AGU PUBLICATIONS

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

10.1002/2017MS001128

Key Points:
- Weights for an optimal weighted-average lagged ensemble are examined analytically
- Weights are not always positive nor do they always decay monotonically with lead time
- The reason for this surprising and counter-intuitive behavior is explained

Supporting Information:
- Supporting Information S1

Correspondence to:
T. DelSole,
tdelsole@gmu.edu

Citation:
DelSole, T., Trenary, L., & 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

Received 22 JUL 2017
Accepted 1 NOV 2017
Accepted article online 10 NOV 2017
Published online 29 NOV 2017

The Weighted-Average Lagged Ensemble

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

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

Abstract A lagged ensemble is an ensemble of forecasts from the same model initialized at different times but verifying at the same time. The skill of a lagged ensemble mean can be improved by assigning weights to different forecasts in such a way as to maximize skill. If the forecasts are bias corrected, then an unbiased weighted lagged ensemble requires the weights to sum to one. Such a scheme is called a weighted-average lagged ensemble. In the limit of uncorrelated errors, the optimal weights are positive and decay monotonically with lead time, so that the least skillful forecasts have the least weight. In more realistic applications, the optimal weights do not always behave this way. This paper presents a series of analytic examples designed to illuminate conditions under which the weights of an optimal weighted-average lagged ensemble become negative or depend nonmonotonically on lead time. It is shown that negative weights are most likely to occur when the errors grow rapidly and are highly correlated across lead time. The weights are most likely to behave nonmonotonically when the mean square error is approximately constant over the range forecasts included in the lagged ensemble. An extreme example of the latter behavior is presented in which the optimal weights vanish everywhere except at the shortest and longest lead times.

1. Introduction

The problem of combining forecasts into a superior forecast is of considerable interest at the present time. Many studies focus on combining forecasts from different models, yielding so-called multimodel ensembles. In this paper, we consider combining forecasts from the *same model* initialized at different times but verifying at the same time. Such an ensemble is called a lagged ensemble (Hoffman & Kalnay, 1983). A lagged ensemble is very attractive in operational forecasting because the members are generated routinely and therefore provide an ensemble prediction system with low peak computational cost. Whether a lagged ensemble improves forecast skill is not always clear, since some studies find positive impacts (Ben Bouallegue et al., 2013; Chen et al., 2013; Dalcher et al., 1988; Lu et al., 2007; Yuan et al., 2009) while others find negligible or negative impacts (Buizza, 2008). Nevertheless, because numerous studies do find benefits, and the computational savings associated with a lagged ensemble can be immense, further study of the lagged ensemble method is warranted.

An important aspect of a lagged ensemble is that different members have different skills because of the different lead times. Recent studies have attempted to improve the lagged ensemble by developing a *weighted* lagged ensemble, whereby different members are assigned different weights in such a way as to maximize skill. Whether unequal weights lead to significantly better forecasts than equal weighting is not always clear, especially in the context of multimodel combinations (DelSole et al., 2012). Also, these weights may vary in time (Raynaud et al., 2015) or be fixed (Lu et al., 2007), although Raynaud et al. (2015) found no noticeable gain due to allowing time-variations in the weights. Here we consider only weights that are fixed in time. A natural first guess, which is optimal when the errors are uncorrelated, is to assign a weight that is inversely proportional to the mean square error of each forecast. Such weights are positive and decay monotonically with lead time. In particular, the longest-lead forecast has the smallest weight, since it has the smallest skill, at least on average. However, this behavior is not always observed in practice. Lu et al. (2007) found that weights from multiple regression of hourly forecasts did not decay monotonically with lead time. Indeed, in some cases, the weight assigned to the largest-lead forecast was larger than the weights at shorter lead times. Trenary et al. (2017) found similar behavior in a weighted-average lagged ensemble for the Madden-Julian Oscillation. Of course, since the weights are derived from statistical methods, such results might be explained by random sampling error. Here we show that such behavior can occur even in the absence of sampling error and under a broad range of statistical models for the errors. In particular, we present a series of analytic examples designed to illuminate conditions under which the weights of an optimal lagged ensemble become negative or depend nonmonotonically with lead time. Indeed, we present a remarkable analytic example in which the weights of an optimal lagged ensemble vanish everywhere except those at the shortest and longest leads.

2. The Weighted Lagged Ensemble

The problem of combining forecasts has a long history stretching back at least to Bates and Granger (1969; see Clemen, 1989, for later references). The starting point of these methods is a set of forecasts. Let $f(t,\tau)$ denote a forecast of lead τ and verifying at time t; thus the initialization time is $t-\tau$. A weighted lagged ensemble is defined as

$$\bar{f}(t,\tau;L)=\sum_{l=1}^{L} w_l f(t,\tau+l-1), \tag{1}$$

where w_l are the weights and L is the number of forecasts being combined. The unit of time is chosen to coincide with the interval between initialization times, which is assumed to be uniform (as is common in operational forecasting). Although there exist methods for deriving time-dependent weights (e.g., Raynaud et al., 2015), we seek weights that are time independent.

Let the observation at the verification time be denoted o_t. We assume the forecasts are biased corrected, which means $E[f(t,\tau)-o_t]=0$, where $E[\cdot]$ is an expectation operator. This assumption is equivalent to

$$E[f(t,\tau)]=E[o_t] \quad \text{for all } t \text{ and } \tau. \tag{2}$$

It is desirable for the weighted lagged ensemble also to be unbiased. This condition requires

$$E\left[\bar{f}(t,\tau;L)\right]=\sum_{l=1}^{L} w_l E[f(t,\tau+l-1)]=\sum_{l=1}^{L} w_l E[o_t]=E[o_t]. \tag{3}$$

