed in the parameters).

As mentioned above, a discontinuity exists between the diagonal and the first off-diagonal element of the cross-lead error covariance matrix. For a first-order autoregressive process, this discontinuity is such that the diagonal element is a factor of 2 larger than the value extrapolated from the off-diagonal elements (Trenary et al., 2017). However, such a discontinuity is not expected at very short lead times because the difference between errors is not random for dynamical systems, whereas for autoregressive processes the difference always includes a white noise term (see DelSole, 2000, for further discussion). Therefore, we assume that no discontinuity exists at short lead times and that a factor-of-two discontinuity emerges at sufficiently large lead time. We further assume that the transition between the two regimes can be captured by an exponential function. These considerations lead us to parameterize the diagonal elements as:

$$\widehat{C}_{\text{diag}}(i,i)=(2-e^{-\alpha i})\widehat{C}_{\text{off-diag}}(i,i), \tag{5}$$

where i is the lead time.

The result of the above changes is a new parameterization that involves only six parameters, compared to eight parameters for the model of Trenary et al. (2017).

3. Data

3.1. Forecast and Observation Data

We analyze real-time forecasts and hindcasts from the CFSv2, a fully coupled atmosphere-ocean-land forecast model (Saha et al., 2014). The real-time forecasts are initialized 4 times per day (0, 6, 12, and 18Z) beginning 1 February 2011 and we restrict our analysis to the period 2011–2016. The seasonal hindcasts span the years 1999–2010 and are initialized every 5th day, starting on 1 January 1999, with initializations at the same 6 h intervals. The forecast (hindcasts) targets are monthly averages, up to a 9 month lead, where the first forecast target is the average of the first complete calendar month following the initialization date, and so on. For example, the first target month available for a forecast initialized on 1 January is February, and the second target month is March, and so forth. Here the lead time (τ, refer to equation (2)) of the forecast (hindcast) is defined to be the number of days from the initialization date to the start of the target calendar month. As an example, a forecast initialized on 27 July provides a 5 day lead forecast targeting August, whereas a forecast initialized on 28 July is a 4 day lead forecast for the same target month. Anomalies are computed relative to the 1999–2010 climatology conditioned on lead and forecast target date. A detailed discussion of the climatology is provided in section 3.2.1. Real-time forecasts, hindcast data, and the associated climatology are provided by the International Research Institute for Climate and Society (http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.EMC/.CFSv2/).

The phase and magnitude of ENSO are measured using the standard Niño 3.4 index, a time series found by area averaging the forecast and hindcast sea surface temperature in the eastern central tropical Pacific (5°N–5°S, 170°W–120°W). Unless otherwise noted, our analysis is based on forecasts and hindcasts of the Niño 3.4 index initialized at 0Z.

Forecasts and hindcasts are verified against the observed Niño 3.4 index over the same period, computed using monthly values from the NOAA Optimum Interpolation Sea Surface Temperature version 2 (https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html) (Reynolds et al., 2002).

3.2. Data Issues

3.2.1. Climatology

To evaluate forecast skill, we must first estimate the forecast anomalies by subtracting a forecast climatology from each forecast. The climatology provided by NCEP matches the forecast format and is a function of

initialization day and lead, such that the first lead is the first complete calendar month following the initialization date, and so on. To compute forecast errors as a function of lead time, the climatology must be rearranged into a function of lead time and target month. The result of rearranging the NCEP climatology into lead-target format is shown in Figure 1a. The first striking feature displayed in this figure is the lead-dependent biases in climatology for forecasts targeting months in boreal fall and early winter. For example, the climatological value for forecasts targeting September drops nearly 2°C from the 1 day to the 100 day lead forecast. These cold biases are primarily associated with forecasts initialized during months May–August. For other target months, the forecast climatology shows limited lead time dependence, as seen by the nearly uniform temperatures in Figure 1a. Also prominently displayed in this figure are discontinuous changes in the climatological values at approximately 30 day intervals superimposed on these lead-dependent biases. For example, consider the September climatology, wherein the 31 and 32 day lead forecast climatological values differ by almost 1°C. These periodic discontinuities and model biases are clearly visible when the climatology is plotted as a function of lead time for a few select months, shown as the dashed lines in Figure 1c. The reason for this behavior is that the climatology provided by NCEP is a smooth function of start time for a given monthly lead. The impact of these discontinuities are captured by the MSE, shown as the red curve in Figure 2, which exhibits peaks at approximately 30 and 60 days. Peaks in MSE at the longer lead times are difficult to identify individually due to large background error. In recent work, Tippett et al. (2018) provide additional examples of how the NCEP climatology can impact forecast skill of variables other than the Niño 3.4 index.

