AGU PUBLICATIONS

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

10.1002/2016MS000838

A new method for determining the optimal lagged ensemble

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

[1]George Mason University, Fairfax, Virginia, USA, [2]Center of Ocean-Land-Atmosphere Studies, Fairfax, Virginia, USA, [3]Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, USA, [4]Department of Meteorology, Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah, Saudi Arabia

Key Points:
- Method for determining the optimal forecast protocol for a lagged ensemble is proposed
- Method can infer skill for arbitrary ensemble size and initialization frequency
- For CFSv2 hindcasts of MJO, optimal size is one member for lead < week, and four members for lead > week

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

Citation:
Trenary, L., T. DelSole, M. K. Tippett, and K. Pegion (2017), A new method for determining the optimal lagged ensemble, *J. Adv. Model. Earth Syst.*, *9*, 291–306, doi:10.1002/2016MS000838.

Received 18 OCT 2016
Accepted 4 JAN 2017
Accepted article online 7 JAN 2017
Published online 31 JAN 2017

Abstract We propose a general methodology for determining the lagged ensemble that minimizes the mean square forecast error. The MSE of a lagged ensemble is shown to depend only on a quantity called the cross-lead error covariance matrix, which can be estimated from a short hindcast data set and parameterized in terms of analytic functions of time. The resulting parameterization allows the skill of forecasts to be evaluated for an arbitrary ensemble size and initialization frequency. Remarkably, the parameterization also can estimate the MSE of a burst ensemble simply by taking the limit of an infinitely small interval between initialization times. This methodology is applied to forecasts of the Madden Julian Oscillation (MJO) from version 2 of the Climate Forecast System version 2 (CFSv2). For leads greater than a week, little improvement is found in the MJO forecast skill when ensembles larger than 5 days are used or initializations greater than 4 times per day. We find that if the initialization frequency is too infrequent, important structures of the lagged error covariance matrix are lost. Lastly, we demonstrate that the forecast error at leads $\geq$10 days can be reduced by optimally weighting the lagged ensemble members. The weights are shown to depend only on the cross-lead error covariance matrix. While the methodology developed here is applied to CFSv2, the technique can be easily adapted to other forecast systems.

1. Introduction

Dynamical model forecasts of the Madden-Julian Oscillation (MJO) have become more skillful over the past 15 years [*National Academy of Sciences*, 2016]. Despite routine production of subseasonal forecasts by many operational centers, currently there is no standard protocol, with some centers providing forecasts at monthly intervals and others on a weekly or daily basis. The lack of a standard procedure makes model intercomparisons difficult. Likewise, different operational centers run their subseasonal prediction systems using different ensemble configurations. Computational considerations motivate the use of forecasts based on lagged ensembles, whereby forecasts initialized at different start dates but verifying on the same target date are averaged. In a lagged ensemble, an optimal ensemble size exists because increasing ensemble size can improve skill through ensemble averaging, but it can also degrade skill by including forecasts with longer lead times. Previous work by *Chen et al*. [2013] addressed this question for seasonal lagged ensemble forecasts using a brute force method, whereby forecast skill was estimated explicitly from an existing large-ensemble hindcast data set. However, not all operational centers have the necessary resources available to perform the large number of hindcast experiments needed to find the optimal lagged ensemble size. Here we present a method that can estimate the skill of an *arbitrary* lagged ensemble given only a finite number of hindcasts and no additional numerical experiments. Specifically, we develop a parametric model for the covariances of forecast error that generalizes the *Lorenz* [1982] parametric model for error growth. We apply this methodology to the Climate Forecast System version 2 (CFSv2), although we note that the technique can be adapted for use with other forecast systems. A caveat of this methodology is that the optimal forecast configuration is estimated with respect to the MSE. Estimates based on other skill measures, such as probabilistic skill scores, may yield different optimal ensemble sizes.

This paper shows that the mean square error (MSE) of a lagged ensemble can be inferred from a quantity called the *cross-lead error covariance matrix*. We propose a parametric model for the covariance matrix whose parameters can be estimated from a single ensemble member initialized at a constant frequency. Then error covariances at arbitrary lead times and start frequencies can be estimated by extrapolation.

Indeed, we can even estimate the MSE of a "burst" ensemble, in which many forecasts are initialized at the same start date, from a data set with only one forecast per start date. Further, our parameterization also allows us to estimate the optimal *weighted* lagged ensemble.

The proposed method is applied to CFSv2 forecasts of standard MJO indices. We find that a five member ensemble, based on one-per-day initialization, is reasonably close to the optimal ensemble size for subseasonal forecasts with lead times greater than 7 days. For shorter leads, the forecast error is smallest when one member is used. Operationally, the subseasonal forecasts issued by National Center Environmental Prediction (NCEP) use 16 ensemble members per day, here we show that four initializations per day is sufficient for providing a skillful MJO forecast in the CFSv2, in so far as MSE is concerned. If the initialization frequency is too low (e.g., $>$5 days), then the estimated parametric model does not capture detailed structures of the lagged error covariance matrix, making it difficult to evaluate forecast skill empirically. Lastly, we show that an optimally weighted lagged ensemble can be determined from the cross-lead error covariance matrix and can improve skill even further at long leads ($>$10 days).

This paper is laid out as follows: We first describe the CFSv2 hindcast data used for this study and outline the methodology for computing the forecast MJO indices. In section 3, we derive the relation between MSE and the lagged error covariance matrix. We then describe our parameterization procedure for the cross-lead error covariance matrix and apply it to CFSv2 hindcasts initialized once per day. The methodology is tested by showing that it can accurately predict the MSE of CFSv2 hindcasts initialized four times per day. Having demonstrated the accuracy of the methodology, we use the procedure to investigate the sensitivity of forecast error to changes in ensemble size and initialization frequency. We then demonstrate the reduction in forecast error achieved by optimal weighting of lagged ensemble members. The paper concludes with a summary of our findings.

2. Data

2.1. Forecast Data

We analyze the 45 day hindcasts (reforecasts) from the CFSv2, a fully coupled atmosphere-ocean-land forecast model [*Saha et al*., 2013]. The hindcasts were initialized 4 times per day (0Z, 6Z, 12Z, and 18Z). Each initialization was run forward 45 days, with model output provided at 6 h intervals. Here we use the daily averages of the 6 h data and our analysis focuses on winter months (November to February) for the period 1999–2010.

2.2. MJO Indices

MJO indices are computed from the CFSv2 hindcasts using the methodology laid out by *Gottschalck et al*. [2010], which is based on the combined empirical orthogonal function (EOF) analysis of *Wheeler and Hendon* [2004]. We compute the latitudinally averaged (15°S–15°N) hindcast fields of outgoing longwave radiation, 850 hPa zonal winds, and 200 hPa zonal winds. Each latitudinally averaged field is then normalized by the standard deviation of the corresponding observed globally averaged field. The data are then projected onto the *observed* EOFs to obtain the forecast RMM1 and RMM2 indices. The bias corrected anomalies of both time series are found by subtracting a climatology that is a function of initial month, initial day, and lead time. Because the projection is linear, anomalies can be computed either before or after the operation. Frequently, low frequency variability is removed from the RMM indices by removing the mean of the most recent 120 days [*Gottschalck et al*., 2010]. Here we retain the interannual variability in the RMM indices because subseasonal variability of El Niño-Southern Oscillation will likely impact the predictability of the MJO. Unless otherwise stated, our analysis is based on the lagged ensemble forecasts of the MJO initialized on 0Z.

Forecasts are verified against observed RMM indices over the same period. As described above, the RMM indices are based off the combined EOF analysis of latitudinally averaged (15°S–15°N) outgoing longwave radiation, zonal winds at 850 and 200 hPa. The observed indices are found by projecting daily time series of latitudinally averaged anomalies of these same fields onto the multivariate EOF patterns. The outgoing longwave radiation data are measured by the NOAA polar-orbiting satellite and zonal winds are taken from NCEP/NCAR reanalysis. Consistent with forecast indices, the anomalies are computed relative to seasonal cycle and interannual variability is retained. For more details refer to *Wheeler and Hendon* [2004]. The

observed indices used in this study were made available by Dr. Matthew Wheeler at http://poama.bom.gov.au/project/maproom/RMM/.