The last equality implies that the weights must sum to one:

$$\sum_{l=1}^{L} w_l=1. \tag{4}$$

Thus, if the forecasts are unbiased, then a weighted combination is unbiased if and only if the weights sum to one. We call a linear combination in which the weights sum to one a *weighted average*. We seek a weighted-average lagged ensemble that minimizes the mean square error. The error of the weighted-average lagged ensemble can be written as

$$\bar{\epsilon}(t,\tau;L)=\bar{f}(t,\tau;L)-o_t=\sum_{l=1}^{L} w_l \epsilon(t,\tau+l-1), \tag{5}$$

where $\epsilon(t,\tau)=f(t,\tau)-o_t$ is the error of forecast $f(t,\tau)$. The mean square error is

$$MSE(\tau,L)=E\left[\left(\sum_{l=1}^{L} w_l \epsilon(t,\tau+l-1)\right)^2\right]=\sum_{l=1}^{L}\sum_{l'=1}^{L} w_l w_{l'} \Sigma_{\tau+l-1,\tau+l'-1}, \tag{6}$$

where we define the cross-lead error covariance matrix:

$$\Sigma_{\tau_1,\tau_2}=E[\epsilon(t,\tau_1)\epsilon(t,\tau_2)]. \tag{7}$$

In deriving this expression, errors are assumed to be stationary and hence error covariances are independent of time t. The mean square error can be written in concise matrix form as

$$MSE(\tau,L)=\mathbf{w}^T \Sigma \mathbf{w}, \tag{8}$$

where $\mathbf{w}$ is the vector formed by the weights $w_1,\ldots,w_L$, superscript T denotes the transpose operation, and Σ is the covariance matrix obtained from the appropriate elements in the sum in (6). To minimize the mean square error subject to the constraint (4), we apply the method of Lagrange multipliers using the objective function:

$$\theta = MSE(\tau, L) - 2\lambda \mathbf{j}^T \mathbf{w}, \tag{9}$$

where $\mathbf{j}$ is an L-dimensional vector consisting of ones. Differentiating with respect to $\mathbf{w}$ gives

$$\frac{\partial \theta}{\partial \mathbf{w}} = 2\Sigma \mathbf{w} - 2\lambda \mathbf{j}. \tag{10}$$

Setting this to zero and simplifying gives

$$\Sigma \mathbf{w} = \lambda \mathbf{j}. \tag{11}$$

Inverting the covariance matrix, and choosing the multiplier λ to satisfy the constraint (4), gives the optimal weights:

$$\mathbf{w}_{\mathrm{opt}} = \frac{\Sigma^{-1}\mathbf{j}}{\mathbf{j}^T \Sigma^{-1} \mathbf{j}}. \tag{12}$$

If the errors are uncorrelated, then the error covariance matrix is diagonal and the weights can be solved explicitly in terms of the mean square error as

$$w_\tau \propto \frac{1}{MSE(\tau, 1)}. \tag{13}$$

In this case, the weights are positive and decay monotonically with lead time (provided *MSE* decays with lead time, as it usually does).

3. Simple Examples

To gain insight into how the optimal weights depend on the correlation between errors, consider the generic problem of choosing the value of w that minimizes the mean square of

$$z = wx + y, \tag{14}$$

where x and y are random variables with zero mean and unit variance. The mean square of z is

$$E[z^2] = w^2 + 1 + 2\rho_{XY} w, \tag{15}$$

where ρ_{XY} is the correlation between x and y. The minimum mean square is found by setting the derivative of (15) to zero and solving, which yields $w = -\rho_{XY}$. Thus, the sign of the optimal value of w is opposite to that of the correlation between x and y. This solution should make sense: regressing x out of y leaves the residual $y - \rho_{XY} x$, which no longer is correlated with x. The minimum mean square obtained with $w = -\rho_{XY}$ is

$$E[z^2]_{\mathrm{min}} = 1 - \rho_{XY}^2. \tag{16}$$

Thus, the greater the correlation ρ_{XY}, the smaller the mean square of the combination. This too should make sense—in the extreme case of a perfect correlation $\rho_{XY} = 1$, choosing $w = -1$ results in the difference $y - x$, which vanishes because x and y are perfectly correlated.

Now consider a *weighted average* of two random variables:

$$z = w\epsilon_1 + (1-w)\epsilon_2 = w(\epsilon_1 - \epsilon_2) + \epsilon_2. \tag{17}$$

Note that the far right-hand side of this equation is of the form (14), but the random variables are combined differently because of the constraint that the weights sum to one. Assuming ϵ_1 and ϵ_2 have zero mean and respective variances σ_1^2 and σ_2^2, the mean square of z is

$$E[z^2] = w^2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}) + \sigma_2^2 + 2w(\sigma_1\sigma_2\rho_{12} - \sigma_2^2), \tag{18}$$

where ρ_{12} is the correlation between ϵ_1 and ϵ_2. The minimum mean square occurs at

$$w_{opt} = \frac{\sigma_2^2 - \sigma_1\sigma_2\rho_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}. \tag{19}$$

This solution also could have been obtained from (12) using the covariance matrix for (ϵ_1, ϵ_2):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \Rightarrow \quad \Sigma^{-1} \propto \begin{pmatrix} \sigma_2^2 & -\rho_{12}\sigma_1\sigma_2 \\ -\rho_{12}\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}. \tag{20}$$

Importantly, at least one weight is negative if $w_{opt} < 0$ or $w_{opt} > 1$, which translates to

$$\frac{\sigma_2}{\sigma_1} < \rho_{12} \quad \text{or} \quad \frac{\sigma_1}{\sigma_2} < \rho_{12}. \tag{21}$$

This condition effectively states that the weights can become negative if the correlation ρ_{12} is sufficiently large *and* the difference in variance is sufficiently large. Conversely, the weights are always positive if the variances are equal (i.e., $\sigma_1 = \sigma_2$) or the errors are uncorrelated (i.e., $\rho_{12} = 0$). The negative weights arise because the sign of w_{opt} must be opposite to that of the covariance between $\epsilon_1 - \epsilon_2$ and ϵ_2 in (17), but the latter covariance becomes positive when $\sigma_1\rho_{12} > \sigma_2$. Similarly, we could have arbitrarily attached the weight w to ϵ_2 instead of ϵ_1 in (17), in which case the sign of w_{opt} must be opposite to that of the covariance between $\epsilon_2 - \epsilon_1$ and ϵ_1, which becomes positive when $\sigma_2\rho_{12} > \sigma_1$. This example reflects a theme that emerges in the next section, namely that the weights are most likely to become negative when the errors change rapidly with lead time and are highly correlated with each other.