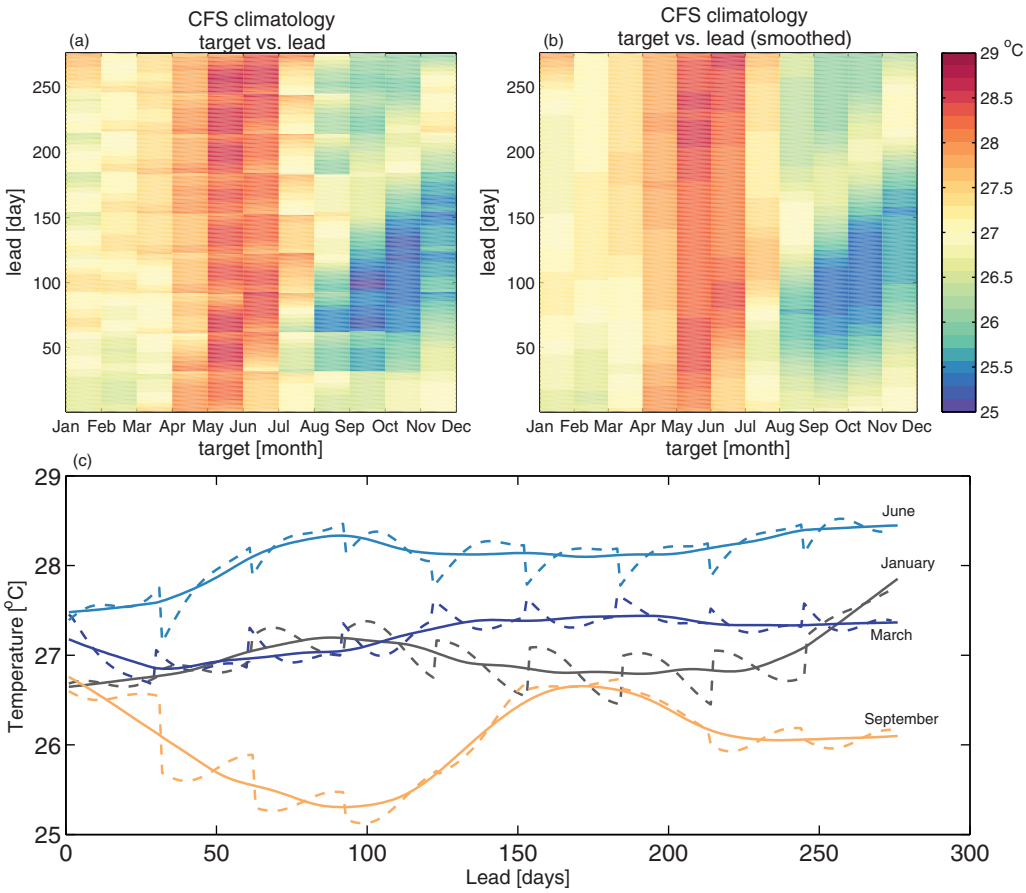

Figure 1. Climatology of the forecast Niño 3.4 index in CFSv2 as a function of lead time and target. (a) The climatology of Niño3.4 provided by NCEP but rearranged into lead and target format. (b) The climatology smoothed using a 60 point running average. The raw climatology in Figure 1a is the 1999–2010 climatology obtained from the CFSv2 Retrospective forecasts. (c) Climatology as a function of lead time for target months January, March, June, and September. The raw climatology is shown as the dashed line and the smoothed climatology from Figure 2b is shown as the solid curve.

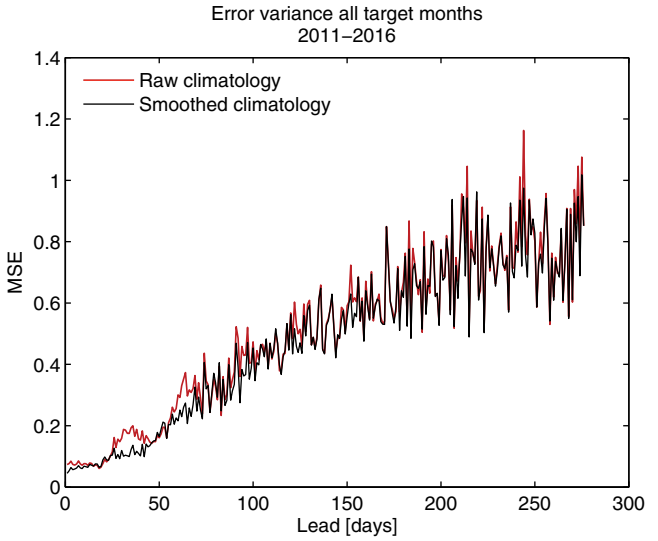

Figure 2. Error variance for the Niño 3.4 index in CFSv2 as a function of lead time when forecast anomalies are evaluated with respect to the raw (red) and the smoothed climatologies (black) shown in Figures 1a and 1b, respectively. The MSE is estimated for all target months during 2011–2016 for forecasts initialized at 0Z.

To remove the resulting discontinuities, we apply a 60 point running average in lead time to the lead-target climatology. The smoothed climatology, shown in Figure 1b, captures the lead-dependent biases identified in Figure 1a, but removes the spurious 30 day variations in the lead-dependent climatology, as clearly shown by the smooth curves in Figure 1c. Moreover, the MSE estimated using the smoothed climatology, shown as the black curve in Figure 2, no longer shows peaks at 30 and 60 days.

Additional analysis found that the climatology fit in lead-target space using only the real-time data, yielded similar structure to that found in Figure 1b. Since the real-time data spans only a limited number of years, we elected to use NCEP climatology with smoothing since more data was used to construct the climatology. Here we estimate the lead-dependent climatology by smoothing the CFS forecast climatology for given target months. Tippett et al. (2018) propose an alternative method for estimating a climatology that is continuous in lead and target. While these approaches do not give exactly the same results, both forecast climatologies capture a similar structure of the lead-dependent model climatology (not shown).

3.2.2. Discontinuities in Errors

Forecast errors for a few select target months are shown as the color curves in Figure 3. For a given target month, the time series for each year are centered relative to the forecast period 2011–2016. To facilitate comparisons, each curve has been offset by adding increments of 2°C to each time series, and the horizontal line in this figure shows the mean for each respective time series. Note that since the data are centered relative to the entire forecast period, the mean line is not necessarily zero. Since, the first initialization in the real-time forecast record is 1 February 2011, the first available forecasts are for March 2011. A striking feature in these plots is the discontinuity in errors in the second half of 2016 (see red curves in Figures 3d–3f which show errors for forecasts targeting July, September, and November). These sharp jumps in forecast error are related to the abrupt correction of a bias in the initial conditions of the operational CFSv2 in March of 2016 (Shawki et al., 2017). Further discussion on the forecast re-initialization is provided by NCEP (http://www.nco.ncep.noaa.gov/pmb/changes/downloads/CFSv2_Atlantic_cold_bias_problem.pdf).

The impacts of the 2016 error discontinuity are most pronounced for forecasts targeting the last 6 months of the year, particularly at intermediate and long lead times. The inclusion of this discontinuity inflates the rate of error growth at long leads (not shown). Consequently, we exclude 2016 from our analysis. In addition to the obvious issues with the forecasts reinitialized in 2016, there may also be also some issues with forecast initialized in 2015. After removing forecasts initialized in 2016, we tested the sensitivity of our results to the inclusion of 2015, by repeating our analysis with and without forecasts initialized in 2015. We found no discernible impact of the 2015 forecasts on our results. Moreover, there is no clear discontinuity present in the forecasts for monthly targets during 2015. As such, we retain the years 2011–2015 for all subsequent analysis.

4. Results

4.1. Fitting the Cross-Lead Error Covariance Matrix

The parametric model of the cross-lead error covariance matrix for the real-time Niño 3.4 forecasts is estimated across all target months for the forecast period 2011–2015. This approach is consistent with how operational forecasts are made, since the ensemble size is held fixed regardless of when the model is initialized. We consider this reasonable, since we are interested in identifying the optimal lagged ensemble that provides the best overall forecast regardless of target and irrespective of ENSO strength.

It should be noted that by combining forecasts in this way, we neglect seasonal dependence in ENSO forecast skill. While it is likely that seasonal differences exist for the cross-lead error covariance matrix, we were unable to confirm with certainty that these differences are larger than sampling errors. Consequently, our analysis focuses on forecast skill across all target months.