3. Results

3.1. Mean Square Error and the Cross-Lead Error Covariance Matrix

We now present a general method for computing the forecast skill of a lagged ensemble forecast system. Let $f_k(v, v-\tau)$ denote the forecast anomaly of the kth MJO index at verification time v and lead time τ. The corresponding forecast error is defined as

$$\epsilon_k(v, v-\tau) = f_k(v, v-\tau) - o(v), \tag{1}$$

where $o(v)$ is the observation anomaly at verification time v. The mean of the lagged-ensemble forecast is defined as

$$\tilde{f}_k(v, v-\tau, L) = \frac{1}{L}\sum_{l=0}^{L-1} f_k(v, v-\tau-l\times\delta t), \tag{2}$$

where L is the size of the lagged ensemble, δt is the time interval between initialization times (assumed equal for simplicity), and the corresponding error is

$$\tilde{\epsilon}_k(v, v-\tau, L) = \tilde{f}_k(v, v-\tau, L) - o(v). \tag{3}$$

The mean square error (MSE) of the lagged ensemble forecast is

$$MSE_k(L, \tau) = \langle(\tilde{\epsilon}_k(v, v-\tau, L)^2\rangle, \tag{4}$$

where the brackets denote an average over verification v. As indicated above, MSE depends on the size of the lagged ensemble L and lead time τ. A direct calculation shows that

$$MSE_k(L, \tau) = \frac{1}{L^2}\sum_{m=0}^{L-1}\sum_{n=0}^{L-1} C_k(\tau+m\times\delta t, \tau+n\times\delta t), \tag{5}$$

where $C_k(i, j)$ is the *cross-lead error covariance matrix* of the forecast for the kth index:

$$C_k(i, j) = \langle\epsilon_k(v, v-i)\epsilon_k(v, v-j)\rangle. \tag{6}$$

The cross-lead error covariance matrices for RMM1 and RMM2 hindcasts are shown in Figures 1a and 1b. The diagonal elements of each matrix give the MSE of the corresponding RMM index as a function of lead for a single ensemble member. Since the MJO forecast skill depends on both indices, the total MSE for a single ensemble member at a given lead is obtained by finding the diagonal element of the matrix at that lead for RMM1 and RMM2 and then summing the two numbers. The off-diagonal elements give the covariance between forecast errors at different lead times but verifying on the same day. According to (5), the MSE of a lagged ensemble is obtained by summing all elements that fall within an $L \times L$ square, whose lower left corner is anchored to the diagonal at lead time τ, as illustrated schematically in Figure 2.

The MSE of MJO hindcasts as a function of lagged-ensemble size for CFSv2 hindcasts initialized once per day (at 0Z) are shown in Figure 3a for varying lead times. Each colored curve in Figure 3a corresponds to the MSE for a fixed lead time in days, as specified by the corresponding label. The size of the lagged-ensemble is indicated on the horizontal axis and is quantified by the number of days spanned by all start dates in the ensemble. For this example, we use forecasts initialized only on 0Z, so a 1 day lagged ensemble has one member, a 2 day lagged ensemble has two members, and so on. The MSE is normalized by the climatological variance of the respective index, thus values less than 1 (horizontal black lines) indicate a skillful forecast, conversely values greater than 1 are associated with a no-skill forecast. From this figure, we see that at short leads (leads $<$7 days), skill is degraded after averaging forecasts beyond the most current one. Thus, for lead times less than a week, no improvement in skill results from using a lagged ensemble. For longer lead times (leads $>$7days), the MSE reaches a minimum at about a 5 day lagged ensemble. The MSE obtained when all four (0Z, 6Z, 12Z, and 18Z) initializations are used in the lagged ensemble is shown in Figure 3c. The primary impact of including more initializations is a reduction in MSE at leads greater than 10

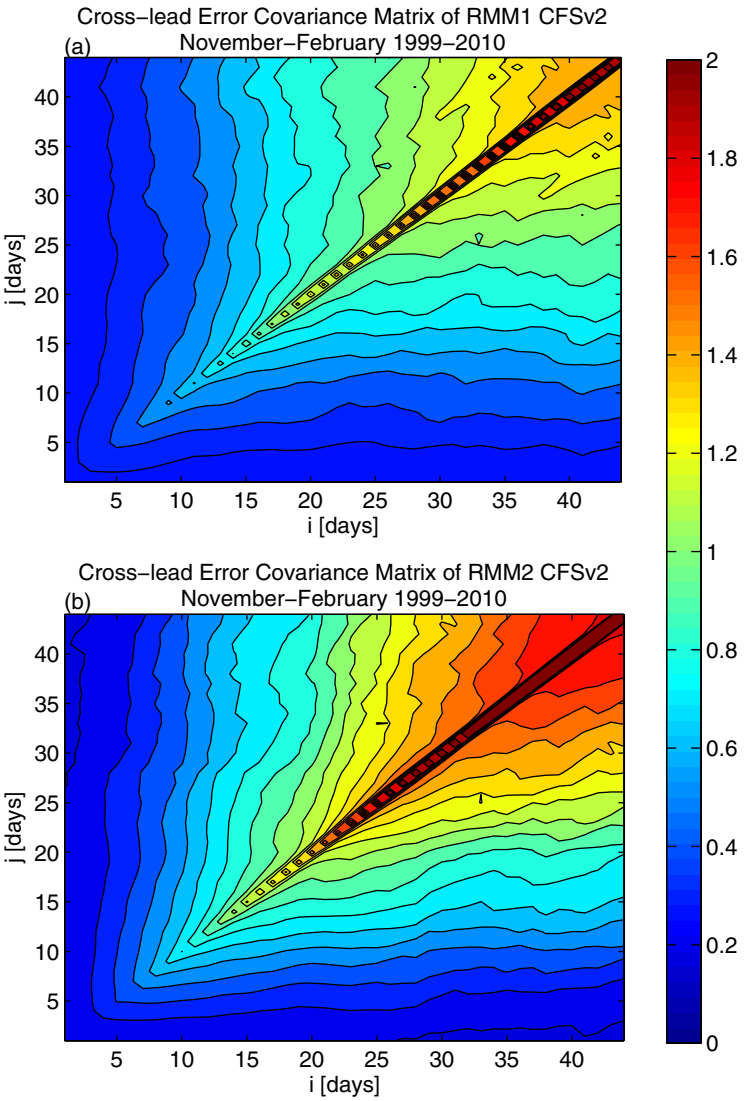

Figure 1. Cross-lead error covariance matrix $C_k(i,j)$ given by equation (6) for the boreal winter CFSv2 hindcasts of (a) RMM1 and (b) RMM2. The axes give the lead time in units of days.

days. Depending upon the ensemble size, both Figures 3a and 3c, shows that boreal winter MJO can be predicted with skill out to 25 days in CFSv2.

3.2. Parameterizing the Cross-Lead Error Covariance

Equation (5) shows that if the cross-lead error covariance matrix is known as a function of the two leads *i* and *j*, then the MSE of a lagged ensemble of any size could be computed. With this in mind, we seek a simple parametric model of the cross-lead covariance matrices of the RMM1/RMM2 indices for the boreal winter 0Z CFSv2 hindcasts for the period 1999–2010.

To develop some intuition for the structure of the cross-lead error covariance matrix, a quantity which has not been studied before, we compute it for a first-order autoregressive model (AR(1)) in Appendix and show the result in Figure 4. For this model, the diagonal elements grow exponentially for short lead times and saturate at twice the climatological variance at long lead times. In contrast, the off-diagonal elements of $C(i, j)$ depend only on the minimum lead (i.e., minimum of *i* or *j*). Importantly, there is a discontinuity between the diagonal element and the neighboring off-diagonal elements. Qualitatively, the error covariance matrices of the CFSv2 RMM1/RMM2 shown in Figures 1a and 1b, possess some of these same features. For instance, at long leads a sharp discontinuity between diagonal and off-diagonal elements can be seen, and at short leads the off-diagonal elements tend to depend only on the minimum lead. However, there are differences. Specifically, at long leads the off-diagonal elements tend to decrease along a row or column as one moves away from the diagonal. Also, the diagonal elements do not grow exponentially with lead time (not shown). These differences are sufficiently great as to make the AR(1) parametric model too simple to use for parameterizing RMM1/RMM2 forecast errors.

