Final Report

Investigator:	Timothy DelSole
co-Investigator:	Michael K. Tippett and Kathleen Pegion
Recipients Name:	George Mason University
Award Number:	NA14OAR4310187
Reporting Period:	Final report
Award Period:	08/01/2014 - 07/31/2018
Program Office:	OAR Climate Program Office (CPO)
NOAA Person of Contact:	Arun Kumar
Program Officer:	Annarita Mariotti
Participants:	Laurie Trenary
Project Title:	Subseasonal NMME Forecasts: Skill, Predictability, and Multi-model Combinations

1. Proposed Goals

The goals of our proposed work were to systematically diagnose predictability and skill on subseasonal timescales in a suite of North American Multi-Model Ensemble (NMME) models, and to develop a statically informed protocol for generating subseasonal predictions. Specific goals include:

- Development of an objective criteria for diagnosing and comparing subseasonal predictability and skill in NMME for all relevant lead times, target days, and models.
- Objectively determine the number of lags that should be included in a lagged ensemble to provide the most skillful subsesaonal forecast.
- Assess whether different initialization frequencies improve subseasonal forecast skill.

2. Results and Accomplishments

This document summarizes our major accomplishments for the award period August 2015 to May 2018.

Statistical Methods for Comparing Forecast Skill

We developed rigorous methods for comparing forecast skill in a previous CTB research project (DelSole and Tippett, 2014). These methods represent a "breakthrough" in the long-standing problem of deciding whether one forecast is more skillful than another. In particular, differences in skill have been assessed in some previous studies using Fisher's test for equality of correlations, or the F test for equality of variances. However, these tests assume that the skill estimates are independent, which is never satisfied when the skills are computed on a common period or use a common set of observations. DelSole and Tippett (2014) proposed four new tests that can be applied over a common period.

During the term of this project, we have built on the above results by developing a "random walk" method of forecast comparison. The procedure is simple: the squared error of forecasts A and B are computed, then the score is increased by one if A beats B, or decreased by one if B beats A. The score is accumulated as a function of initial start time for a fixed lead time. To test significance, a natural null hypothesis is that the forecasts are equally skillful, in the sense that the probability that one model beats the other is 50%. It follows that the average score should vanish, and the 95% confidence interval can be computed from a Bernoulli distribution with p = $1/2$. In particular, the accumulated score should evolve according to a classical random walk. Note that this methodology avoids the statistical pitfalls associated with testing differences in correlation skill or mean square error discussed above. As an example, we compare seasonal forecasts of NINO3.4 by CFSv2 with those by other models in the NMME. The result is shown in fig. 1 (see figure caption for details). The figure reveals an abrupt decrease in skill around 2000, presumably due to the discontinuity in climatology of the CFSv2 associated with the introduction of ATOVS satellite data around 1998 (Kumar et al., 2012). This result demonstrates that this technique has the potential to detect abrupt changes in skill in real-time forecasting, which may prove useful for routine monitoring of skill among NMME forecasts. We anticipate that this technique will be a useful foundation for some of the research in this project.

Extended-range forecasts of areal-averaged Saudi Arabia rainfall

The climate of Saudi Arabia is arid to semiarid, and rainfall occurrence is infrequent and scattered. However, when rainfall does occur in the region, it is sometimes intense and causes flooding, loss of life and damage to property. Several studies have related precipitation in the region to remote forcing, especially tropical convection related to ENSO and the MJO. These climate signals are predictable, and based on the predictability of the MJO, it has been proposed that rainfall in the region should be predictable up to 2-3 weeks in advance. We have documented the ability of the CFSv2 to forecast areal averaged Saudi Arabia rainfall (Tippett et al., 2015). We find that the CFSv2 was able to predict features of a 2013 flooding event up to 10 days in advance. The reforecasts show that the CFS can skillfully predict rainfall amount, number of days exceeding a threshold, and probability of heavy rainfall occurrence for varying forecast averaging windows. Logistic regression improves forecast skill and reliability. Forecast probability signals have a clear relation with MJO phase during the 6-month wet season (November-April), with the forecasts for wet conditions being significantly more common during phases 1,2,7 and 8 (fig. 2). No such relation is seen during the dry season.