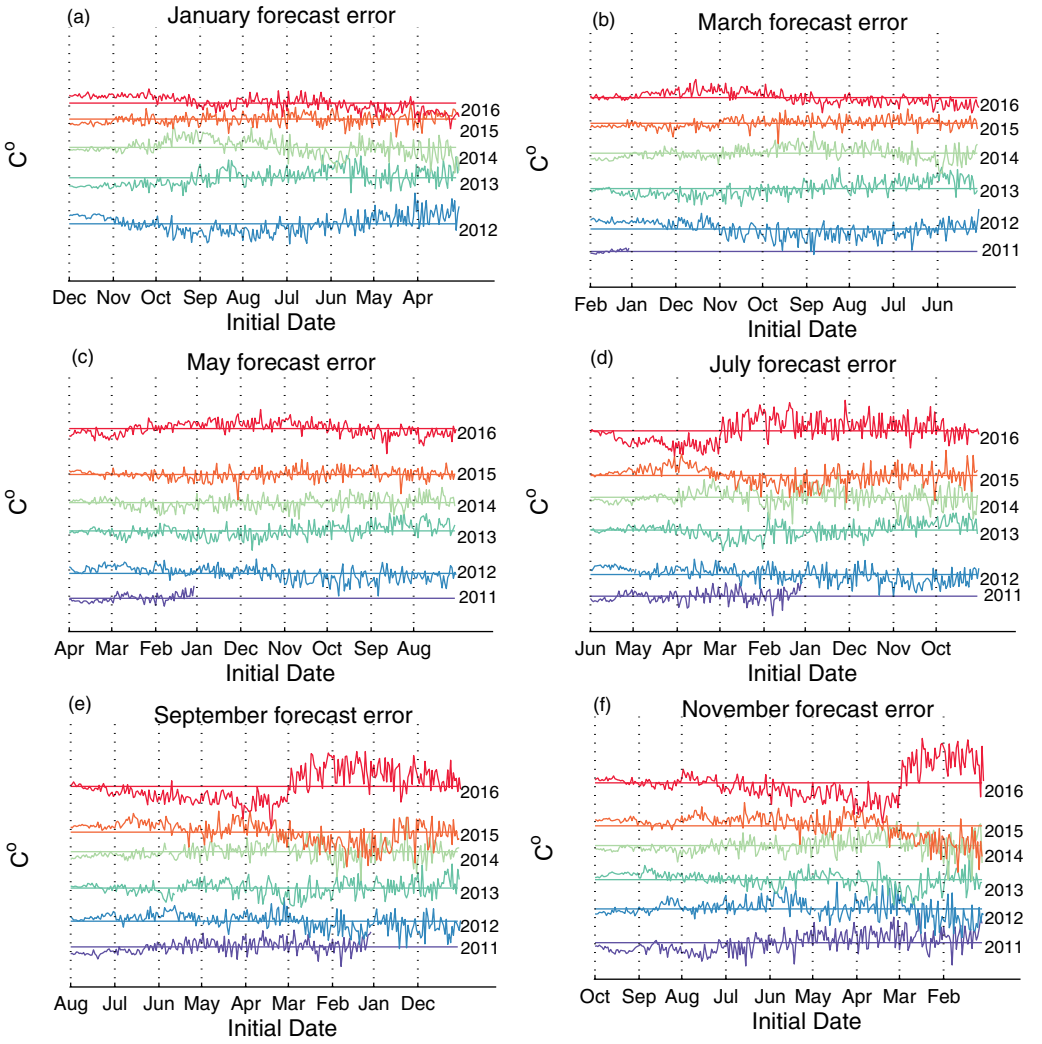

Figure 3. Real-time forecast errors for the Niño 3.4 index in CFSv2 as a function of lead time for target months (a) January, (b) March, (c) May, (d) July, (e) September, and (f) November for the years 2011–2016 for forecasts initialized at 0Z. The year is listed next to the corresponding curve. The errors are in degrees Celsius and the horizontal axis shows the lead time in days. The errors are offset by adding incremental values of 2°C to each time series. The horizontal lines show the mean for each respective time series. Each error time series for a given target month are centered relative to the forecast period 2011–2016.

The cross-lead error covariance matrix for real-time forecasts of the Niño 3.4 index for CFSv2 forecasts initialized at 0Z is shown in Figure 4a. We emphasize that the forecasts are initialized *daily* but the targets are *monthly averages*. We use all available daily initializations for a given monthly target to estimate the cross-lead error covariance according to equation (3). Including all four initializations (0, 6, 12, and 18Z), produces the error covariance matrix shown in Figure 4c, with similar error growth properties. The structure of these cross-lead error covariance matrices both have a patchwork appearance that is noisier in comparison to the error growth found for 45 day lead forecasts of MJO indices from hindcasts of the same model (Trenary et al., 2017). There is no simple answer as to why the error growth properties differ between MJO and ENSO. There are likely a multitude of factors that account for these differences, including a fundamental difference in predictability between the two phenomena.

We fit our new parametric model (described in section 2) to the CFSv2 forecasts of the Nino3.4 index. The covariance matrix obtained from fitting the parametric model to the error growth of real-time CFSv2 monthly Niño 3.4 forecasts can be seen in Figure 4b. The parameters of the fit are listed in Table 1.