The lead time dependence of the cross-lead error covariances is shown as the blue curves in Figure 5. The off-diagonal elements generally decrease exponentially away from the diagonal element. One might be tempted to start from the diagonal element and fit the entire curve by an exponential, but the off-diagonal elements differ fundamentally from the diagonal elements in that the former is a covariance while the latter is a variance. As a result, there exists a fundamental discontinuity between the diagonal and off-diagonal elements. Accordingly, we fit the off-diagonal elements by themselves, and then fit the diagonal elements

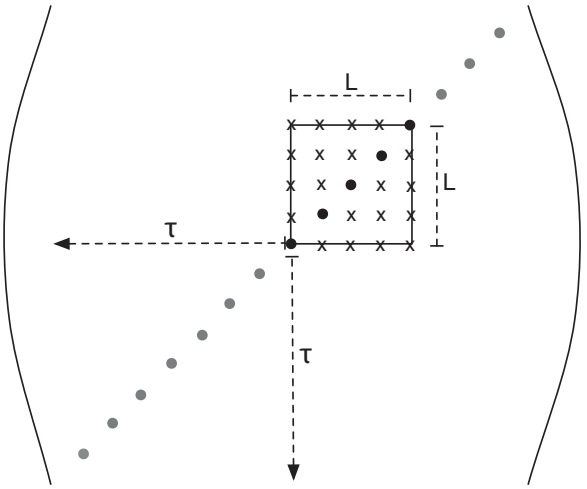

Figure 2. Schematic showing the elements of the cross-lead error covariance matrix that are summed when computing the MSE of a lagged ensemble forecast of size L and lead τ. Circles and crosses indicate diagonal and off-diagonal elements, respectively. Lead time increases to the right and upward.

in a manner that depends on the off-diagonal fit. The above considerations lead us to adopt the following parametric model for the off-diagonal elements:

$$\hat{C}(i,j) = a_{\min[i,j]} e^{-\gamma_{\min[i,j]}|i-j|} + b_{\min[i,j]}, \quad (7)$$

where $\min(i,j)$ denotes the minimum of i and j and the caret ($\hat{}$) denotes a parameterized quantity. The parameter γ defines the exponential decay rate, a defines the initial amplitude of the exponential decay, and b defines the saturation value of the off-diagonal elements. These parameters are assumed to depend on the minimum lead τ linearly as follows:

$$\begin{aligned} a_\tau &= \beta_a \cdot \tau + \alpha_a, \\ \gamma_\tau &= \beta_\gamma \cdot \tau, \\ b_\tau &= \beta_b \cdot \tau + \alpha_b. \end{aligned} \quad (8)$$

We do not claim that the above parameterization works for all error covariances. Here the error growth is sufficiently close to linear that the above parameterization works well. The zero intercept of γ ensures that β_γ is positive for all lead times, corresponding to exponential decay away from the diagonal element.

The above parameters were estimated for each MJO index separately by minimizing the sum square difference between the observed error covariances and the parameterized values derived from (7). Only the off-diagonal elements were used to fit the parametric equation—no diagonal elements were included in the estimation. As a result, the diagonal element extrapolated from (7) (i.e., the sum of the a and b parameters) does not match the diagonal element recovered from the CFSv2 data. This discontinuity is a real feature of the error covariance and not an artifact of our fitting procedure. Accordingly, we fit a separate function to the residual between the CFSv2 diagonal elements and the diagonal elements extrapolated from (7). Further, if we directly fit the diagonal elements, rather than the above residuals, the resulting curve may result in a cross-lead error covariance matrix that is not positive definite, which has implications for the finding the optimal weights (see section 4). Visual inspection of these residuals (not shown) suggest that they are well fit by the logistic growth law,

$$\hat{C}_{residual}(\tau) = \frac{\epsilon_o}{1 + e^{-\alpha(\tau - \tau_o)}}, \quad (9)$$

where ϵ_o is the maximum error, α is the initial error growth rate, and τ_o is the inflection point. The error curve described by (9) is characterized by slow initial growth, followed by a period of rapid growth that gradually plateaus to saturation, giving the curve its distinctive "S" or sigmoid shape. This model for error growth was proposed by *Lorenz* [1982]. The above parametric model implies that the diagonal elements are given by

$$\hat{C}(\tau,\tau) = \hat{C}_{residual}(\tau) + a_\tau + b_\tau. \quad (10)$$

Because $\hat{C}_{residual}$ is positive, the diagonal element will be larger than any other element along the corresponding row or column. This parameterization does not guarantee a positive definite matrix, but our experience indicates that it does in practice.

Consolidating the above equations leads to an eight-parameter model for the cross-lead error covariance matrix. This matrix is a generalization of the *Lorenz* [1982] error growth model by accounting for cross-lead covariances between ensemble members.

The parameterized error covariances for both RMM1 and RMM2 for select leads are shown as the red curves in Figure 5. Comparison with the estimated values from CFSv2 (i.e., the blue curves) shows that the fits are generally quite good for both variables, though the fits for RMM2 estimate a sharper transition between the off and along diagonal error, particularly at longer lead times. In any event, with minor exception, the

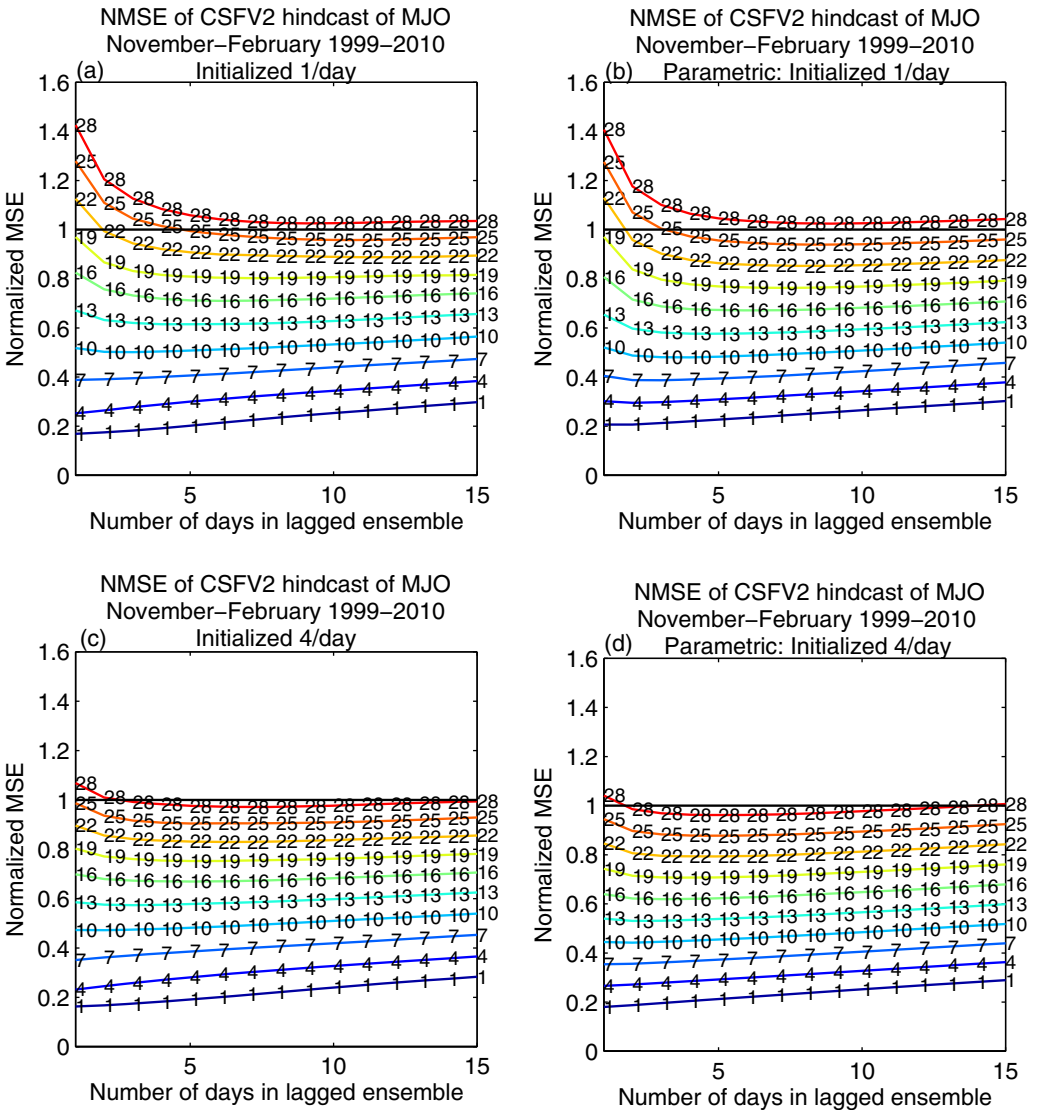

Figure 3. Normalized MSE of the boreal winter (1 November to 28 February) MJO forecast as function of lagged ensemble size (horizontal axis) and lead (colored curves—the number denotes the forecast lead in days) in the CFSv2. The MSE is computed in terms of the standard *Wheeler and Hendon* [2004] RMM1/RMM2 indices. (a) Normalized MSE for MJO forecast from the 0Z initialization of CFSv2. (b) Empirically derived normalized MSE computed using the fit shown in Figure 1. (c) Normalized MSE for MJO forecast when 0Z, 6Z, 12Z, and 18Z initializations of CFSv2 are used. (d) Empirically derived normalized MSE computed using the fit shown in Figure 1 interpolated to include four separate initializations.

empirical models capture the lead-dependent structure of the error covariance matrices for the RMM1 and RMM2 indices. Because the sample size in our study is large (e.g., a total of 58,080 forecasts used to estimate an eight-parameter model), it is unlikely that overfitting or statistical significance is a major concern. Therefore, we have not performed cross validation—all available data are used to estimate the covariance matrix. If we were to subset the data to identify flow-dependent skill, such as estimating the MSE separately for events of a certain strength or phase, then the risk of overfitting and lack of significance increases. We argue that the holistic approach pursued here, in which the entire cross-lead covariance matrix is fitted over all lead times, is much more attractive than estimating covariances separately. For instance, we note in section 4 that the optimal weights can be negative when sample covariances are used directly. Also, the covariances at different leads and ensemble sizes smoothly connect with those at other leads and ensemble sizes, which would not be the case when sample covariances are estimated separately.