Predictability at week 3-4 over the contiguous United States

One of the goals of the proposed research was the diagnosis of predictability and skill on subseasonal timescales. To this end, we evaluated the 3-4 week predictability of temperature and precipitation over the United States in the Climate Forecast System version 2 (CFSv2). The results are summarized in DelSole et al. (2017). We demonstrate that the CFSv2 is capable of producing skillful forecasts of temperature and precipitation in the 3-4 week range

for regions of the continental United States. Skill is established by grid-point correlation and Predictable Component Analysis, a statistical method that finds the components of each respective fields that maximize predictability in the model. As an example, fig. 3 shows the point-wise correlation skill for CFSv2 forecasts of January and July temperature and precipitation over the United States. This figure shows that significant forecast skill can be found over nearly half of the area of the United States for January/July temperature and January precipitation. Additional analysis presented in DelSole et al. (2017) indicates that these predictable structures are linked to variability of the El Niño Southern Oscillation (ENSO) and the Madden Julian Oscillation (MJO). Another important contribution from this work was development of a new significance test that accounts for serial correlation in daily data.

This work was presented at the 41st NOAA Climate Diagnostics and Prediction Workshop and was recently featured in Inside Science[1] and highlighted on the NOAA Climate Program Office website.[2]

Developing a Protocol for Subseasonal Forecasting

Often operational centers do not have the resources needed to perform the large number of hindcasts required to identify the optimal forecast ensemble size. To address this issue, we developed an objective method that can be used to estimate the skill of an arbitrary lagged ensemble given only a finite number of hindcasts. A brief description of the methodology is presented here, a more detailed can be found in Trenary et al. (2017).

If the errors are stationary, the mean square error (MSE) of the lagged ensemble as a function of lead time (τ) and lagged ensemble size (L) is

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

where Δ is the time interval between initialization times (assumed fixed for simplicity) and $C(\tau_1, \tau_2)$ is the covariance between forecast errors at leads τ_1 and τ_2 for forecasts verifying on the same target date. We proposed a parametric model of C and use re-forecast data to estimate the parameters of this model. Once the parameters are determined, the MSE can be evaluated for arbitrary τ, L, and Δ (note that Δ can be fractions of the original time interval between initializations).

This methodology was used to estimate the optimal lagged ensemble size for subseasonal forecasts of the Madden Julian Oscillation (MJO). To do this, we first estimated the error covariance matrices for the MJO indices RMM1 and RMM2 (Wheeler and Hendon, 2004) from CFSv2 hindcasts initialized once per day and then fit a parametric model to each. A representative example of the fit is shown in figs. 4a and b and demonstrates that the parameterized covariance function is capable of capturing with great accuracy the lagged

[1] https://www.insidescience.org/news/breaking-new-ground-weather-forecasting

[2] http://cpo.noaa.gov/AboutCPO/AllNews/TabId/315/ArtMID/668/ArticleID/725267/NOAA-model-shows-significant-forecast-skill-3-4-weeks-in-advance.aspx

error growth as a function of lead and ensemble size for both indices. The skill derived from this parameterized covariance function is shown in fig. 5b and is in excellent agreement with the values computed directly from CFSv2 output fig. 5a.

Having estimated a parametric covariance function from once-per-day initialized forecasts, we now use it to compute the skill for 4-per-day initialized forecasts. The resulting empirically derived MSE is shown in fig. 5d, which is in remarkable agreement with the actual mean square error derived from lagged ensembles initialized six hours apart, shown in fig. 5c. Note that the empirically derived normalized MSE captures the reduction of the MSE with the inclusion of more initializations, this is particularly evident at the longer lead times. These results demonstrate the we are able to fit an empirical model to the lagged error covariance for a single initialization and interpolate this analytic function to different frequencies. We can then accurately estimate of the impacts of sampling frequency on forecast error as a function of lead time and ensemble size.