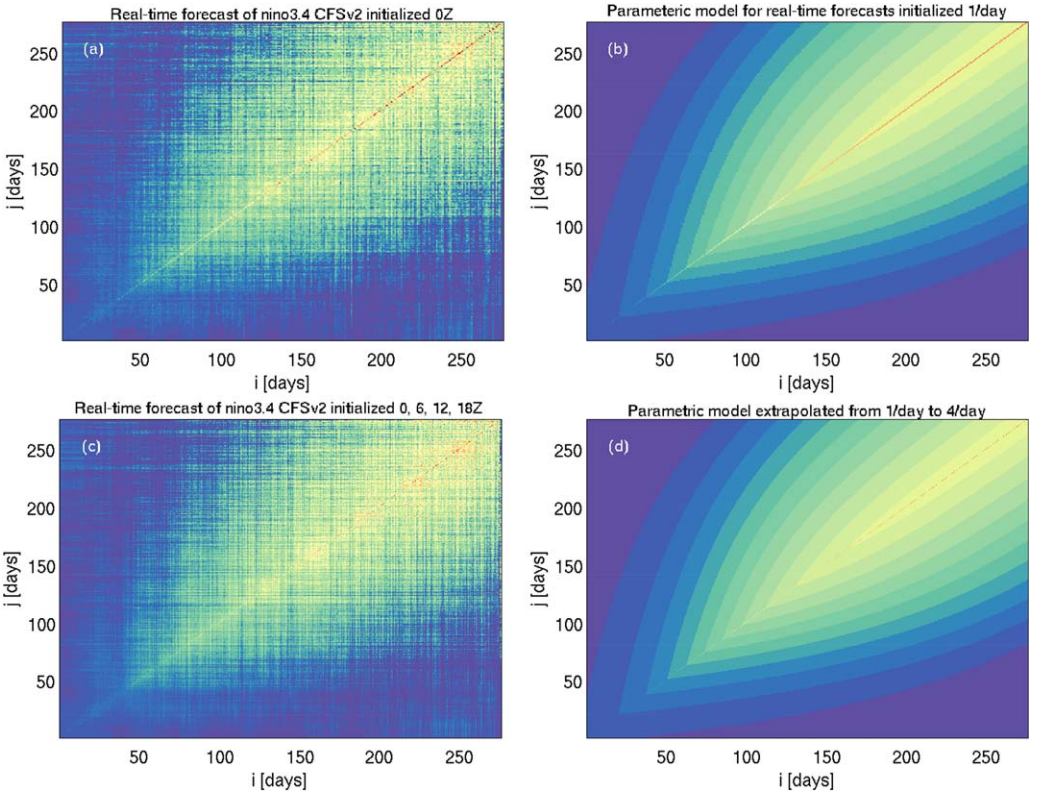

Figure 4. Cross-lead error covariance matrix $C_{(i,j)}$ given by equation (3) for the real-time CFSv2 forecasts of the Niño 3.4 index for monthly forecasts during the period 2011–2015. The axes give the lead time in units of days. Plot (a) shows the cross-lead error covariance matrix for Niño 3.4 estimated directly from CFSv2 real-time forecasts initialized at 0Z. Plot (b) shows the corresponding parametric model for cross-lead error covariance matrix shown in Figure 4a. Similarly, plot (c) shows the cross-lead error covariance matrix as in Figure 4a, but for real-time forecasts initialized 4 times per day (0, 6, 12, 18 Z). Plot (d) shows the cross-lead error covariance matrix for four initializations per day inferred from the parametric model fit at 1 day intervals.

The diagonal and cross section of the cross-lead error covariance matrix, along with the corresponding parametric fits are shown in Figures 5a and 5b, respectively. Note that the goodness of fit of the parametric model is evidenced by the fact that the smooth curves representing the parametric fit, pass through the forecast based estimates, represented by the noisy curves. In essence, the parametric model provides a smoothed estimate of the cross-lead error covariance, accurately capturing the rapid error growth along the diagonal and the linear decay of the error covariance away from the diagonal. By reducing the noise, the detailed structure of the cross-lead error covariance matrix, shown in Figure 4b, is clearer. The parametric model fit for the cross-lead error covariance is a piece-wise continuous function of time and can be evaluated at arbitrary times. We exploit this property and estimate the cross-lead error covariance matrix for an ensemble forecast initialized every six hours using the parametric equations with parameters estimated from forecasts initialized every 24 h (see Trenary et al., 2017, for details of this extrapolation procedure). The cross-lead error covariance matrix from our parametric model of error growth extrapolated to account for 4 per day initialized forecasts is shown in Figure 4d, and is in good agreement with the actual cross-lead error covariance matrix shown in Figure 4c.

Table 1
Parameters of the Parametric Model of the Cross-Lead Error Covariances

	Real-time forecasts 2011–2015	Hindcasts 1999–2010
α	0.008	0.021
a	−0.003	−0.002
b	0.59	0.61
$\epsilon_\circ$	0.6	0.5
κ	0.03	0.02
$\tau_\circ$	72.20	70.2

To test robustness of our estimates, we recompute the fit using hindcast data of the Niño 3.4 index for the period 1999–2010. The parameters recovered for the real-time and hindcast fits are displayed in the left and right columns of Table 1, respectively. The estimated parameters for the sigmoid ($\epsilon_\circ$, κ, and $\tau_\circ$) and the linear decay (a and b) are

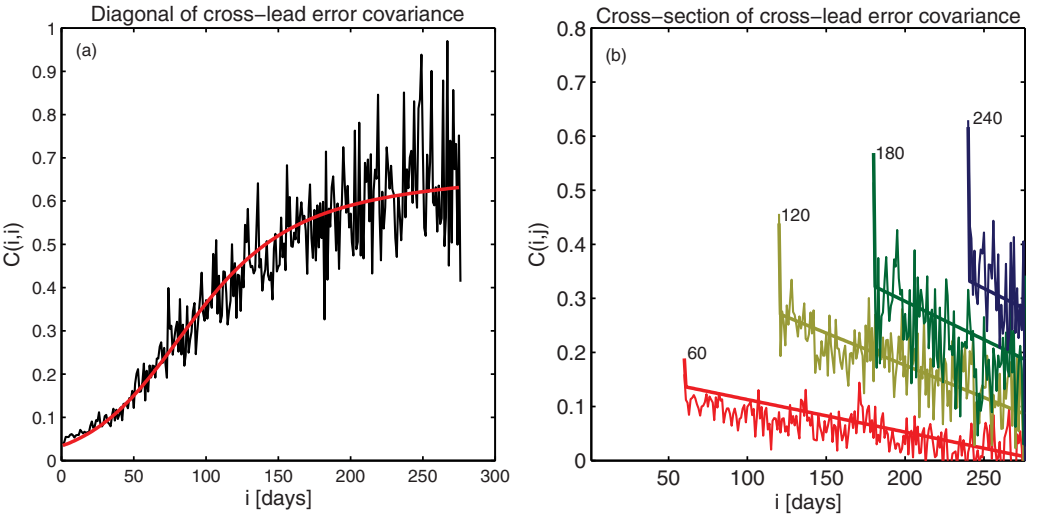

Figure 5. (a) Diagonal and (b) cross section of the cross-lead error covariance matrix for CFSv2 forecasts of Niño 3.4 index initialized at 0Z. Estimates based on real-time forecasts are shown as the noisy curves and the corresponding parametric fit by the smooth curves, passing through the real-time data. The number next to each curve in Figure 5b denotes the lead time in days.

comparable across the data sets. The greatest difference is found for the parameter α, which is almost three times greater than the estimate from the real-time fit. To test if this difference could be due to differences in initialization frequency, we subsample the error covariance of the real-time data, shown in Figure 4a, to have the same 5 day coverage as the hindcast and then recalculate the fit. The recovered fit for the sub-sampled data, shown as the red curve in Figure 6a, passes through the real-time data shown in black, and is in excellent agreement with the fit found using real-time data with daily coverage, which is overlaid in blue. Note that given the similarity in the real-time 1 and 5 day fit, the fitted curves overlap and are not visible separately. In fact the parameters are nearly identical (not shown), indicating that error growth properties are not sensitive to changes in initialization frequency.