Suppose now that the variances are equal (i.e., $\sigma_1 = \sigma_2$). Then, the minimum mean square of z (18) occurs at $w_{opt} = 1/2$, *regardless of the value of* ρ_{12}. This solution, in particular its independence from ρ_{12}, differs dramatically from the solution obtained by minimizing (15). Furthermore, the minimum mean square obtained from $w_{opt} = 1/2$ is

$$E[z^2]_{min} = \frac{1 + \rho_{12}}{2}. \tag{22}$$

This result also differs qualitatively from (16): the minimum mean square (16) *decreases* with correlation ρ_{XY}, whereas the minimum mean square (22) *increases* with correlation ρ_{12}. Thus, if the combination is of the form (14), then the mean square is minimized by choosing the two *most dependent* variables (i.e., variables with the highest correlation), whereas if the combination is a weighted average of the form (17), then the mean square is minimized by choosing the two *least dependent* variables (i.e., variables with smallest correlation). In essence, when the weights sum to one, the mean square is reduced when errors cancel, and this cancellation is most likely when the errors are least dependent.

4. More Complex Examples

In this section, we consider a weighted average of three or more random variables. A natural approach to exploring the relation between a weighted average and the error covariance function of the variables is to construct a mathematical model of a stochastic process and then compute the cross-lead error covariance function of the associated forecasts. This can be done analytically for a first order autoregressive process (AR1). We show in the appendix that the resulting weights are positive and decrease monotonically with lag. This result is consistent with the empirical results of Trenary et al. (2017).

The above result implies that to obtain weights that are negative or depend nonmonotonically on lead time, a more general model than an autoregressive model is necessary. To develop an intuitive understanding, it is desirable to explore the sensitivity of the weights to changes in the covariance matrix of the errors. However, it is unlikely that much insight will come from directly perturbing specific elements of the covariance matrix. For instance, the covariance matrix must always be positive definite, so perturbations applied to matrix elements must be constrained to ensure that the full matrix remains positive definite. Also, perturbing elements of a covariance matrix implies changing the growth and correlation properties of the errors. We would not want to perturb the covariance matrix in a way that generates kinks or other unusual characteristics in the error growth or correlation properties of the errors. A more instructive approach is to separate variance-correlation dependencies using the decomposition:

$$\Sigma = \mathbf{DRD}, \tag{23}$$

where $\mathbf{D}$ is a diagonal matrix with positive diagonal elements $d_1, \ldots, d_L$ equal to

$$d_\tau = \sqrt{MSE(\tau, 1)}, \tag{24}$$

and $\mathbf{R}$ is a correlation matrix. Any positive definite error covariance matrix can be decomposed in this form. This decomposition allows us to perturb the error growth and correlation properties separately. Inverting (11) to derive the weights yields

$$\mathbf{w} = \lambda \mathbf{D}^{-1}\mathbf{R}^{-1}\mathbf{D}^{-1}\mathbf{j}. \tag{25}$$

In general, mean square error increases with lead time. Thus, we consider three different error growth functions: logistic, linear, and constant. Also, the correlation between errors tends to decrease with the difference in lead times (Trenary et al., 2017). For simplicity, we assume that the error correlation depends *only* on the difference in lead times. Thus, $\mathbf{R}_{ij} = f(|i-j|)$. This assumption implies that the correlation matrix has the Toeplitz form:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{L-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{L-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{L-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{L-1} & \rho_{L-2} & \rho_{L-3} & \cdots & 1 \end{pmatrix}. \tag{26}$$

This matrix is uniquely specified by the parameters $\rho_1, \ldots, \rho_{L-1}$. We consider three types of decay functions for the correlation ρ_τ: constant, power law, and linear.

We emphasize that the above Toeplitz form is assumed merely for analytic simplicity and should not be construed as a preferred parameterization of $\mathbf{R}$. In practice, the eight-parameter model of Trenary et al. (2017) yields a more accurate parameterization. Our focus in this paper is conceptual understanding. As will become clear, the Toeplitz form leads to interesting insights into the weights of an optimal lagged ensemble.

4.1. Constant Off-Diagonal Elements

First, consider the case in which the errors have the same correlation regardless of lead time (an admittedly unrealistic example for a lagged ensemble). In this case, the correlation matrix has constant off-diagonal elements and hence takes the form:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}. \tag{27}$$

This covariance matrix can be written equivalently as

$$\mathbf{R} = (1-\rho)\mathbf{I} + \rho \mathbf{j}\mathbf{j}^T, \tag{28}$$

where $\mathbf{I}$ is the identity matrix. The inverse of (28) is

$$\mathbf{R}^{-1} = (1-\rho)^{-1}\left(\mathbf{I} - \gamma \mathbf{j}\mathbf{j}^T\right), \tag{29}$$

where $\gamma = \rho/(L\rho + \rho - 1)$. Substituting (29) into (25) yields

$$\mathbf{w} = \lambda(1-\rho)^{-1}\left(\mathbf{D}^{-2}\mathbf{j} - \gamma\left(\mathbf{j}^T\mathbf{D}^{-1}\mathbf{j}\right)\mathbf{D}^{-1}\mathbf{j}\right). \tag{30}$$

Using (24), the weights can be expressed alternatively as

$$w_k \propto \frac{1}{MSE_k} - \frac{\gamma \sum_k \frac{1}{\sqrt{MSE_k}}}{\sqrt{MSE_k}} \quad \text{for } \mathbf{R}=(27). \tag{31}$$

In the special case of uncorrelated errors, $\rho=0$, $\gamma=0$, and $w_k \propto 1/MSE_K$; that is, the weights are inversely related to MSE, as might be anticipated intuitively (or derived in some weighting schemes, as in Johnson & Swinbank, 2009). It proves convenient to define

$$x_k = \frac{1}{d_k} = \frac{1}{\sqrt{MSE_k}}. \tag{32}$$

Then the solution (31) can be expressed equivalently as

$$w_k = \lambda(1-\rho)^{-1}\left(x_k^2 - L\gamma\bar{x}x_k\right), \tag{33}$$

where $\bar{x}$ is the average of $x_1, \ldots, x_L$, hence $1/\bar{x}$ is the harmonic mean of the root-mean-square error. It is straightforward to show that if MSE_k is a monotonically increasing function of k, then the weights decrease monotonically with k. However, the weights can become negative if $x_k < L\gamma\bar{x}$; i.e., the weights become negative if the correlation ρ is sufficiently large and the mean square errors have sufficient spread, as found in the two-variable case in section 3.