We note that applying this methodology to other variables and different forecast systems might require identification and fitting of a new parametric model. Furthermore, caution should be

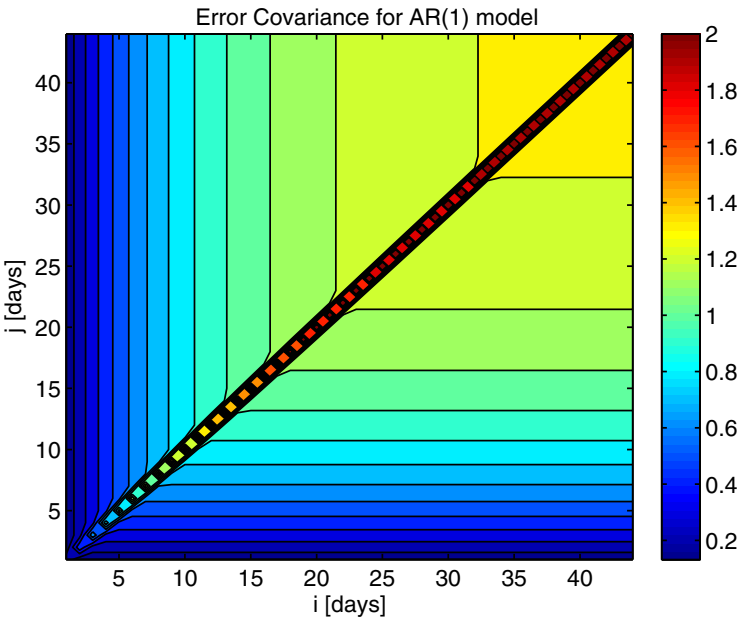

Figure 4. Cross-lead error covariance matrix for an AR(1) process, with $\phi = 0.95$. See Appendix for details.

exercised when fitting nonlinear functions, such as those fit in this study, since some trial-and-error may be required when specifying the initial guesses for the parameters.

3.3. Estimates of MSE Based on Parametric Model

Having parameterized the cross-lead error covariance matrix, we can substitute the parameterized values into (5) to compute the MSE of a lagged ensemble. The resulting values are shown in Figure 3b for the once-per-day initializations. Comparison between Figures 3a and 3b indicates good agreement, especially at longer lead times. At short leads, the empirical estimates show more curvature with lead time than that found directly from CFSv2 data.

Because the parametric model is a piece-wise continuous function of time, it can be evaluated at arbitrary times. This property allows us to estimate the MSE at times other than those used to fit the model. As an example, we estimate the MSE of an ensemble forecast initialized every 6 h using the parametric equations with parameters estimated from forecasts initialized every 24 h. The MSE extrapolated from our parametric model for four-per-day initialized forecasts is shown in Figure 3d, and is in remarkable agreement with the actual MSE estimated from four-per-day initialized CFSv2 forecasts shown in Figure 3c. These results demonstrate that forecasts initialized at one frequency can be used to accurately predict the skill at different (here, higher) initialization frequencies. This capability allows us to estimate the sensitivity of skill to changes in initialization frequency, lead time, and ensemble size without the need of costly numerical experiments.

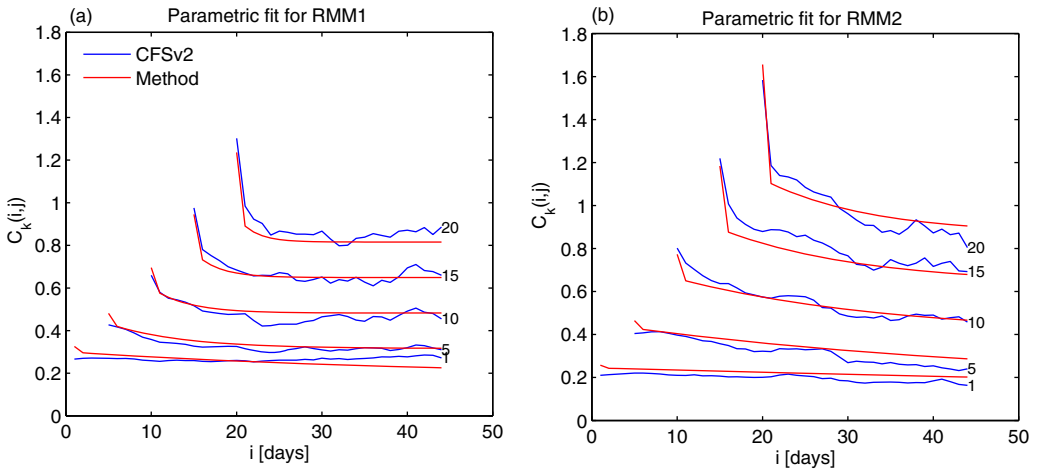

Figure 5. Cross sections of the error covariance matrices for (a) RMM1 and (b) RMM2. For each lead time, indicated by the number next to each curve, the error covariance matrix is shown along a row, starting from the diagonal element. The values estimated from CFSv2 are shown in blue while those from the parametric model are shown in red.

It is instructive to consider the above extrapolation procedure in terms of the schematic Figure 2. The original matrix elements (i.e., the dots and crosses) are separated by 24 h. To estimate MSE for 6 h initialization frequency, the parametric equations can be evaluated at 6 h intervals to obtain values *between* the original matrix elements. We can visualize this procedure as averaging over the appropriate *fractional* elements within a square. Equivalently, we could derive a new covariance matrix in which elements are separated by 6 h, and then average over the elements in the "new" square, which would look equivalent to Figure 2. For fixed ensemble size L, decreasing the time between initial starts corresponds to decreasing the sides of the averaging square. In the next section, we will consider the implications of approaching the limit of infinitesimal time between initial starts, in which case the square shrinks to a point and the resulting MSE approximates that of a burst ensemble.

3.4. MSE of an Infinite Burst Ensemble

Our parametric model can also be used to estimate the MSE of a "burst" ensemble, a fact that may seem surprising given that only one forecast per initialization time is available from the CFSv2 hindcasts. A burst ensemble is characterized by initial errors that are approximately equal, in contrast to a lagged ensemble in which the amplitude of the initial errors grow with ensemble size. Thus, a burst ensemble can be approximated by considering ensemble members initialized an infinitesimal time step apart (e.g., 1 s apart). Note that we are not suggesting that a real forecast system should create a burst ensemble in this way, rather, we are simply noting that an ensemble created this way will have error characteristics similar to that of a burst ensemble. Accordingly, consider the MSE of the lagged ensemble (5) in the limit $\delta t \to 0$, in which case L is equal to the burst ensemble size. Because the parametric model is discontinuous near the diagonal, this limit must be taken separately for diagonal and off-diagonal elements, yielding

$$\lim_{\delta t \to 0} MSE_k(L,\tau) = \frac{1}{L}\hat{C}(\tau,\tau) + (a_\tau + b_\tau)\left(1 - \frac{1}{L}\right). \tag{11}$$

Equivalently, from (10),

$$\lim_{\delta t \to 0} MSE_k(L,\tau) = \frac{1}{L}\hat{C}_{residual}(\tau) + (a_\tau + b_\tau). \tag{12}$$

Thus, the MSE decreases with ensemble size (as expected) until it saturates at $(a_\tau + b_\tau)$, which is the diagonal element extrapolated from the off-diagonal elements. This gives a physical interpretation of these extrapolated quantities: the off-diagonal elements extrapolated to the diagonal give the MSE of an infinite ensemble and defines the upper bound on skill at the specified lead time τ. Also, $\hat{C}_{residual}$ is proportional to the rate of reduction in MSE per ensemble member. Note that if the parameterized model had not included a discontinuity at the diagonal, then the limit would have reduced to the diagonal element, a clearly incorrect result because the diagonal element is the mean square error of a single member. This argument can be reversed to prove that there must be a discontinuity at the diagonal.