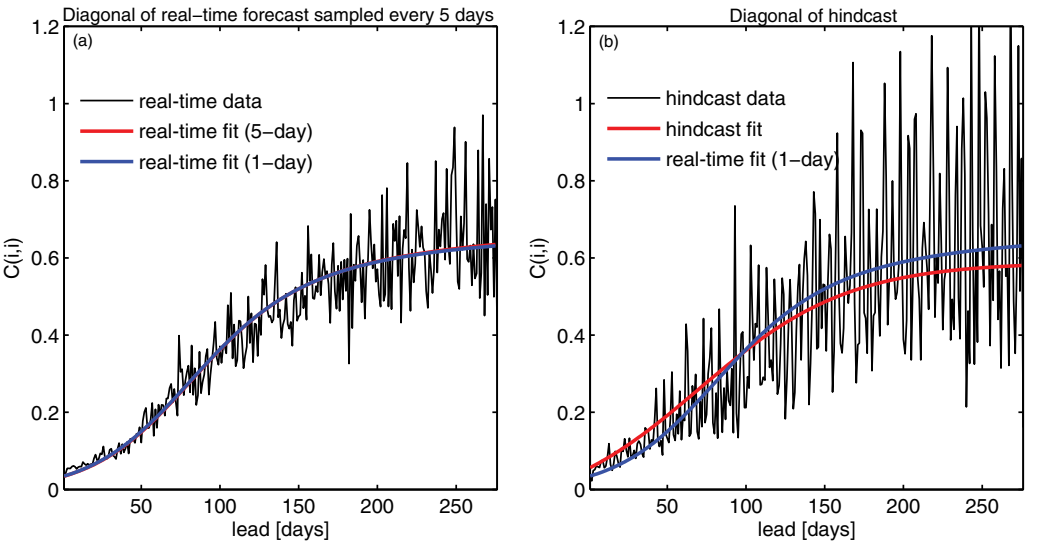

Figure 6. (a) Diagonal of the cross-lead error covariance matrix for real-time CFSv2 forecast of the Niño 3.4 index (black) and corresponding fits with an initialization of once per day (blue) and every 5th day (red). Note, because the fits are nearly identical, it appears as if only one fit is plotted. (b) Diagonal of the cross-lead error covariance matrix for CFSv2 hindcasts of the Niño 3.4 index (black) and the resulting fit (red). For comparison, the fit estimated from real-time data with daily initialization is overlaid in blue.

To gain some insight into why the estimates of α differ and how this might effect the fit, we compare the hindcast MSE (or equivalently, the diagonal of the hindcast cross-lead error covariance matrix), shown as the black curve in Figure 6b, with fits estimated from hindcast and real-time data, shown in the same figure as the red and blue curves, respectively. Comparing the black curves in Figures 6a and 6b, it is clear that the hindcast estimates of the MSE are significantly noisier than those recovered from real-time data. This may seem somewhat counter intuitive, since the number of years making up the hindcast data set is more than twice that of the real-time record. However, since the hindcasts are initialized every 5th day as opposed to daily, fewer samples are available in the hindcast dataset to estimate the covariances. For example, consider a 10 day lead forecast. Because the hindcast data is initialized every 5 days, the appropriate initial condition for a 10 day lead forecast does not exist for every single month. In contrast, the real-time data is initialized every day, so a 10 day lead forecast exists for each month. When all the initial conditions are collected for a given lead, there can be 58 samples from the (shorter) real-time set but only 12 samples from the (longer) hindcast set. This counter-intuitive result is a direct consequence of not initializing the hindcast every day.

Considering the large uncertainty in the hindcast estimate of MSE (black curve in Figure 6b), differences in the estimates of α are indistinguishable. Visually, we can see this by comparing the two fits shown in Figure 6b—while different, both curves provide a reasonable fit to the hindcast data, in fact it could be argued that the real-time estimate, shown in blue, provides a better fit at the longer lead times. It should be noted that while the difference in fit appear to be fairly minor, the percent difference in MSE can be as large as 20% (not shown). Nonetheless, within the limits of the data, parameter estimates appear to be consistent across datasets.

4.2. Lagged Ensemble Size and Minimization of MSE

Referring to equation (2), the MSE can be estimated as a function of lead time and lagged ensemble size by summing the elements of the cross-lead error covariance matrix. Estimates using the real-time and parametric error covariance matrices for Niño 3.4 forecasts are shown in Figure 7, where each colored curve represents the MSE for the specified lead. The size of the lagged ensemble, shown along the horizontal axis in each plot, quantifies the number of days between initialization and target month. For example, a 1 day lagged ensemble is composed of one member when considering a single initialization and four members when four initializations per day are used. The MSE has been normalized by the variance of the observed monthly Niño 3.4 index over the period 2011–2015, such that values less than 1 indicate a skillful forecast. The left column, Figures 7a and 7c, shows the MSE for real-time forecasts initialized 1/d and 4/d, respectively. The corresponding parametric based estimates are shown in Figures 7b and 7d, respectively. By and large the forecast and parametric based estimates of MSE are in remarkable agreement. Comparing the estimates for the 1/d initializations, shown in Figures 7a and 7b, we see that the parametric model is able to capture both the magnitude and curvature of the lead-dependent MSE. Moreover, both the data and parametric model results indicate that the CFSv2 forecasts of Niño 3.4 index can be skillful for more than 150 days. This long lead time skill in ENSO forecasts with the CFSv2 is well documented (e.g., Barnston et al., 2012, 2017; Saha et al., 2014). Agreement between forecast and parametric based estimates of MSE is likewise captured when four initializations per day are included in the estimate. Comparing Figures 7a and 7c, we see that the inclusion of four initializations reduces the magnitude and curvature of the MSE, especially at longer lead times. These same characteristics are similarly captured by the parametric model, as shown in Figure 7d. Note that the parametric estimates of MSE for the 4/d initializations are extrapolated from the parametric model estimated from the 1/d initializations. This indicates that the parametric model can skillfully estimate error growth for initializations not included in the original fit.