4.2. Correlations Decay With Lead

Next we consider correlations that decay as a power law of the form:

$$\rho_\tau = \rho^\tau \quad \Rightarrow \quad R_{ij} = \rho^{|i-j|}, \tag{34}$$

where $|\rho| < 1$. This correlation function is illustrated in Figure 1a for three different values of ρ. Each function completely specifies the correlation matrix. We then compute the weights from (25) for the case in which mean square errors grow as a logistic function (Figures 1b and 1c), as a linear function (Figures 1d and 1e), and as a constant function (Figures 1f and 1g; the parameter values for error growth are not important). In the case of the logistic function (Figures 1b and 1c), the weights become negative for small lead times, but do not develop a U-shape at the tail. For linear error growth, the weights are positive and develop positive curvature at the tail, with the curvature increasing with increasing ρ (Figure 1e). Finally, for constant mean square error, the weights show a very significant curvature at the tail (Figure 1g).

To understand the above results, we take advantage of the fact that the correlation matrix formed from (34) has an analytic inverse (Dow, 2003), which is

$$\mathbf{R}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1+\rho^2 & -\rho & 0 \\ 0 & 0 & 0 & \ldots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \ldots & 0 & -\rho & 1 \end{pmatrix}. \tag{35}$$

Substituting (35) into (25), and using (32), yields

$$w_k = \lambda \begin{cases} x_1(x_1 - \rho x_2) & k=1 \\ x_k(-\rho x_{k-1} + (1+\rho^2)x_k - \rho x_{k+1}) & 1 < k < L \\ x_L(-\rho x_{L-1} + x_L) & k=L \end{cases}. \tag{36}$$

The interior weights (i.e., $1 < k < L$) can be manipulated into the suggestive form:

$$w_k \propto (1-\rho)^2 x_k^2 - \rho x_k \left(\nabla^2 x\right)_k, \tag{37}$$

where $(\nabla^2)_k = x_{k-1} - 2x_k + x_{k+1}$ is a finite difference approximation of the second derivative with respect to lag. Substituting (32) gives

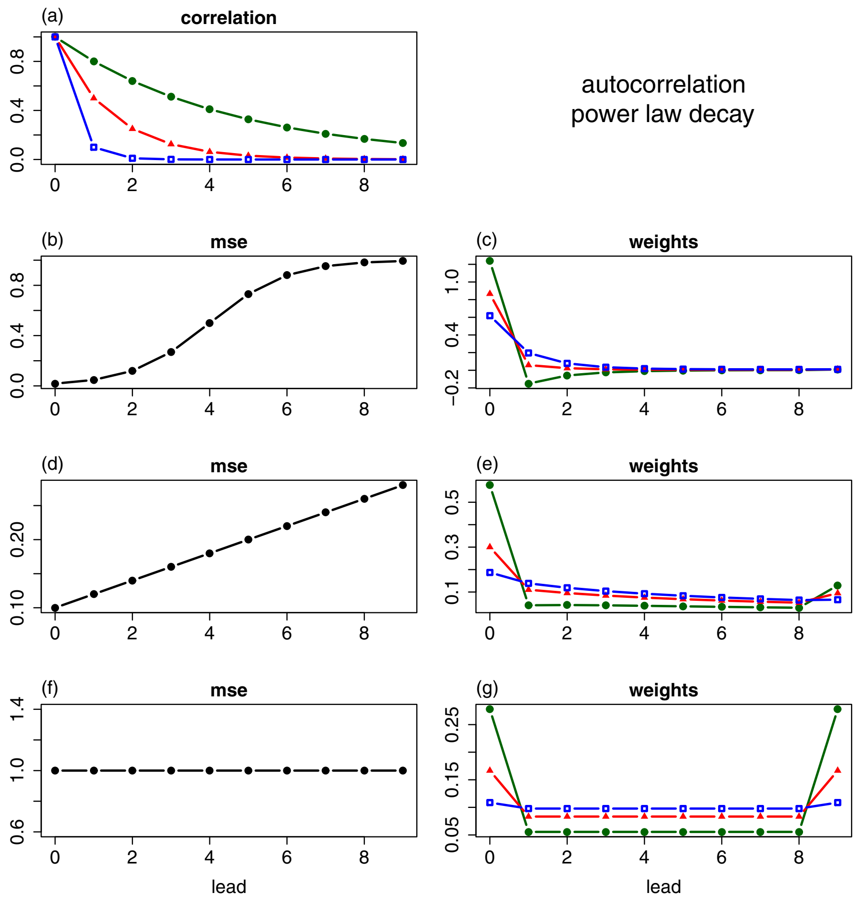

Figure 1. Illustration of the cross-lead error covariance matrix and corresponding weights of an optimal lagged ensemble. The covariance matrix is decomposed as (23), where the correlation matrix is of the Toeplitz form (26), with correlations shown in (a), parameterized as (34) using the values $\rho=(0.8, 0.5, 0.1)$ (green, red, blue, respectively). The mean square error (i.e., diagonal element of **D**) is parameterized as (b) a sigmoid, (d) linear, and (f) constant function of lead, and the respective weights are shown in (c), (e), and (g). The color of the curve for the weights coincides with the color of the correlation function in Figure 1a used to define the covariance matrix.

$$w_k \propto \frac{(1-\rho)^2}{MSE_k} - \frac{\rho}{\sqrt{MSE_k}}\left(\nabla^2 \frac{1}{\sqrt{MSE}}\right)_k. \tag{38}$$

This solution is a linear combination of $1/MSE$ and $1/\sqrt{MSE}$, as in (31), except that the last term in (38) includes an extra factor that depends on the *curvature* of x_k.

Building on the above results, we partition the full solution (36) as

$$\mathbf{w} = \mathbf{w}_{\text{non-gradient}} + \mathbf{w}_{\text{gradient}}, \tag{39}$$

where

$$w_{k,\text{gradient}} = \rho\lambda \begin{cases} -x_1(\nabla x)_2 & k=1 \\ -x_k(\nabla^2 x)_k & 1<k<L, \\ x_L(\nabla x)_L & k=L \end{cases} \tag{40}$$

where $(\nabla x)_k = x_k - x_{k-1}$ is a finite difference approximation to the first derivative, and

$$w_{k,\text{non-gradient}} = (1-\rho)\lambda \begin{cases} x_1^2 & k=1 \\ x_k^2(1-\rho) & 1<k<L. \\ x_L^2 & k=L \end{cases} \tag{41}$$

A plot of these terms is shown in Figure 2. It can be seen that the negative weight is caused entirely by the gradient term, while tail curvature is caused entirely by the nongradient term.