Operationally NCEP uses a total of 16 ensemble members per forecast day, which is generated from 4 "burts" ensemble members at 0Z, 6Z, 12Z, 18Z. We use our parametric covariance function to test whether increasing the ensemble size beyond once-per-day significantly improves the MJO forecast skill. In particular, we evaluate the covariance function at 16 and 1000 equally spaced intervals per day. The resulting NMSE as a function of lead time and lagged ensemble size for the 16 and 1000 ensemble members are shown in fig. 6a and b, respectively. Comparing these two figures, it is evident that there is little improvement in forecast skill when more than 16 ensemble members are used. By comparing, fig.5d and fig. 6a, we see that there is only marginal improvement in MJO forecast when the ensemble size is increased from 4 to 16.

Monthly ENSO forecast skill and lagged ensemble size

We adapted the methodology described above to the lagged ensemble for real-time CFSv2 forecasts of Niño3.4 (Trenary et al. 2018). The main surprise in the new application to Niño3.4, relative to our application to the MJO, is that an identifiability problem emerged when estimating parameters in the covariance model because Niño3.4 has a much longer time scale compared to the MJO. This problem was resolved by removing a parameter and modifying the covariance model slightly. The resulting covariance model was robust regardless as to whether it was fit using real-time forecast data or hindcast data. We confirmed that our covariance model could recover the MSE of a lagged ensemble initialized 4/day over a wide range of ensemble sizes and lead times, even though it was estimated from 1/day initialization data

Our major result from this part of the project is shown in figure 7. This figure shows the normalized MSE for burst and lagged ensembles, as estimated by our method. Comparison is made for given ensemble size. For example, when considering the lagged ensemble, the initializations are included in ensemble size count (shown as the horizontal axis in fig. 7), such that an ensemble size of 16 is equivalent to 4 lagged members with 4 initializations per day (denoted by the vertical dashed line in fig. 7). The MSE of the burst ensemble was

estimated from our covariance model by considering lagged ensemble members initialized an infinitesimal time step apart. We find that for all leads and an ensemble size less than 30 days, the forecast skill as a function of ensemble size is roughly equal for the lagged (with four initializations) and burst ensemble configuration. When more than 30 ensemble members are included, the MSE of the burst ensemble saturates. This saturation of the MSE provides an estimate of the infinite ensemble MSE, since no reduction in MSE occurs with the addition of more members. In contrast, the MSE of the lagged ensemble continues to grow with ensemble size since the addition of each lagged member introduces forecasts initialized further from the target date. Applying this method to real-time forecasts, we find that the MSE consistently reaches a minimum for a lagged ensemble size between one and eight days, when four initializations per day are included. This ensemble size is consistent with the 8-10 day lagged ensemble configuration used operationally and is close to the estimated skill of the infinite ensemble. *As such, the current ensemble configuration for the operational monthly ENSO forecast appears to be nearly optimal. Improvements in forecast skill will most likely come from increased spatial resolution and improvements to parameterized physics, rather than increases in ensemble size.* Moreover, we find that the skill of the weighted, lagged, and burst ensembles are nearly the same.

This work was reported in Trenary et al. (2018).

Weighted-average lagged ensemble

During this award period, we also made theoretical strides in understanding the *weighted* lagged ensemble. It is natural to consider a weighted ensemble because each member of a lagged ensemble is initialized at different times, so intuitively one should down-weight members that have smaller skill owing to their longer lead times. However, when we determined the optimal weights for minimizing mean square error, sometimes the member with the longest lead time did not have the smallest weight (Trenary et al., 2017). To understand this behavior, we examined a series of analytic examples designed to illuminate conditions under which the weights of an optimal lagged ensemble become negative or depend non-monotonically with lead time.