As detailed in Trenary et al. (2017), the optimal lagged ensemble is defined as the ensemble size that minimizes the MSE. Referring to Figure 7, there clearly exists an optimal size that is lead-dependent for CFSv2 forecasts of the Niño 3.4 index. In the discussion that follows, we will consider only the optimal lagged ensemble inferred from the MSE computed using the parametric model of the cross-lead error covariance. We do this since the parameterized error covariance reduces the noise in the MSE estimates, making it easier to identify the optimal lagged ensemble. For forecasts initialized once per day, Figure 7b shows the optimal lagged ensemble size increases from 2 at a 30 day lead to 8 at a lead of 150 days. For reference, the 2 and 8 member ensemble size is indicated by the dotted vertical lines in 7. Comparing Figures 7b and 7d, we see that the primary impact of including more initializations is a reduction in MSE at leads greater than 30 days. When the initialization frequency is increased to 4 times per day, the optimal ensemble size

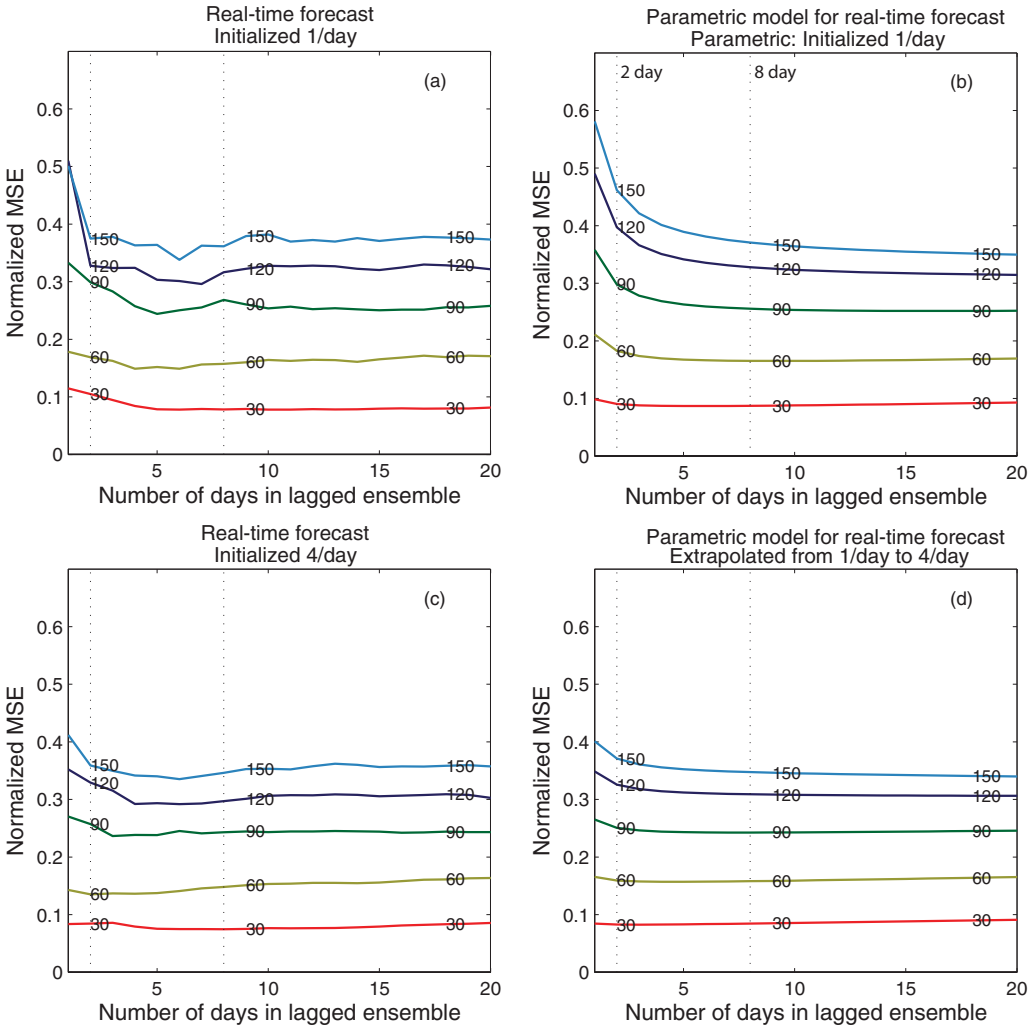

Figure 7. Normalized MSE of the monthly mean Niño 3.4 forecast as function of lagged ensemble size (horizontal axis) and lead (colored curves—the number denotes the forecast lead in days) in the CFSv2. (a) normalized MSE for Niño 3.4 CFSv2 forecast initialization at 0Z. (b) Empirically derived normalized MSE computed using the fit shown in Figure 4b. (c) Normalized MSE for Niño 3.4 forecasts when 0, 6, 12, and 18Z initializations of CFSv2 are used. (d) Empirically derived normalized MSE computed using the fit shown in Figure 4b interpolated to include four separate initializations. The dotted vertical lines denote the location of the 2 and 8 day lagged ensemble. All MSE estimates are normalized relative to 2011–2015.

decreases for 30–60 day leads, but remains unchanged for lead times greater than 90 days, as can be seen in Figure 7d. Generally, for leads greater than 30 days and less than 150 days, the optimal size ranges between 1 and 8 days, depending on the lead time. The optimal ensemble identified here for CFSv2 forecasts with four initializations per day is similar to the configuration used for real-time forecasts issued as part of the North American Multi-Model Ensemble (NMME; Kirtman et al., 2013) or operational forecasts issued by Climate Prediction Center. In both cases, an ensemble size of 8 and 10 lagged members are used, respectively.

Once the minimum in the MSE has been reached at the optimal lagged ensemble size, the addition of each subsequent ensemble member will cause the errors to grow, albeit very slowly. That being the case, from a deterministic standpoint, a large ensemble size will not produce a more skillful forecast, but the large lagged ensemble size may prove to be beneficial for probabilistic forecasts. Since the MSE experiences only minimal growth with increased ensemble size, the forecasts can be viewed as being drawn from the same probability density function. Consequently, larger lagged ensembles may be useful in probabilistic forecasts, where the skill tends to increase with ensemble size.