The MSE of an arbitrary mix of burst ensembles and lagged ensembles can be computed straightforwardly by generalizing the ensemble mean forecast (1) to include a time step δt_l that depends on ensemble member:

$$\tilde{f}_k(v, v-\tau, L) = \frac{1}{L}\sum_{l=0}^{L-1} f_k(v, v-\tau-\delta t_l). \tag{13}$$

In the previous examples, for instance, a one-per-day initialization corresponds to $\delta t_l = l$ while a four-per-day initialization corresponds to $\delta t_l = l/4$. In contrast, a single burst corresponds to $\delta t_l = \Delta l$, where Δ is a small number, whereas a four member burst every 0Z corresponds to

$$\delta t_l = [\,0 \quad \Delta \quad 2\Delta \quad 3\Delta \quad 1 \quad 1+\Delta \quad 1+2\Delta \quad 1+3\Delta \quad 2 \quad \ldots\,]. \tag{14}$$

The MSE of the generalized ensemble is then

$$MSE_k(L,\tau) = \frac{1}{L^2}\sum_{m=1}^{L-1}\sum_{n=1}^{L-1} C_k(\tau+\delta t_m, \tau+\delta t_n). \tag{15}$$

In this way, the MSE of even an inhomogeneous mix of burst and lagged ensembles can be estimated simply by specifying the appropriate time steps δt_l.

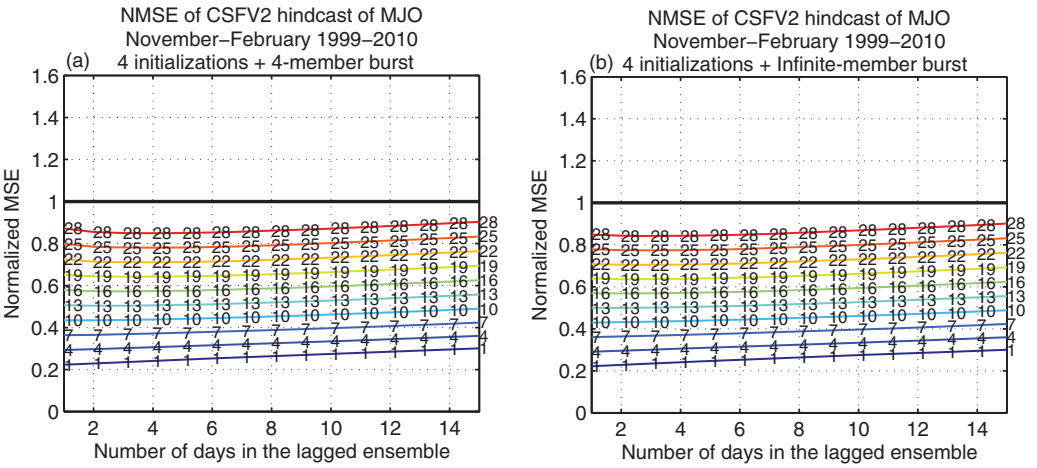

Figure 6. Empirically derived normalized MSE for CFSv2 hindcasts of MJO during boreal winter (1 November to 28 February) as a function of lagged ensemble size (horizontal axis) and lead (colored curves—the number denotes forecast lead in days). MSE is found using the parametric model fitted to error covariance matrices of RMM1 and RMM2 for hindcasts initialized 1 day apart and then interpolated to (a) four burst initializations and (b) for infinite burst for each 4 day initializations.

To illustrate the above procedure, we compute the MSE of the operational NCEP forecasts, which consist of four burst ensemble members 4 times a day—at 0Z, 6Z, 12Z, and 18Z—giving a total of 16 members per day. This protocol corresponds to choosing time steps based on a pattern similar to that illustrated in (14). The resulting NMSE estimated for the NCEP operational configuration (four initializations per day—four bursts) as a function of lead time and lagged ensemble size are shown in Figure 6a. Comparing these results with the NMSE from a four-per-day forecast ensemble (Figure 3d) shows only marginal improvement in MJO forecast at long lead times when the ensemble size is increased from beyond 4. To illustrate the impacts of a burst ensemble on the forecast skill, we show the NMSE for an infinite burst ensembles with four initializations per day in Figure 6b. Comparing Figures 6a and 6b, it is clear that increasing the burst ensemble beyond 4 does not improve the forecast skill.

3.5. Applicability to Other Forecast Systems

Many forecast systems initialize subseasonal forecasts every 5 days rather than every day [*Vitart et al*., in press]. To address the question as to whether such forecasts are sufficient to estimate an accurate parametric model for the cross-lead error covariance matrix, we subsample the 0Z daily forecast of the MJO indices to mimic a 5 day initialization frequency and repeat our analysis using empirical fits based on the 5 day data to estimate 1 day error growth. The covariance matrices estimated directly from the CFSv2 hindcasts are shown in Figures 7a and 7b. Decreasing the initialization frequency blurs out the fine scale features of the error covariance matrices, but some of the general characteristics remain intact. In particular, at the shorter leads, the off-diagonal error decays more slowly away from the diagonal than at higher leads, and the error grows most rapidly along the diagonal. The empirical fits of the 5 day error covariances shown in Figures 7c and 7d capture these main features remarkably well. The empirical formulas recovered from the 5 day fit (Figures 7c and 7d) are extrapolated to a 1 day initialization frequency, the resulting error covariance matrices for RMM1 and RMM2 are shown in Figures 7e and 7f. Comparing these estimates with error covariance of the 1 day forecast (Figures 1a and 1b), it is clear that the empirical models based on a 5 day initialization frequency cannot reproduce the tilt of the contours near the diagonal. Rather, the 1 day estimates of the off-diagonal error growth at short leads are very reminiscent of the structure found for the AR(1) model (see Appendix). At longer leads, the off-diagonal error growth increases away from the diagonal, which is the opposite of what was found in the CFSv2 data (see Figure 1).

While the 1 day estimates fail to capture the complete structure of the error covariance matrices, the parametric descriptions do broadly capture the lead-dependent error growth for both indices, and these estimates, while less accurate, may still be useful for approximating the influence of initialization frequency on forecast error. To test this, we calculate the normalized MSE for the 1 day error estimate. The