The optimal weights can be estimated directly from the error covariance matrix Σ as follows

$$\mathbf{w}_{opt} = \frac{\Sigma^{-1}\mathbf{j}}{\mathbf{j}^T\Sigma^{-1}\mathbf{j}}, \tag{2}$$

where $\mathbf{j}$, is a vector consisting of ones. Since the error covariance matrix is positive definite, eqn. (2) can be further decomposed into variance and correlation dependancies:

$$\Sigma = \mathbf{DRD}, \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix, with elements equal to the lead dependent MSE, and $\mathbf{R}$ is a correlation matrix. This decomposition allows us to examine changes in behavior of the

weights with respect to perturbations made to error growth ($\mathbf{D}$) and correlation ($\mathbf{R}$) properties separately.

Figure 8 shows results for three different error growth functions—logistic, linear, and constant and three types of decay functions for the correlation—constant, power law and linear. The three different autocorrelation functions are shown in figure 8a, and the error growth functions are shown in figures 8b, d, f. In the case of the logistic function (figures 8b and 8c), the weights become negative for small lead times, but do not develop a U-shape at the tail. For linear error growth (fig. 8d), the weights (fig. 8e) are positive and develop positive curvature at the tail, with the curvature increasing with increasing correlation. Finally, for constant mean square error (fig. 8f), the weights show a very significant curvature at the tail (fig. 8g). Because these results were obtained from an exactly solvable covariance model, it eliminates questions about whether such behavior is caused by sampling errors. Several mathematical properties of this model were derived which clearly demonstrate that the optimal weights do not always decay monotonically with lead time.

Generally we find that the weights are most likely to behave non-monotonically when the mean square error is approximately constant over the range of forecasts included in the lagged ensemble and when the forecast errors at different lead times are highly correlated.

This work was reported in DelSole et al. (2018).

Climatology bias and forecast skill

In the course of our work with Niño3.4 lagged ensembles, we discovered significant biases and discontinuities in the NCEP climatology for real-time forecasts. Subsequent investigation revealed that the monthly forecast climatology provided by the NCEP Environmental Modeling Center (EMC) for the CFSv2 were biased in the sense of systematically differing from the hindcasts that are used to compute it. These biases, which are unexpected, are primarily due to fitting harmonics to hindcast data that have been organized in a particular format, which on careful inspection is seen to introduce discontinuities. A further undesirable consequence of this fitting procedure is that the EMC forecast climatology varies discontinuously with lead time for fixed target month. For example, these discontinuities and model biases are clearly visible when the EMC climatology for the Niño3.4 index is plotted as a function of lead time for a few select months, shown as the colored curves in figure 9a and b. Two alternative methods for computing the forecast climatology are proposed and the resulting fits are shown as the black and grey curves in figures 9a and b. We found that the choice of forecast climatology can have a large impact on the resulting forecast anomalies and, therefore, forecast skill. As an example, figures 10a,b show two forecast anomalies from the CFSv2 initialized 6 hours apart, where the anomalies are defined with respect to the EMC forecast climatology. The figure shows that the earlier forecast for the monthly mean (fig. 10a) is substantially cooler than the later forecast, despite being initialized only 6 hours earlier. If, however, the anomalies are defined with respect to one of our proposed methods, then the resulting forecast anomalies are shown in figures 10c, d and show much less variation. This result demonstrates that the difference in forecast anomalies seen in figures 10a,b

are due primarily to the discontinuity in the EMC forecast climatology, not to the change in forecast. Overall, the proposed methods more accurately fit the hindcast data and provide a clearer representation of the CFSv2 model climate drift.

This work was reported in Tippett et al. (2018).

3. Highlights Accomplishments

- Development of a rigorous method for comparing forecast skill.

- Demonstration of forecast skill of temperature and precipitation over the contiguous United States on the 3-4 week timescale. These results provide a scientific basis for predictability on these timescales.

- Formalization of a rigorous significance test for serially correlated daily data.

- Development of a statistically informed methodology capable of identifying the optimal ensemble using only preexisting data. The procedure is applicable to lagged and burst ensembles and is capable of estimating the infinite ensemble in terms of the Mean Square Error.

- Identification of the optimal lagged ensemble for subseasonal forecasts of the MJO and seasonal forecasts of ENSO in CFSv2.