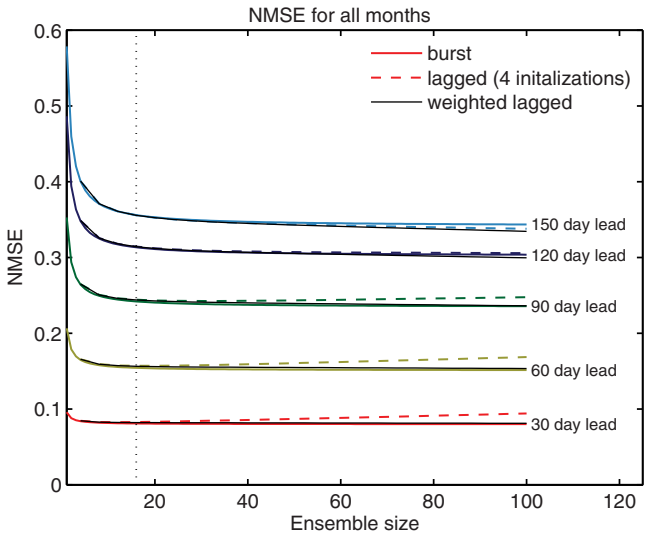

Figure 8. Parametrically derived normalized MSE for CFSv2 forecasts of Niño 3.4 as a function of ensemble size for a "burst" ensemble (solid color curves), lagged ensemble with four initializations per day (color dashed curves), and optimally weighted lagged ensemble for forecasts initialized 4 times per day (black curves). Each set of color curves corresponds to MSE estimates for the specified lead time. Estimates for the "burst" ensemble are computed assuming the ensemble members are initialized an infinitesimal time step apart. The dotted vertical line denotes the location of the 8 day lagged ensemble when four separate initializations are included.

4.2.1. Burst Versus Lagged Ensemble

Trenary et al. (2017) show how the parametric model of the cross-lead error covariance can be adapted to estimate the MSE of a burst ensemble. Specifically, a burst ensemble can be approximated by considering ensemble members initialized an infinitesimal time step apart. In this section, we use our previously developed methodology to evaluate the relative benefits of using a burst versus lagged ensemble in Niño 3.4 forecasts from the CFSv2.

The normalized MSE estimates for a burst and lagged ensemble with four initializations per day for specified lead times, are displayed as the solid and dashed curves in Figures 8, respectively. We find that for all leads and an ensemble size less than 30 days, the forecast skill as a function of ensemble size is roughly equal for the lagged (with four initializations) and burst ensemble configuration. When more than 30 ensemble members are included, the MSE of the burst ensemble saturates. This saturation of the MSE provides an estimate of the infinite ensemble MSE, since no reduction in MSE occurs with the addition of more members. In contrast, the MSE of the lagged ensemble continues to growth with ensemble size since the addition of each lagged member introduces forecasts initialized further from the target date.

This analysis suggests that burst and lagged ensembles can have comparable ENSO forecast skill, when an ensemble size of 30 members or less is used. However, it is possible that an optimally weighted lagged ensemble is capable of out-performing a burst ensemble. In the next section, we evaluate the impacts of optimal weighting on the error growth and optimal ensemble size for the lagged ensemble.

4.2.2. Optimal Weights of a Lagged Ensemble

We now consider weighting the lagged ensemble members. If the forecast is bias corrected, and we desire an unbiased weighted forecast ensemble, then the weights should sum to one and the MSE can be written as,

$$MSE_k(L,\tau)=\mathbf{w}^T\mathbf{C}\mathbf{w}, \tag{6}$$

where $\mathbf{w}$ is vector containing the weights, $\mathbf{C}$ is the covariance matrix appropriate for the forecast configuration, and the superscript T denotes the transpose operation (see DelSole et al., 2017, for more details). If the ensembles are equally weighted, $w=1/L$ and equation (2) is recovered. Using the method of Lagrange multipliers to impose the constraint that the weights sum to one, Trenary et al. (2017) derived the following expression for the optimal weights

$$\mathbf{w}=\frac{\mathbf{C}^{-1}\mathbf{z}}{\mathbf{z}^T\mathbf{C}^{-1}\mathbf{z}}, \tag{7}$$

where $\mathbf{z}$ is a vector of all ones. If we were to relax the constraint that the forecast be unbiased, then lower MSE is possible but, the optimal weights would depend on more than just the cross-lead error covariance and therefore would require estimating additional covariance information.

Applying equation (7) to the parametric error covariance for Niño 3.4 CFSv2 forecast (see Figure 4) optimal weights are found as a function of ensemble size for lead times of 60 and 90 days, and are shown in Figures 9a and 9b, respectively. Each colored curve shows the optimal weights for the ensemble size specified in days. The recovered weights display a distinctive structure, with greater weight given to the forecasts initialized closest to the target date. Moreover, as the ensemble size increases, each additional ensemble member is taken from a forecast initialized further from the target date and the optimal weights decreases to account for the loss is skill. Interestingly, we find the weights increase when the largest ensemble size is approached. DelSole et al. (2017) discuss the reasons for this curious behavior and why it is not an artifact of our parametric model.

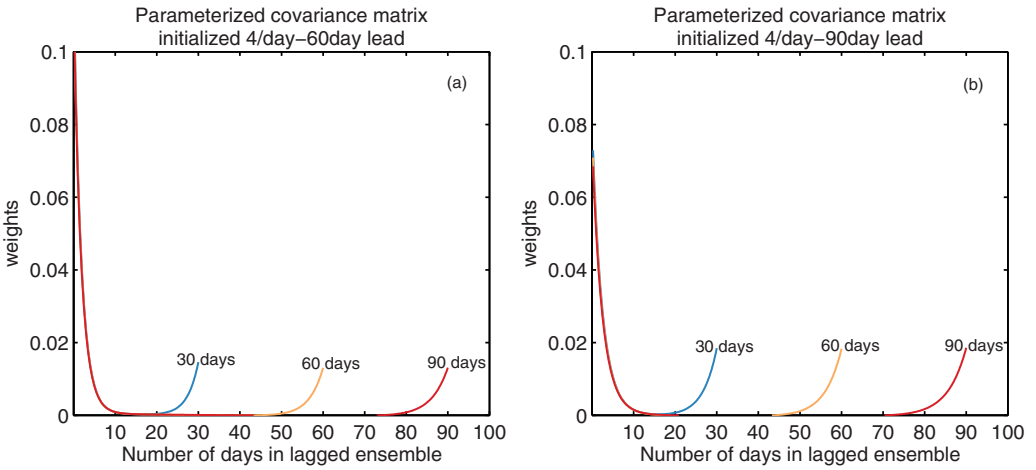

Figure 9. Optimal weights computed from the parametric cross-lead error covariance matrix of Niño 3.4 forecasts with an initialization frequency of 4 times per day. Optimal weights are found according to equation (7). The results are shown for a (a) 60 day lead and a (b) 90 day lead. Each curves shows the optimal weights for a given lagged ensemble size, measured in units of days and listed at right.