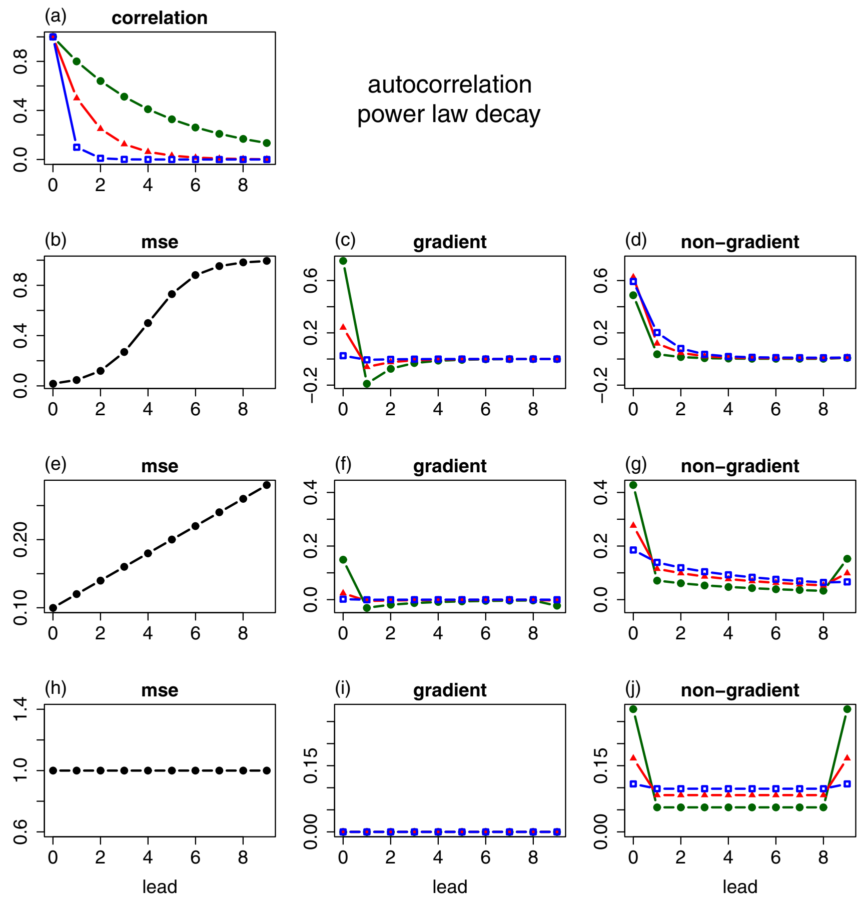

Figure 2. Same as Figure 1, except that the weights are decomposed into (c, f, i) a gradient term and (d, g, j) a nongradient term.

Lastly, suppose the correlations decay linearly as

$$R_{ij} = 1 - d|i-j|, \tag{42}$$

where d is a slope parameter. The matrix (42) has the analytic inverse:

$$\mathbf{R}^{-1} = -\frac{1}{2d}\begin{pmatrix} -\zeta_{L-1}/\zeta_L & 1 & 0 & 0 & \dots & 0 & 0 & d^2/\zeta_L \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \\ d^2/\zeta_L & 0 & 0 & 0 & \dots & 0 & 1 & -\zeta_{L-1}/\zeta_L \end{pmatrix}, \tag{43}$$

where $\zeta_L = -2d + d^2(L-1)$ (Dow, 2003). Substituting (43) into (25) yields

$$w_k = -\frac{\lambda}{2d}\begin{cases} x_1(-x_1\zeta_{L-1}/\zeta_L + x_2 + x_L d^2/\zeta_L) & k=1 \\ x_k(x_{k-1} - 2x_k + x_{k+1}) & 1 < k < L \\ x_L(x_1 d^2/\zeta_L + x_{L-1} - x_L\zeta_{L-1}/\zeta_L) & k=L \end{cases}. \tag{44}$$

Again, this solution can be partitioned into the form (39) with

$$w_{k,\text{gradient}}=\frac{\lambda}{2d}\begin{cases}-x_1(\nabla x)_2 & k=1\\ -x_k(\nabla^2 x)_k & 1<k<L\\ x_L(\nabla x)_L & k=L\end{cases}, \tag{45}$$

$$w_{k,\text{non-gradient}}=-\frac{\lambda d}{2\zeta_L}(x_1+x_L)\begin{cases}x_1 & k=1\\ 0 & 1<k<L\\ x_L & k=L\end{cases}. \tag{46}$$

In this case, if the curvature term vanishes, then the weights in the interior *vanish* and all of the weight would be placed *at the end points*. This tendency for all the weight to be placed at the end points is evident in Figure 1g, especially for large ρ where the power law decay approximates a linear decay law.

5. Constant Mean Square Error

The above results show that the tendency toward a U-shape is most obvious when the mean square error is nearly constant over the range of lags in question (i.e., when the gradient term is small). Let us therefore focus on this particular case, which is tantamount to assuming $\mathbf{D}=\alpha\mathbf{I}$, where $\alpha>0$. In this case, the covariance matrix becomes

$$\Sigma=\alpha^2\mathbf{R}, \tag{47}$$

which implies that the covariance matrix itself, rather than just the correlation matrix, is Toeplitz. In this case, we are effectively considering the linear combination:

$$\epsilon=w_1\epsilon_1+w_2\epsilon_2+\ldots+w_L\epsilon_L, \tag{48}$$

where the errors $\epsilon_1,\ldots,\epsilon_L$ have *equal variances*. In this case, the weights depend entirely on the correlations between the errors. Our goal is to develop an intuitive understanding of the optimal choice of weights.