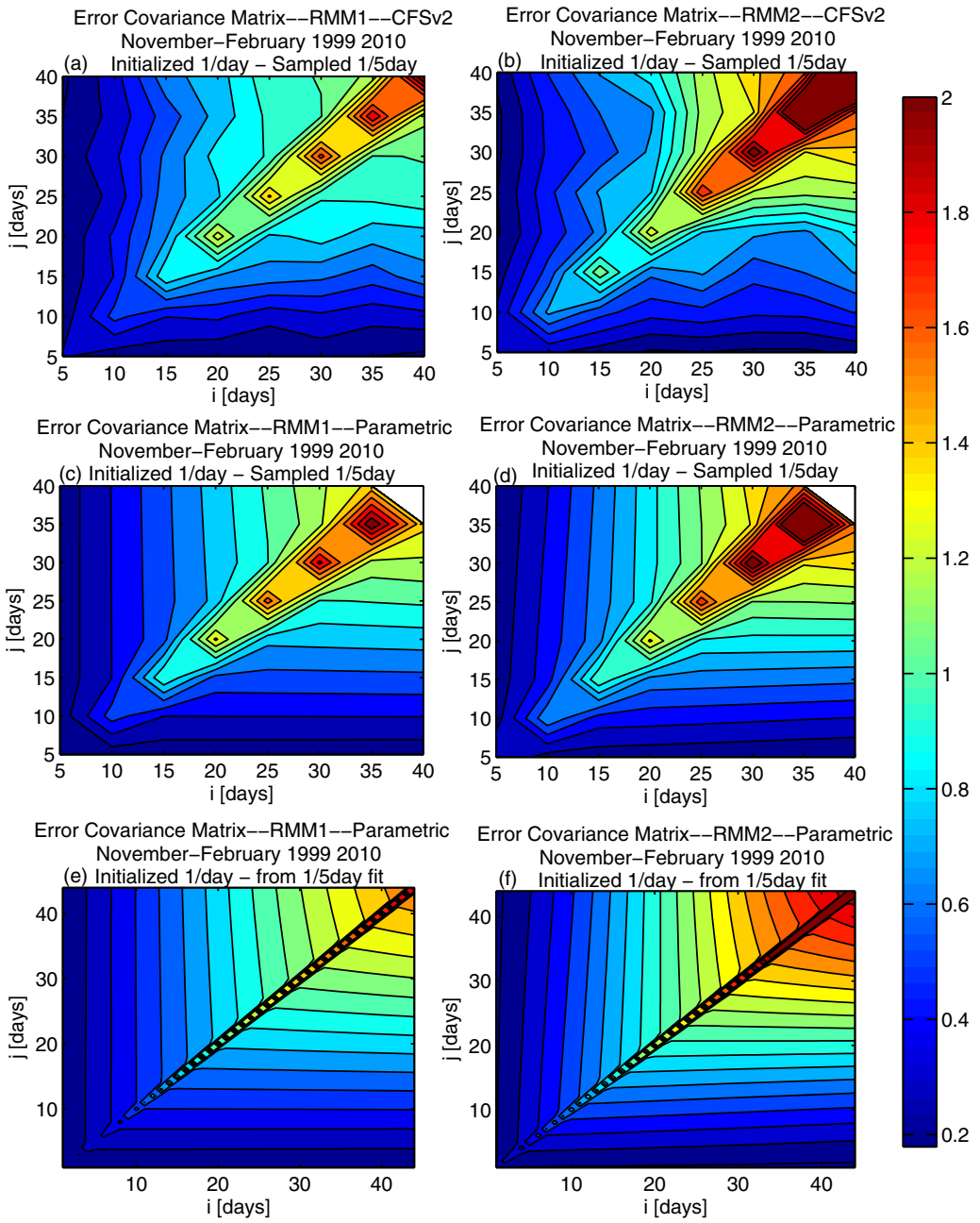

Figure 7. Cross-lead error covariance matrices for CFSv2 hindcasts of MJO during boreal winter (1 November to 28 February) for a 5 day initialization frequency. Figures 7a and 7b show the cross-lead error covariance matrices for RMM1 and RMM2 estimated directly from CFSv2 hindcasts. Figures 7c and 7d are covariance matrices derived from the parametric model fitted to the 5 day subsampled data. Figures 7e and 7f are covariance matrices derived from the same parametric model but evaluated at 1 day intervals.

resulting normalized MSE curves, shown in Figure 8, accurately estimates the magnitude of the normalized MSE at leads less than 10 days. However, the optimal ensemble size is not correctly predicted: at these short leads, the estimated optimal ensemble size is 2 days, whereas the direct estimates suggest an optimal ensemble size of 1 day (see Figure 3). At longer leads, the magnitude of normalized MSE is underestimated, but the optimal lagged ensemble size of 5 days agrees with the direct estimates (see Figure 1).

These results show that if the initialization is too infrequent, the detailed structure of the error growth is lost. Further, we show that these finer scale features of the error covariance matrices significantly impact

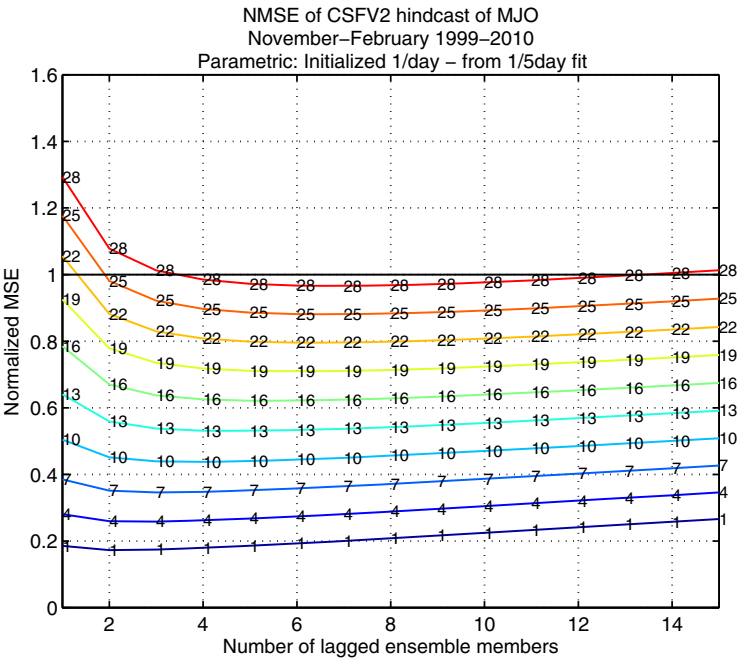

Figure 8. MSE of MJO hindcasts derived from the parametric model, as in Figures 3b, except that the parametric model is fitted to the 5 day error covariance matrices of RMM1/RMM2 (Figures 7c and 7d) and then extrapolated to 1 day initialization frequency (Figures 7e and 7f). The MSE is shown as function of lagged ensemble size (horizontal axis) and lead (colored curves—the number denotes the forecast lead in days).

the MSE magnitude and optimal lagged ensemble size. Specific to the CFSv2, we find that a 5 day initialization frequency is insufficient for capturing the forecast error growth.

4. Weighted Lagged Ensemble

So far, each forecast has been weighted equally, regardless of lag. The assumption of equal weights is clearly not optimal since different members have different lead times. The optimal weights that minimize the MSE depend on precisely the same cross-lead covariance matrix as does the MSE itself. Therefore, once this covariance matrix is known, the optimal weights can be determined. Let us define a weighted lagged ensemble for the forecast anomaly of the kth MJO index at verification time v and lead time τ as

$$\tilde{f}'_k(v, v-\tau, L) = \sum_{l=0}^{L-1} w_l f_k(v, v-\tau-\delta t_l), \tag{16}$$

where w_l are the weights. If the weights are constrained to sum to one, the mean square error is

$$MSE_k(L,\tau) = \langle (\tilde{f}'_k(v, v-\tau, L) - o(v))^2 \rangle. \tag{17}$$

If the weights are collected into the vector $\mathbf{w}_k$, then the MSE can be written as

$$MSE_k(L,\tau) = \mathbf{w}_k^T \mathbf{C}_k \mathbf{w}_k, \tag{18}$$

where superscript T denotes the transpose operation. We recover (5) after substituting $w_l = 1/L$. To find the *optimal* weights, we want to minimize the mean square error subject to the constraint that weights sum to one. To do so, we use the method of Lagrange multipliers to construct the objective function

$$\Theta = MSE_k(L,\tau) + \lambda \mathbf{z}^T \mathbf{w}_k, \tag{19}$$

where λ is the Lagrange multiplier and $\mathbf{z}$ is a vector of all ones. Differentiating (19) with respect to $\mathbf{w}$ gives

$$\frac{\partial \Theta}{\partial \mathbf{w}} = 2\mathbf{C}_k \mathbf{w} + \lambda \mathbf{z}. \tag{20}$$

Setting (20) to zero, and solving for the weights, such that the multiplier λ satisfies the constraint that the weights sum to one, gives

$$\mathbf{w}_k = \frac{\mathbf{C}_k^{-1}\mathbf{z}}{\mathbf{z}^T \mathbf{C}_k^{-1} \mathbf{z}}. \tag{21}$$

From (21), it is clear that the optimal weights for a given lead and ensemble size depend only upon the cross-lead error covariance matrix.

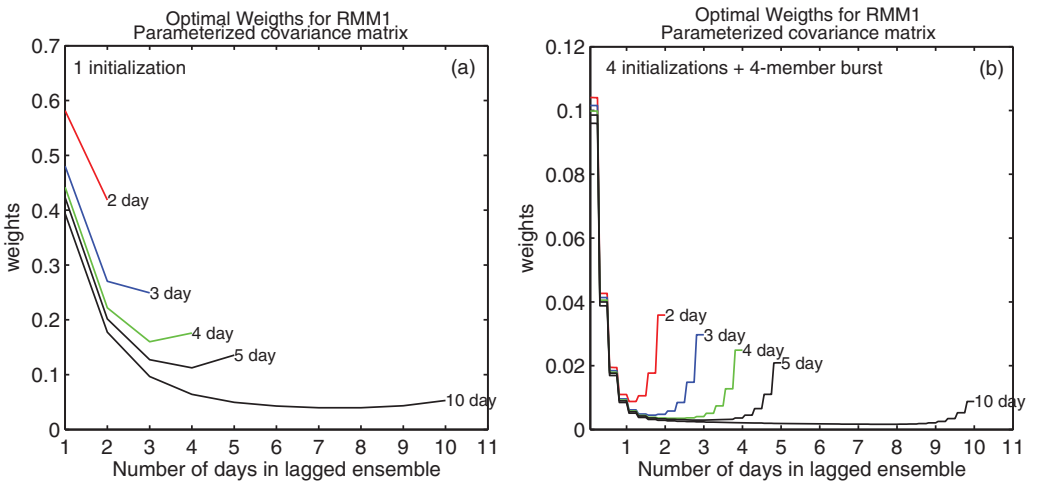

Figure 9. Optimal weights computed from the parametric cross-lead error covariance matrices of RMM1 for an initialization frequency of (a) once per day and (b) four initializations per day, with four burst members per initialization. Optimal weights are found according to equation (21). The results are shown for a 10 day lead. Each curves show the optimal weights for a given lagged ensemble size, measured in units of days, and listed on the right.