- Weights of the optimally weighted-average lagged ensemble do not always decay with lead time. Using a series of analytic examples we clarify the conditions that lead to this surprising and counterintuitive behavior.

- We demonstrate that the forecast climatology provided by the EMC introduces bias into the forecast, thereby reducing skill. We propose two alternative methods for computing the forecast climatology, both of which more accurately fit the hindcast data and provide a clearer representation of the CFSv2 model climate drift.

4. Transitions to Operations

This project has produced a methodology that can be used to inform the protocol for operational lagged-ensemble forecasting. This methodology requires much less data for estimating the error of a lagged ensemble forecast than the usual brute force approach of producing long-term re-forecasts at a high initialization frequency. Also, the re-forecasts need not include burst ensembles because our method can estimate the skill of a burst ensemble of arbitrary size. This methodology is most useful at the planning stage prior to operations, rather than after the forecasts become operational (in which case the protocol is frozen and

not changeable). Our analysis indicates that the current ensemble configuration is optimal for MJO and ENSO prediction, in terms of mean square error performance, and that further improvements are very unlikely to be obtained by increasing ensemble size.

This project has discovered biases and discontinuities in the NCEP climatology for real-time forecasts. In addition, we propose two simple methods for estimating climatologies that avoid these biases and discontinuities. These methods represent a very simple modification of the regression analysis that is already used in constructing the NCEP climatologies.

This project has produced compelling evidence that CFSv2 produces skillful forecasts of temperature and precipitation in the 3-4 week range for regions of the continental United States. An important contribution of this work is the development of a new significance test that accounts for serial correlation in the data. Thus, the procedure developed in this project can be applied to operational forecasts from future models. The results of our work provide a foundation for developing operational forecasts beyond two weeks based on CFSv2 forecasts.

5. Publications from Project

DelSole T. and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658-4678.

DelSole T. and M. K. Tippett, 2015: Forecast comparison based on random walks. *Mon. Wea. Rev.*, **142(2)**, 615- 626.

Trenary L., T. DelSole, M. K. Tippett, and B. Doty, 2016: Extreme eastern US winter of 2015 not symptomatic of climate change [in "Explaining Extremes of 2015 from a Climate Perspective"]. *Bull. Amer. Meteor. Soc.*, **97**, S31-S35, doi:10.1175/BAMS-D-16-0156.1.

Barnston A. G., M. K. Tippett, M. Ranganathan, and M. L. LHeureux, 2017: Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Clim. Dyn.*, doi:10.1007/s00382-017-3603-3.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3-4 average temperature and precipitation over the contiguous United States. *J. Clim.*, **30 (10)**, 3499-3512, doi:10.1175/JCLI-D- 16-0567.1.

Shawki D., R. D. Field, M. K. Tippett, B. H. Saharjo, I. Albar, D. Atmoko, and A. Voulgarakis, 2017: Long-lead prediction of the 2015 fire and haze episode in Indonesia. *Geophys. Res. Lett.*, **44**, 9996-10,005, doi:10.1002/2017GL073660.

Tippett M. K., M. Ranganathan, M. L'Heureux, A. G. Barnston, and T. DelSole, 2017: Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Clim. Dyn.*, doi: 10.1007/s00382-017-3721-y.

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:10.1002/2016.

DelSole, T., L. Trenary, and M. K. Tippett, 2018: The weighted-average lagged ensemble. *J. Adv. Model. Earth Syst*, **9**, 2739-2752. doi: 10.1002/2017MS001128.

Tippett M. K., L. Trenary, T. DelSole, K. Pegion, and M. L'Heureux, 2018: Sources of Bias in the Monthly CFSv2 Forecast Climatology. *J. Appl. Meteor. Climatol.*, **57**, 1111-1121, doi: 10.1175/JAMC-D-17-0299.1.

Trenary, L., T. DelSole, M. K. Tippett, and K. Pegion, 2018: Monthly ENSO forecast skill and lagged ensemble size. *J. Adv. Model. Earth Syst.*, **10**, 1074-1086, doi: 10.1002/2017MS001204.