First, we show that the optimal weights in (48) are *symmetric about their midpoint*. To show this, first note that relation (11) becomes

$$\mathbf{Rw}=\lambda'\mathbf{j}, \tag{49}$$

where $\lambda'=\lambda/\alpha^2$. This equation can be written in index notation as

$$R_{ik}w_k=\lambda', \tag{50}$$

where we invoke the implicit rule that indices that appear twice are summed. Applying the transformations $i\to L-i+1$ and $k\to L-k+1$ gives the equivalent set of equations:

$$R_{L-i+1,L-k+1}w_{L-k+1}=\lambda'. \tag{51}$$

Because the covariance matrix is Toeplitz (i.e., depend only on the *difference* in indices),

$$R_{L-i+1,L-k+1}=R_{ik}. \tag{52}$$

Substituting (52) into (51), and then differencing the resulting equation with (50), yields

$$R_{ik}(w_k-w_{L-k+1})=0. \tag{53}$$

Because $\mathbf{R}$ is positive definite, this system of equations is satisfied if and only if

$$w_k=w_{L-k+1}\quad \text{for } k=1,2,\ldots,L, \tag{54}$$

which proves that *the optimal weights from a Toeplitz matrix are symmetric about their midpoint*.

The above symmetry could have been anticipated by physical reasoning. For instance, it would be difficult to explain why the first weight w_1 could differ from the last weight w_L. After all, each random variable in the sum has equal variance, and the correlation of ϵ_1 with each of its neighbors are the same as those for ϵ_L, so there is nothing in the properties of the random variables to distinguish the first random variable from the last random variable. This symmetry immediately implies that the optimal weights from a Toeplitz covariance matrix must be either constant or exhibit a U-shape somewhere (which could be either concave upward or downward). In particular, the weights cannot decay monotonically with lead.

When are the weights constant? To answer this question, suppose $\mathbf{w} \propto \mathbf{j}$. Substituting this into (49) reveals that each row of $\mathbf{R}$ must sum to the same value. As can be seen from (26), the difference between the sum of the first two rows yields the condition $\rho_1 = \rho_{L-1}$; the difference between the sum of the second and third rows of $\mathbf{R}$ yields the condition $\rho_2 = \rho_{L-2}$. More generally, the difference between the k'th and $(k+1)$'th rows is

$$\rho_k = \rho_{L-k} \quad \text{for } \mathbf{w} \propto \mathbf{j}. \tag{55}$$

This result shows that *if the weights from a Toeplitz matrix are equal, then the correlations of the correlation matrix are symmetric about their midpoint*. It follows immediately that if the correlations decay with lag, which is empirically true, then the optimal weights cannot be constant; i.e., the weights must exhibit a U-shape somewhere.

5.1. Constant Off-Diagonal Elements

Consider the case in which the off-diagonal elements of the correlation matrix are constant, as in (27). This matrix satisfies (55), hence it follows immediately that the optimal weights are constant:

$$\mathbf{w}_{\text{opt}} = \frac{1}{L}\mathbf{j} \quad \text{for } \mathbf{R} = (27). \tag{56}$$

This solution also could have been obtained by substituting (29) into (49). Note that this solution is valid regardless of the value of ρ.

5.2. Correlations Decay as a Power Law

Next, consider the case in which the correlation matrix is of the form:

$$R_{ij} = \rho^{|i-j|}. \tag{57}$$

Because the correlations do not satisfy (55), the weights cannot be uniform. Condition (49) is equivalent to the condition:

$$\mathbf{w} = \lambda \mathbf{R}^{-1}\mathbf{j}. \tag{58}$$

Substituting (35) into (58), and choosing λ to satisfy the constraint (4), yields

$$w_i = \frac{1}{2+(L-2)(1-\rho)} \begin{cases} 1 & i=1 \\ 1-\rho & 1 < i < L \\ 1 & i=L \end{cases}. \tag{59}$$

Because $|\rho| < 1$, the weights are largest at the end points $i=1, L$, thus the weights form a U-shape. Moreover, the U-shape becomes more extreme as $\rho \to 1$. These features are easily seen in Figure 1g.

5.3. Correlations Decay Linearly

Next, consider the case in which the correlations in the correlation matrix decay linearly, as in (42). Substituting (43) into (58) yields

$$\mathbf{w}_{\text{opt}} = \begin{pmatrix} 1/2 \\ 0 \\ \vdots \\ 0 \\ 1/2 \end{pmatrix} \quad \text{for } \mathbf{R} = (42). \tag{60}$$

Remarkably, this solution holds regardless of the slope coefficient d. This result shows that when the correlations decay linearly, *all of the weight* is placed on the first and last lags. This feature also is easily seen in Figure 1g.

5.4. Condition for the Interior Weights to Vanish

To understand the above result, consider the question of when the interior weights vanish. Suppose all but the first and last weights w_1 and w_L vanish. The system of equations (49) then become

$$w_1+\rho_{L-1}w_L=\lambda', \tag{61}$$

$$\rho_1 w_1+\rho_{L-2}w_L=\lambda', \tag{62}$$

$$\rho_2 w_1+\rho_{L-3}w_L=\lambda', \tag{63}$$

$$\vdots \tag{64}$$

By symmetry, $w_1=w_L$, in which case the above equations yield the system of equations:

$$\rho_{k-1}+\rho_{L-k}=\text{constant} \quad \text{for } k=1,2,\ldots,L-1. \tag{65}$$

It is easily verified that $\rho_k=1-\beta k$ solves this system of equations. Moreover, it can be shown that this is the only nontrivial solution to the above linear inhomogeneous difference equation (Efthimiou, 2011). This result shows that if the correlations are a linear function of lag, then the interior weights must vanish, as found in (60), leaving only two nonzero weights at the end points.

The above results partly rationalize the solution (60). In particular, the above results show that if the correlations decay linearly, then the interior weights vanish, leaving two forecasts at the end points for combining. The fact that the two forecasts at the end points have the most extreme difference in lags means that the resulting combination will have the smallest mean square error of any pair of forecasts, as implied by (22). Finally, as discussed in section 3, the optimal combination of two forecasts with equal variances is to combine them equally, i.e., to use the weights $w_1=w_L=1/2$.