Using (21), we find the optimal weights for CFSv2 hindcasts of RMM1 based on parametric models of the cross-lead error covariance when the initialization frequencies are specified as once per day and 4 times per day, with four bursts per initialization. The optimal weights are found for different ensemble sizes at a 10 day lead for each forecast configuration. Note we selected a 10 day lead for illustrative purposes. Importantly, the weights determined from the sample covariance matrix are negative for some leads, which is often seen as undesirable when evaluating skill in a single model, since it implies that some ensemble members are not correctly forecasting the signal. In contrast, the weights determined from the parameterized covariance matrix were always positive. As such, we show only the weights derived from the parametric models of the cross-lead error covariance matrix. Lastly note that only the optimal weights for RMM1 are shown, since the weights for RMM2 are qualitatively similar.

The optimal weights for RMM1 with initialization frequencies of once per day and 4 times per day, with four bursts per initialization are shown in Figures 9a and 9b, respectively. The curves in each figure give the optimal weights for different lagged ensemble sizes in increments of a day, with the total size indicated by the text to the right of each curve. When an initialization frequency of once per day is used, the optimal weights shown in Figure 9a are largest when the fewest number of lagged ensemble members are used, with weights decreasing as the number of days in lagged ensemble grows. The general decrease in the optimal weights as a function of lagged ensemble size indicates that most of the forecast skill comes from the forecast initialized closest to the target date. The optimal weights found for four initializations per day—four bursts, shown in Figure 9b, are similarly characterized by a general decrease in weights with lagged ensemble size. The inclusion of burst members introduces stepwise changes in the weights, with each burst member receiving equally weighting.

A curious result is that the weights tend to increase at the last ensemble member. This result is counterintuitive: one would expect the ensemble weights to decrease with increasing ensemble size, since the inclusion of each additional lagged ensemble member introduces older information into the forecast, thereby degrading skill. However, since the weights are derived from a parameterized covariance matrix, this behavior cannot be dismissed as an artifact of sampling error. To try to understand the source of this behavior in the weights, we perform a set of sensitivity experiments, wherein the optimal weights are found for error covariance matrices with distinct structure. Specifically, we find the optimal weights assuming the error covariance is defined as one of the following: (i) AR(1) process (see Figure 4 and equation (A16)). (ii) A covariance matrix that is AR(1)-like, except that the diagonal elements are replaced by a sigmoid function. For consistency, the off-diagonal elements are defined to be half the value at the appropriate diagonal element. (iii) A covariance matrix that is like that of (ii), except that the off-diagonal elements decay exponentially away from the diagonal elements. We use the parametric fit recovered for the RMM1 in CFSv2 to

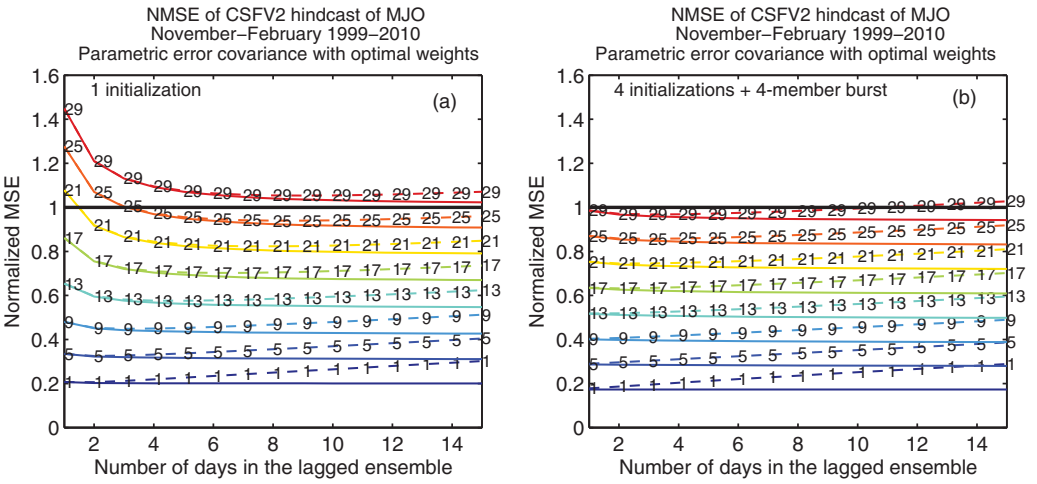

Figure 10. Normalized MSE of the optimally weighted lagged ensemble (solid) and for the equally weighted lagged ensemble (dashed) derived from the parameterized error covariance matrix of RMM1/RMM2 for an initialization frequency of (a) 1 per day and (b) four initializations times per day, with four burst members per initialization. The MSE is shown as a function of lagged ensemble size (horizontal axis) and lead (colored curves—the number denotes the forecast lead in days).

represent these features (see Figure 5, red curve—off-diagonal only). These sensitivity experiments indicate that the decay of off-diagonal error growth is the primary cause of the increase in weight for the oldest ensemble member. This same conclusion holds when burst ensemble members are included. The physical explanation for this behavior does not seem to be simple.

Substituting the optimized weights and the corresponding empirically based cross-lead error covariance matrices for RMM1 and RMM2 into (18), we examine the impact of the optimal weighting on the forecast MSE for forecasts initialized once per day and 4 times per day, with four burst members per initialization. The resulting MSE curves for the optimally weighted forecast are shown as the solid curves in Figures 10a and 10b, respectively. For comparison, the dashed curves denote the corresponding MSE for forecasts that are not weighted. Comparing the solid curves with the dashed, we see that the optimal weights have the greatest impact on the magnitude of the MSE for increasing ensemble sizes. Moreover, we see that the weights are also dependent on lead. For example, considering forecasts initialized once per day, we see from Figure 10a that the MSE of the weights and nonweighted forecast diverge quite rapidly at short leads, which indicates that only the first few ensemble members contribute to forecast skill. At longer lead times, the MSE curves diverge when a larger ensemble size is used. Similar behavior is found when optimal weighting is applied to forecasts initialization 4 times per day with four burst members per initialization, as shown in Figure 10b. This suggests that magnitude of the weights can also serve as a useful benchmark for determining the optimal lagged ensemble. Moreover, it is surprising that optimal and equal weights identify nearly the same optimal ensemble size.

In this section, we have shown that the optimal weights needed to produce the most skillful forecast depends only on the cross-lead error covariance matrix. Analytic functions of RMM1 and RMM2 cross-lead error covariance matrices skillfully estimate the optimal weights for ensemble forecasts of both indices. We find that the largest weights are associated with lagged ensemble members closest to forecast target date. Essentially, optimal weighting of the ensembles reduces the lagged ensemble size, since the older forecasts are assigned weights close to zero. Generally, weighting the ensembles has the most impact on MJO forecast error at longer lead times. While not addressed here, it is likely that optimal weighting based on forecast MSE will improve the skill of probabilistic forecasts as well.

5. Conclusions and Discussion

This paper describes a general methodology for extrapolating the MSE of a lagged ensemble to initialization frequencies other than those available in a given hindcast data set. This methodology exploits the fact that the MSE of a lagged ensemble depends on a quantity called the cross-lead error covariance matrix. We propose a parametric model for this covariance matrix that captures the dominant lead time dependencies of

the covariances and can be evaluated at lead times different from those available in the data set used to estimate the original covariance matrix.

To demonstrate the utility of the proposed methodology, we first examined the cross-lead error covariance matrices estimated from CFSv2 daily hindcasts of MJO indices RMM1 and RMM2 for hindcasts initialized at 0Z. Qualitatively, the structure for both indices is reminiscent of that for an AR1 model: the off-diagonal elements tend to depend on the smallest lead time and a sharp discontinuity between the diagonal and off-diagonal elements occurs at long leads. However, there are important differences: the discontinuity between diagonal and off-diagonal elements is weak or nonexistent at short leads. The covariance matrices for both indices are used to compute the normalized MSE as a function of lead time and lagged ensemble size. We find that at leads less than a week the MJO forecast error is a minimum when only one lagged ensemble member is used, indicating a lagged ensemble does not improve skill at short leads. At long lead times ($>$7 days), the forecast error reaches a minimum when approximately five lagged members are used.