6. PI Contact Information

Timothy DelSole

Department of Atmospheric, Oceanic, and Earth Sciences
Center for Ocean-Land-Atmospheric Studies
George Mason University
112 Research Hall, Mail Stop 2B3
Fairfax, VA 22030 USA

Voice: 703-993-5715 *Fax:* 703-993-5770 *E-mail:* tdelsole@gmu.edu

7. Budget for Coming Year

Not applicable.

8. Future Work

We have reached the end of the award period and no further work will be pursued under this grant.

References

DelSole, T. and M. K. Tippett, 2014: Comparing Forecast Skill. *Mon. Wea. Rev.*, **142**, 4658–4678.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3–4 average temperature and precipitation over the contiguous united states. *Journal of Climate*, **30 (10)**, 3499–3512, doi:10.1175/JCLI-D-16-0567.1, URL http://dx.doi.org/10.1175/JCLI-D-16-0567.1, http://dx.doi.org/10.1175/JCLI-D-16-0567.1.

Kumar, A., M. Chen, L. Zhang, W. Wang, Y. Xue, C. Wen, L. Marx, and B. Huang, 2012: An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Mon. Wea. Rev.*, **140**, 3003–3016.

Tippett, M. K., M. Almazroui, and I.-S. Kang, 2015: Extended-range forecasts of areal-averaged Saudi Arabia rainfall. *Wea. Forecasting*, **30**, 1090–1105, doi:10.1175/WAF-D-15-0011.1.

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, URL http://dx.doi.org/10.1002/2016MS000838.