5.5. Three-Member Lagged Ensemble

Many of the above results, plus other new features, can be illustrated using a three-member lagged ensemble. The error of a weighted 3-member lagged ensemble is

$$\epsilon=w_1\epsilon_1+w_2\epsilon_2+w_3\epsilon_3, \tag{66}$$

where ϵ_τ is the forecast error at lead τ. The assumption that the cross-lead error covariance matrix is Toeplitz means that the error covariances depend only on the difference in lead times. Assuming the errors are unbiased and have unit variance,

$$E[\epsilon_1\epsilon_2]=\rho_1, \quad E[\epsilon_2\epsilon_3]=\rho_1, \quad \text{and} \quad E[\epsilon_1\epsilon_3]=\rho_2. \tag{67}$$

Thus, only two correlations enter the problem, ρ_1 and ρ_2. When $\rho_1=\rho_2$, the error covariance matrix is of the form (27), hence the weights are equal (see (56)). It is easy to see why the weights are equal in this case: from symmetry considerations, there is nothing in the problem to distinguish the errors, hence the weights are exchangeable. Moreover, the weights sum to one, hence they must each be equal to $1/L$. In fact, this is the *only* case in which the weights are equal to each other, because $\rho_1=\rho_2$ is precisely equivalent to condition (55).

Next, recall that the optimal weights from a Toeplitz error covariance matrix satisfy (54), which in the present example implies $w_1 = w_3$. The fact that the weights are symmetric about their midpoint immediately implies that the weights must be either constant, or U-shaped. As proven above, the weights are constant if and only if $\rho_1=\rho_2$. Therefore, it follows that the weights must be U-shaped if $\rho_1 \neq \rho_2$; that is, if the correlation between ϵ_1 and ϵ_3 differs from the other correlations. This condition "breaks the symmetry" and distinguishes the errors.

Next, recall that the weights sum to one. Since the weights are symmetric about their midpoint, let $w_1=w_3=w$. Then the fact that the weights sum to one implies $w_2=1-2w$. Therefore, the error of the weighted lagged ensemble (66) can be expressed equivalently as

$$\epsilon=w\epsilon_1+(1-2w)\epsilon_2+w\epsilon_3=w(\epsilon_1+\epsilon_3-2\epsilon_2)+\epsilon_2. \tag{68}$$

The mean square error is therefore

$$MSE = E[\epsilon^2] = w^2 \ \mathrm{var}[\epsilon_1 + \epsilon_3 - 2\epsilon_2] + 1 + 2w \ \mathrm{cov}[\epsilon_1 + \epsilon_3 - 2\epsilon_2, \epsilon_2]. \quad (69)$$

The first two terms are positive, while the last term, involving the covariance, is negative (as will be shown below). Were it not for the covariance term, the minimum *MSE* would be obtained by $w = 0$. Evaluating the *MSE* using (67) gives

$$MSE = w^2(6 - 8\rho_1 + 2\rho_2) + 2w(2\rho_1 - 2) + 1. \quad (70)$$

The minimum of this quadratic equation gives the value of w that minimizes *MSE*. This value is

$$w_{\mathrm{opt}} = \frac{1 - \rho_1}{3 - 4\rho_1 + \rho_2}. \quad (71)$$

We consider the following special cases. First, the case of equal weights requires

$$\rho_1 = \rho_2 \quad \text{for } w_{\mathrm{opt}} = 1/3. \quad (72)$$

This recovers our previous results that equal weighting requires equal correlations, and that a U-shaped set of weights must occur when $\rho_1 \neq \rho_2$. Second, suppose the correlations decay as a power law (34); thus, $\rho_1 = \phi$ and $\rho_2 = \phi^2$. Then, the optimal weight (71) can be simplified to

$$w_{\mathrm{opt}} = \frac{1}{3 - \phi} \quad \text{for } \rho_\tau = \phi^\tau. \quad (73)$$

Here the weights are equal for $\phi = 0$, but as ϕ increases from 0, w_{opt} increases, thereby producing (literally) a U-shape in the weights. This example shows that the U-shape in the weights occurs for any power law decay of the correlations. Third, suppose the correlations decay linearly as $\rho_\tau = 1 - \beta\tau$. Then, the optimal weight (71) can be simplified to

$$w_{\mathrm{opt}} = \frac{1}{2} \quad \text{for } \rho_\tau = 1 - \beta\tau, \quad (74)$$

which recovers (60). In this case, all of the weight is placed on the end points and the interior weight vanishes. Remarkably, the weight is independent of the slope parameter β.

None of the above examples give an *concave downward* U-shape for the weights. A concave downward U occurs when $w_{\mathrm{opt}} < 1/3$, which occurs when $\rho_1 < \rho_2$. In other words, the correlations would need to *increase* with the difference in lead times in order to produce an inverted U-shape. Such an increase is rarely observed empirically.

6. Conclusion

This paper examined the weights of an optimal weighted-average lagged ensemble over a range of idealized models for the forecast errors. If the forecast errors are uncorrelated, then the optimal weights are inversely proportional to the mean square error, as might be expected intuitively. The goal of this paper was to understand when the optimal weights deviate from this behavior; in particular, when the weights contain negative values or depend nonmonotonically on lead time, as noted in previous studies (Lu et al., 2007; Trenary et al., 2017). We show that both types of behavior can be reproduced in a broad range of statistical models for the errors, thereby demonstrating that they are not necessarily artifacts of sampling errors.

An instructive approach is to assume that the correlation between forecast errors depends only on the lag between the forecasts. This assumption leads to a Toeplitz correlation matrix, which enjoys special analytic properties. Typically, the correlation between forecast errors decays with lead time. Accordingly, we consider two decay laws, one based on a power law and one based on linearity. In both cases, the weights can be written as the sum of two terms: a term that depends on gradients of error growth, and another term that does not. The term associated with gradients in error growth causes negative weights when the error growth is sufficiently strong and the errors are sufficiently correlated. The term that is not associated with gradients of error growth can contribute a U-shape in the weights, in the sense that the weights eventually increase with lag, contrary to intuition. To understand the latter behavior better, we examined the extreme case in which the mean square error is constant over the range of lags in the ensemble, which leaves the correlation structure as the only detail that determines the weights. In this case, the U-shape follows immediately from symmetry principles. That is, we prove that if the mean square error is constant over the range of lags in the ensemble, and the correlation between errors decreases with the difference in leads, then the weights must have a U-shape. Thus, such U-shaped weights are a ubiquitous phenomenon when the mean square errors are approximately constant over the range of lead times in the lagged ensemble. If the correlation decay also is exactly linear, then, remarkably, the weights *vanish* at intermediate lags, leaving all of the weight to be concentrated on the first and last forecast in the lagged ensemble. These results appear to clarify certain surprising and counter-intuitive aspects of the weights of optimal weighted-average lagged ensembles seen in previous studies.