Certain structures in the cross-lead error covariance matrix vary smoothly as a function of lead time, making it possible to parameterize the error covariances in terms of analytic functions. The off-diagonal elements of the matrix were modeled by an exponential decaying function in which the parameters change linearly as a function of lead time. The along-diagonal elements of the matrix were constrained by the off-diagonal fits and modeled by the logistic equation. Since the parametric equations are continuous, the parametric model can be used to infer skill for arbitrary ensemble size and initialization frequency. As an example, we fitted the parametric model using CFSv2 hindcasts initialized once-per-day at 0Z, and then used the resulting fit to estimate the skill of CFSv2 hindcasts initialized four-times-per-day. The skill inferred from the parametric model compared well with the skill estimated directly from CFSv2 hindcasts initialized at 0Z, 6Z, 12Z, and 18Z. Operationally, NCEP uses four initializations per day for their subseasonal forecasts, with four bursts per initialization, yielding a total of 16 ensemble members per day. We adapt our methodology to account for the inclusion of burst ensembles. In terms of the MSE, we find that increasing the ensemble size from 4 to 16 produces marginal improvement in forecast skill at long leads and very little improvement at short leads. By extension, we show that increasing the number of ensembles beyond 16 members provides negligible benefit for the subseasonal forecast at any lead time.

A number of operational centers produce subseasonal forecasts using weekly or monthly initializations. Given the competing need for computational resources, it may be desirable to initialize subseasonal forecast less frequently while still maintaining a certain threshold of forecast skill. We test the impacts of decreasing the initialization frequency on forecast skill by subsampling the 1 day initialization frequency to mimic forecast data with a 5 day initialization frequency. We then fit an empirical model to the 5 day error covariance matrices and extrapolate back to the 1 day frequency. We find that the 5 day initialization is too coarse for MJO forecast in the CFSv2, and relevant structures of the error covariance matrices are lost. With that said, the successful implementation of this methodology is dependent on hindcast data being generated with frequent enough initializations to capture the error growth of the specified predicted variable. As such, the initialization frequency needed to capture the structure of the forecast errors is likely to depend upon the variable being predicted and the forecast model.

Lastly, we show that the MJO forecast error can be further reduced by finding the optimally weighted ensemble mean. As it turns out the optimal weights depend only on the cross-lead error covariance matrix and can easily be computed from the analytic functions described above. Applying this method to parametric estimates of the RMM1 and RMM2 error covariance matrices, we show that the largest weights are given to lagged ensemble member with the shortest lead time. As a result of optimal weighting, the magnitude of the forecast error is reduced at leads greater than $>$10 days.

Appendix A: Errors of an AR(1) Model

The properties of the cross-lagged error covariance matrix are not evident consequences of any formal theory. To gain insight into the properties of this matrix, we compute the cross-lagged error covariance of a first-order autoregressive process. As before, ϵ refers to the forecast error and v the verification day. An AR1 process is governed by a model of the form

$$X_v = \phi\ X_{v-1} + W_v, \tag{A1}$$

where ϕ is the AR(1) parameter and W_v is the white noise with zero mean and variance σ_W^2. It is well known that the autocovariance function of the stationary solution is

$$\langle X_{v+i} X_v \rangle = \sigma_X^2 \phi^{|i|}, \tag{A2}$$

where σ_X^2 is the stationary variance of X. For any realization of white noise, the solution of (A1) can be found by recursive evaluation:

$$X_v = \phi^i X_{v-i} + \sum_{k=0}^{i-1} \phi^k W_{v-k} \quad for \quad i \geq 0. \tag{A3}$$

To mimic the single forecasts of the CFSv2, we assume the forecast for X_v is drawn from (A3). The error of this forecast is the difference between (A3) and X_v:

$$\epsilon(v, v-i) = \phi^i X_{v-i} + \sum_{k=0}^{i-1} \phi^k W_{v-k} - X_v. \tag{A4}$$

It is important to recognize that the forecast noise associated with initial condition X_{v-i} is independent of that with initial condition X_{v-j}. Therefore, when computing the cross-lead error covariance matrix (6) for off-diagonal elements, the noise terms vanish, yielding:

$$C(i,j) = \langle \left(\phi^i X_{v-i} - X_v\right)\left(\phi^j X_{v-j} - X_v\right) \rangle \tag{A5}$$

$$= \phi^{i+j} \langle X_{v-i} X_{v-j} \rangle - \phi^i \langle X_{v-i} X_v \rangle - \phi^j \langle X_{v-j} X_v \rangle + \langle X_v^2 \rangle \tag{A6}$$

$$= \sigma_X^2 \left(\phi^{i+j} \phi^{|i-j|} - \phi^{2i} - \phi^{2j} + 1\right) \tag{A7}$$

$$= \sigma_X^2 \left(1 - \phi^{2\min[i,j]}\right) \quad for \quad i \neq j. \tag{A8}$$

Note that this expression is identical to that which would be obtained for an infinite ensemble, since the noise of an infinite ensemble vanishes. Thus, the above expression holds for both diagonal and off-diagonal elements for an infinite ensemble. In contrast, the diagonal elements for a single forecast is given by

$$C(i,i) = \left\langle \left(\phi^i X_{v-i} - X_v + \sum_{k=0}^{i-1} \phi^k W_{v-k} \right)^2 \right\rangle \tag{A9}$$

$$= \langle \left(\phi^i X_{v-i} - X_v\right)^2 \rangle + \left\langle \left(\sum_{k=0}^{i-1} \phi^k W_{v-k} \right)^2 \right\rangle \tag{A10}$$

$$= \sigma_X^2 \left(1 - \phi^{2i}\right) + \sigma_W^2 \sum_{k=0}^{i-1} \phi^{2k} \tag{A11}$$

$$= \sigma_X^2 \left(1 - \phi^{2i}\right) + \sigma_W^2 \frac{1 - \phi^{2i}}{1 - \phi^2} \tag{A12}$$

$$= \sigma_X^2 \left(1 - \phi^{2i}\right) + \sigma_X^2 \left(1 - \phi^{2i}\right) \tag{A13}$$

$$= 2\sigma_X^2 \left(1 - \phi^{2i}\right), \tag{A14}$$

where we have used standard summation formulas for geometric series and the fact that the stationary variance of X is

$$\sigma_X^2 = \frac{\sigma_W^2}{1 - \phi^2}. \tag{A15}$$

Combining the above solutions yield

$$C(i,j) = \sigma_X^2 (1 + \delta(i,j)) \left(1 - \phi^{2\min[i,j]}\right). \tag{A16}$$

where $\delta(i,j)$ is the delta function (whose value vanishes when $i \neq j$ and equals one when $i = j$). The above result shows that the diagonal element is *twice* the off-diagonal element—there exists a discontinuity between the diagonal and neighboring off-diagonal elements. This discontinuity arises from the fact that a single forecast contains noise that is uncorrelated for different lead times—there would be no discontinuity if the forecast contained no noise, or equivalently if the forecast was based on an infinite ensemble.

The error variance, given by the diagonal (A14), increases with lead and saturates to $2\sigma_X^2$, which is the variance we would expect for a no-skill forecast. According to (A8), the off-diagonal error for a given lead (i) depends only on the minimum of (i, j) and increases with lead. As an example, the lagged error covariance for AR(1) model with $\phi = 0.95$ is shown in Figure 4.

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), the National Aeronautics and Space Administration (NNX14AM19G), and 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. CFSv2 data can be found http://nomads.ncdc.noaa.gov/data/cfsr-hpr-ts45/. Observed RMM indices were obtained from Dr. Matthew Wheeler and can be found at http://poama.bom.gov.au/project/maproom/RMM/.

References

Chen, M., W. Wang, and A. Kumar (2013), Lagged ensembles, forecast configuration, and seasonal predictions, *Mon. Weather Rev.*, *141*(10), 3477–3497, doi:10.1175/MWR-D-12-00184.1.

Gottschalck, J., et al. (2010), A framework for assessing operational Madden–Julian Oscillation forecasts: A CLIVAR MJO Working Group Project, *Bull. Am. Meteorol. Soc.*, *91*(9), 1247–1258, doi:10.1175/2010BAMS2816.1.

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

National Academies of Sciences, Engineering, and Medicine (2016), *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*, The Natl. Acad. Press, Washington, D. C., doi:10.17226/21873.

Saha, S., et al. (2013), The NCEP climate forecast system version 2, *J. Clim.*, *27*(6), 2185– 2208.

Vitart, F., et al., The sub-seasonal to seasonal prediction (S2S) project database, *Bull. Am. Meteorol. Soc.*, doi:10.1175/BAMS-D-16-0017.1, in press.

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