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

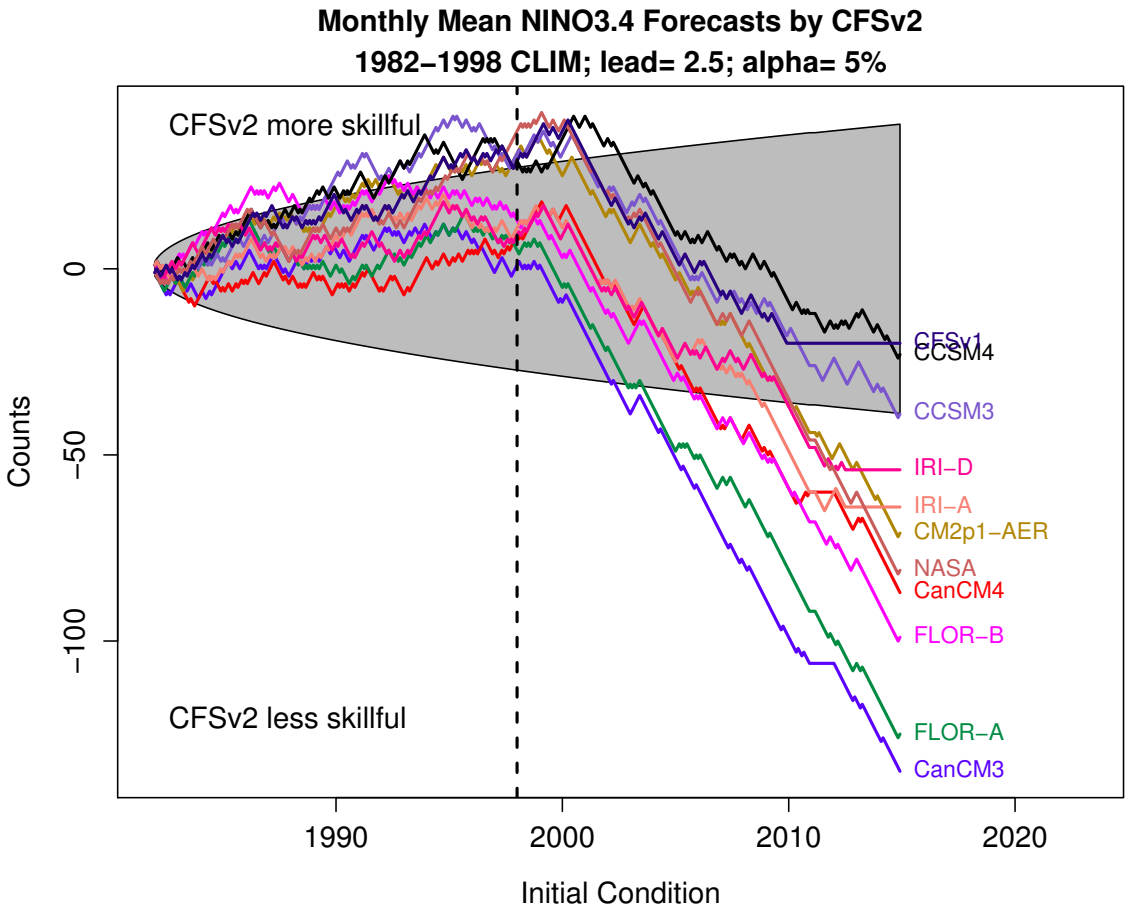

Figure 1: Comparison of monthly mean forecasts of NINO3.4 at lead 2.5 months between CFSv2 and other models in the NMME. All forecasts and observations are centered relative to the 1982-1998 period. The score is defined such that it is increased by one when the squared error of CFSv2 is less than that of another model, and decreased by one otherwise (the mean square errors are never exactly equal to each other). The scores are then accumulated forward in time for each model separately, over all initial months and years, thereby tracing out a random walk. The shaded area indicates the range of scores that would be obtained 95% of the time under the null hypothesis of equally skillful forecasts. A random walk extending above the shaded area indicates that CFSv2 forecasts are closer to observations significantly more often than expected under the null hypothesis (i.e., the CFSv2 is more skillful than the model). Similarly, a random walk extending below the shaded area indicates that CFSv2 forecasts are closer to observations significantly less often than expected under the null hypothesis (i.e., the CFSv2 is less skillful than the model).

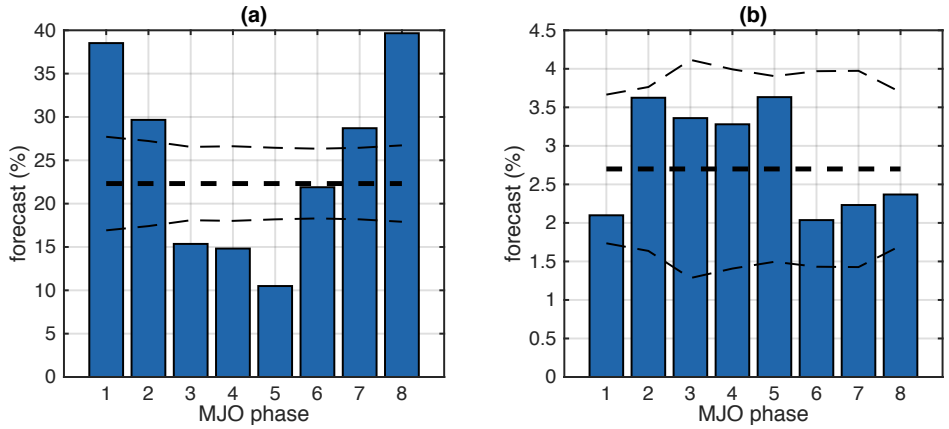

Figure 2: Average first-lead logistic regression forecast probabilities of heavy rainfall during the subsequent 5 days stratified by MJO phase three days prior to start for (a) Nov-Apr and (b) May-Oct season. Heavy and light dashed lines show the unconditional mean and its 95% confidence intervals, respectively, based on the number of forecasts in each phase.