Appendix A: AR1 Model

In this appendix, we compute the cross-lead error covariance function of an AR1 process. Consider the AR1 model:

$$x_t = \phi x_{t-1} + w_t, \tag{A1}$$

where ϕ is a parameter between 0 and 1 and w_t is a white noise process with zero mean and variance σ_W^2. Trenary et al. (2017) showed that the cross-lead error covariance matrix of this process is

$$\Sigma_{\tau_1,\tau_2} = \left(1+\delta_{\tau_1,\tau_2}\right)\left(1-\phi^{\min[\tau_1,\tau_2]}\right)\frac{\sigma_W^2}{1-\phi^2}, \tag{A2}$$

where δ_{ij} denotes the Kronecker delta function. This covariance matrix has the form:

$$\Sigma = \frac{\sigma_W^2}{1-\phi^2}\begin{pmatrix} 2(1-\phi) & 1-\phi & 1-\phi & \dots & 1-\phi \\ 1-\phi & 2(1-\phi^2) & 1-\phi^2 & \dots & 1-\phi^2 \\ 1-\phi & 1-\phi^2 & 2(1-\phi^3) & \dots & 1-\phi^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1-\phi & 1-\phi^2 & 1-\phi^3 & \dots & 2(1-\phi^L) \end{pmatrix}. \tag{A3}$$

Based on this covariance matrix, the ith row of (11) may be written as

$$\sum_j \left(1-\phi^{\min[i,j]}\right) w_j = \lambda\left(\frac{1-\phi^2}{\sigma_W^2}\right) - \left(1-\phi^i\right) w_i. \tag{A4}$$

Taking the difference between the ith and $(i+1)$ th rows, assuming $i \le L-1$, yields

$$\sum_j \left(\phi^{\min[i+1,j]} - \phi^{\min[i,j]}\right) w_j = \left(1-\phi^{i+1}\right) w_{i+1} - \left(1-\phi^i\right) w_i$$

$$\left(\phi^{i+1}-\phi^i\right)\sum_{j=i+1}^{L} w_j = \left(1-\phi^i\right)\left(w_{i+1}-w_i\right) + \left(\phi^i - \phi^{i+1}\right) w_{i+1}$$

$$\left(\phi^{i+1}-\phi^i\right)\left(w_{i+1} + \sum_{j=i+1}^{L} w_j\right) = \left(1-\phi^i\right)\left(w_{i+1}-w_i\right). \tag{A5}$$

For positive weights, the left hand side of (A5) is negative (because $|\phi| < 1$), and therefore,

$$w_{i+1} < w_i. \tag{A6}$$

That the weights are positive can be seen by writing (A5) for $i = L-1, L-2, \ldots$. In each case, w_i can be written as a linear combination of weights at higher lags, $w_{i+1}, w_{i+2}, \ldots$, with positive coefficients. Thus, $w_{L-1} = \beta w_L$ for some positive β, which proves that w_{L-1} and w_L have the same sign. Repeating this argument for $i = L-2, L-3, \ldots$ proves that all weights must have the same sign. Moreover, the case of all negative weights can be excluded because (A5) cannot be satisfied in this case, given that Σ has only positive elements. Therefore, all the weights are positive, and inequality (A6) implies that the weights of an optimally weighted lagged ensemble of an AR1 process *decrease monotonically*. Thus, the optimal weights of a lagged ensemble for an AR1 process cannot be larger at the longest lead times relative to the weights at slightly shorter lead times. This conclusion is consistent with the fact that the error covariances discussed in Trenary et al. (2017) had structure that clearly differed from the pattern seen in (A3).

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), and the National Oceanic and Atmospheric Administration (NA14OAR4310160). This paper complies with the AGU Publications Data Policy in that no data were produced by this study beyond the mathematical model described in the text. The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

References

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research*, *20*(4), 451–468.

Ben Bouallegue, Z., Theis, S. E., & Beghardt, C. (2013). Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, *22*, 49–59.

Buizza, R. (2008). Comparison of a 51-member low-resolution (tl399l62) ensemble with a 6-member high-resolution (tl799l91) lagged-forecast ensemble. *Monthly Weather Review*, *136*(9), 3343–3362. https://doi.org/10.1175/2008MWR2430.1

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

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

Dalcher, A., Kalnay, E., & Hoffman, R. N. (1988). Medium range lagged average forecasts. *Monthly Weather Review*, *116*(2), 402–416. https://doi.org/10.1175/1520-0493(1988)116<0402:MRLAF>2.0.CO;2

DelSole, T., Yang, X., & Tippett, M. K. (2012). Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society*, *139*, 176–183. https://doi.org/10.1002/qj.1961

Dow, M. (2003). Explicit inverses of Toeplitz and associated matrices. *ANZIAM Journal*, *44*(E), E185–E215.

Efthimiou, C. (2011). *Introduction to functional equations: Theory and problem-solving strategies for mathematical competitions and beyond (MSRI Mathematical Circles Library)* (363 pp.). Providence, RI: American Mathematical Society.

Hoffman, R. N., & Kalnay, E. (1983). Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus, Series A*, *35*(2), 100–118. https://doi.org/10.1111/j.1600-0870.1983.tb00189.x

Johnson, C., & Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, *135*(640), 777–794. https://doi.org/10.1002/qj.383

Lu, C., Yuan, H., Schwartz, B. E., & Benjamin, S. G. (2007). Short-range numerical weather prediction using time-lagged ensembles. *Weather and Forecasting*, *22*(3), 580–595. https://doi.org/10.1175/WAF999.1

Raynaud, L., Pannekoucke, O., Arbogast, P., & Bouttier, F. (2015). Application of a bayesian weighting for short-range lagged ensemble forecasting at the convective scale. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 459–468. https://doi.org/10.1002/qj.2366

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*(1), 291–306. https://doi.org/10.1002/2016MS000838

Yuan, H., Lu, C., McGinley, J. A., Schultz, P. J., Jamison, B. D., Wharton, L., & Anderson, C. J. (2009). Evaluation of short-range quantitative precipitation forecasts from a time-lagged multimodel ensemble. *Weather and Forecasting*, *24*(1), 18–38. https://doi.org/10.1175/2008WAF2007053.1