Having identified the optimal weights, we want to know how optimal weighting of the lagged ensemble improves forecast skill relative to burst ensemble configuration or a lagged ensemble with equal weighting. Using the optimal weights derived from equation (7), the weighted MSE is found according to equation (6) and is shown in Figure 8 as the thick black curves for the specified lead times. These results demonstrate that regardless of lead time, the optimally weighted forecast is indistinguishable from the infinite ensemble at all ensemble sizes. Furthermore, we can see that the MSE recovered for the optimal lagged ensemble is close to the infinite ensemble.

5. Conclusions and Summary

In this study, we examine the MSE of a lagged ensemble with respect to ensemble size and forecast configuration for monthly forecasts of the Niño 3.4 index in the CFSv2. As in Trenary et al. (2017), the MSE is estimated from a cross-lead error covariance matrix of the forecast variable. We show that the MSE for arbitrary initialization frequency and ensemble size can be inferred by extrapolating a parametric model fit to the cross-lead error covariance matrix for monthly forecasts of the Niño 3.4 initialized once per day. Robustness of the real-time forecast estimates of the parametric model to sampling errors is established by comparing with fits obtained from the hindcast data set. The parameteric model used in this paper differed from that used in Trenary et al. (2017). The need to modify the parametric model arose because the difference in time scales between ENSO and MJO leads to an identifiability problem when estimating the parametric model.

Applying this methodology to real-time forecasts of the Niño 3.4 index, we find that for forecasts initialized 4 times per day, the optimal ensemble size increases from a 1 day lagged ensemble at a 30 day lead to an 8 day lagged ensemble for leads greater than 60 days. The optimal lagged ensemble identified here is consistent with forecast configuration currently being used operationally as part of the NMME (Kirtman et al., 2013) and the official seasonal forecasts issued by the Climate Prediction Center, where 8 and 10 day lagged members are used, respectively. In contrast to deterministic forecast skill, we speculate that because the change in MSE is so small with increasing lagged ensemble size, the large lagged ensemble may prove beneficial for probabilistic ENSO forecasts. Lastly, it is worth noting that relative to other variables in the climate system, ENSO is highly predictable and the optimal lagged ensemble identified for an ENSO forecast may not be applicable to forecasts of other, less predictable variables (e.g., Kharin et al., 2001).

Adapting our methodology to account for the inclusion of burst ensemble members, we find that the optimal ensemble size is virtually indistinguishable for a burst, lagged, and optimally weighted lagged ensemble configurations. For example, when an eight member ensemble is used (i.e., eight bursts or 2 day lagged member with four initializations per day), the difference in skill between the ensemble configurations is less

than 1%. This difference remains small until an ensemble size of ~7–10 day is reached, at which point the MSE of the lagged ensemble continues to increase and that of the burst saturates. This saturation point of MSE for a burst ensemble denotes the limit of skill for an infinite ensemble, since the addition of more ensemble members no longer improves forecast skill. We find that the current operational configuration of 8–10 day lagged ensemble is close to the infinite ensemble for monthly ENSO forecasts.

Acknowledgments
This research was supported primarily by the National Oceanic and Atmospheric Administration, under the Climate Test Bed program (NA10OAR4310264). Additional support was provided by the National Science Foundation (AGS-1338427), National Aeronautics and Space Administration (NNX14AM19G), the National Oceanic and Atmospheric Administration (NA14OAR4310160 and NA14OAR4310184). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies. Forecast data and the associated climatology are provided by the International Research Institute for Climate and Society from their website http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.EMC/.CFSv2/. Observed SST data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html.

References

Barnston, A., Tippett, M. K., L'heureux, M. L., Li, S., & DeWitt, D. G. (2012). Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society*, *93*, 631–651. https://doi.org/10.1175/BAMS-D-11-00111.1

Barnston, A. G., Tippett, M. K., Ranganathan, M., & L'heureux, M. L. (2017). Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Climate Dynamics*, 1–20. https://doi.org/10.1007/s00382-017-3603-3

DelSole, T. (2000). A fundamental limitation of Markov models. *Journal of Atmospheric Sciences*, *57*(13), 2158–2168. https://doi.org/10.1175/1520-0469(2000)057 < 2158:AFLOMM>2.0.CO;2

DelSole, T., Trenary, L., DelSole, T., & Tippett, M. K. (2017). The weighted-average lagged ensemble. *Journal of Advances in Modeling Earth Systems*, *9*, 2739–2752. https://doi.org/10.1002/2017MS001128

Kharin, V. V., Zwiers, F. W., & Gagnon, N. (2001). Skill of seasonal hindcasts as a function of the ensemble size. *Climate Dynamics*, *17*(11), 835–843. https://doi.org/10.1007/s003820100149

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L. III, Paolino, D. A., Zhang, Q., et al. (2013). The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction, Phase-2 toward developing intra-seasonal prediction. *Bulletin of the American Meteorological Society*, *95*, 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Kumar, A., & Hoerling, M. P. (2000). Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bulletin of the American Meteorological Society*, *81*(2), 255–264. https://doi.org/10.1175/1520-0477(2000)081<0255:AOACMO>2.3.CO;2

Lorenz, E. N. (1982). Atmospheric predictability experiments with a large numerical model. *Tellus*, *34*, 505–513.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. (2002). An improved in situ and satellite SST analysis for climate. *Journal of Climate*, *15*, 1609–1625.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Shawki, D., Field, R. D., Tippett, M. K., Saharjo, B., Albar, I., Atmoko, D., et al. (2017). Long-lead prediction of the 2015 fire and haze episode in Indonesia. *Geophysical Research Letters*, *44*, 9996–10005. https://doi.org/10.1002/2017GL073660

Tippett, M. K., Trenary, L., DelSole, T., Pegion, K., & L'heureux, M. L. (2018), Sources of bias in monthly CFSv2 forecast climatology. *Journal of Applied Meteorology and Climatology*. http://doi.org/10.1175/JAMC-D-17-0299.1

Trenary, L., DelSole, T., Tippett, M. K., & Pegion, K. (2017). A new method for determining the optimal lagged ensemble. *Journal of Advances in Modeling Earth Systems*, *9*, 291–306. https://doi.org/10.1002/2016MS000838