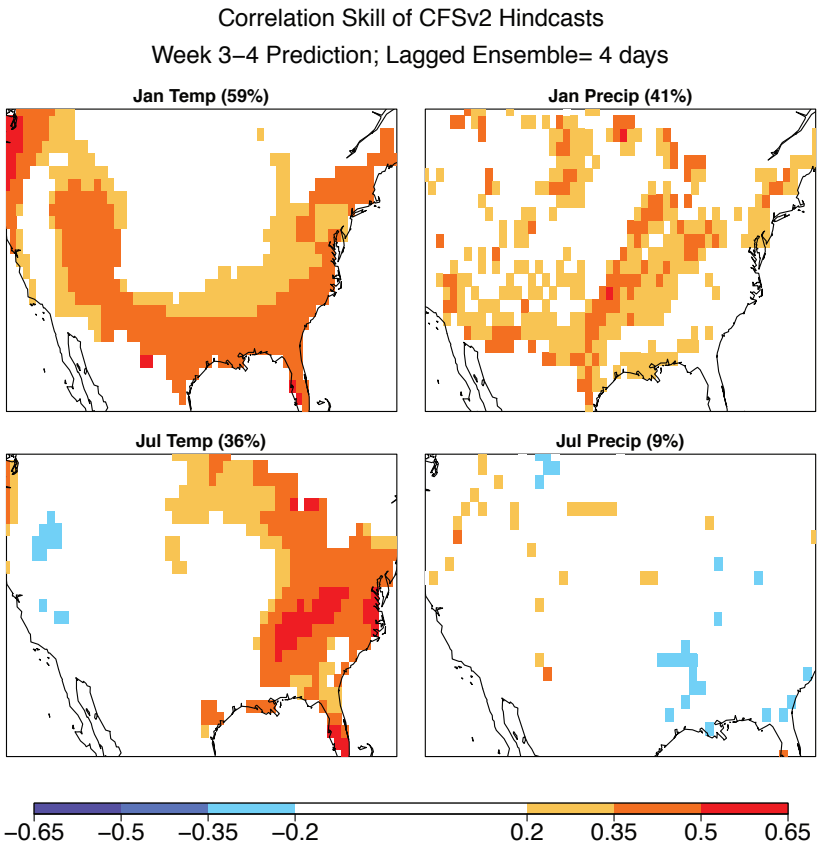

Figure 3: Correlation skill of week 3-4 temperature and precipitation CFSv2 hindcasts over CONUS during January and July, 1999-2010 (12 years). The hindcasts are based on a 4-day lagged ensemble (comprising 16 members drawn from 4x daily hindcasts). Values that are statistically insignificant at the 5% level (according to the permutation test) are masked out. The percentage area with significant correlation skill (positive and negative) is indicated in the title of each panel.

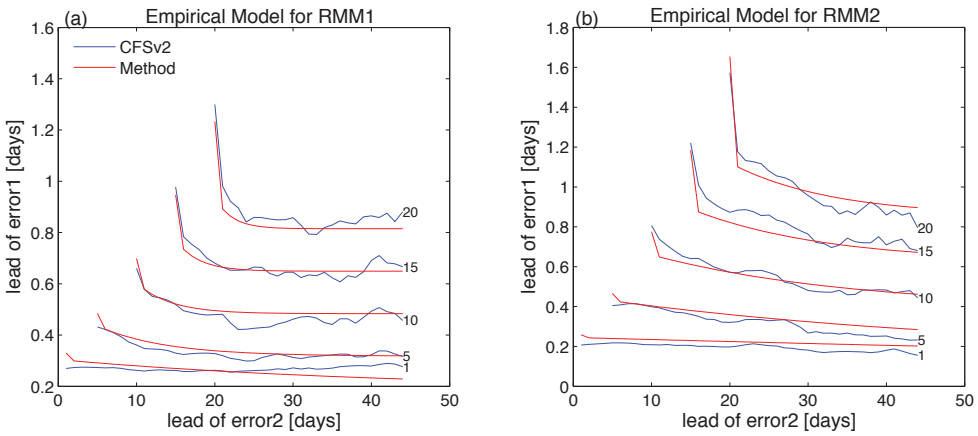

Figure 4: Cross section of the lagged error covariance for (a) RMM1 and (b) RMM2 in the CFSv2 are shown in blue. An 8 parameter empirical fits to the lagged error covariance matrices shown in red. The cross sections are for the 0Z initialization of the lagged ensemble forecast.

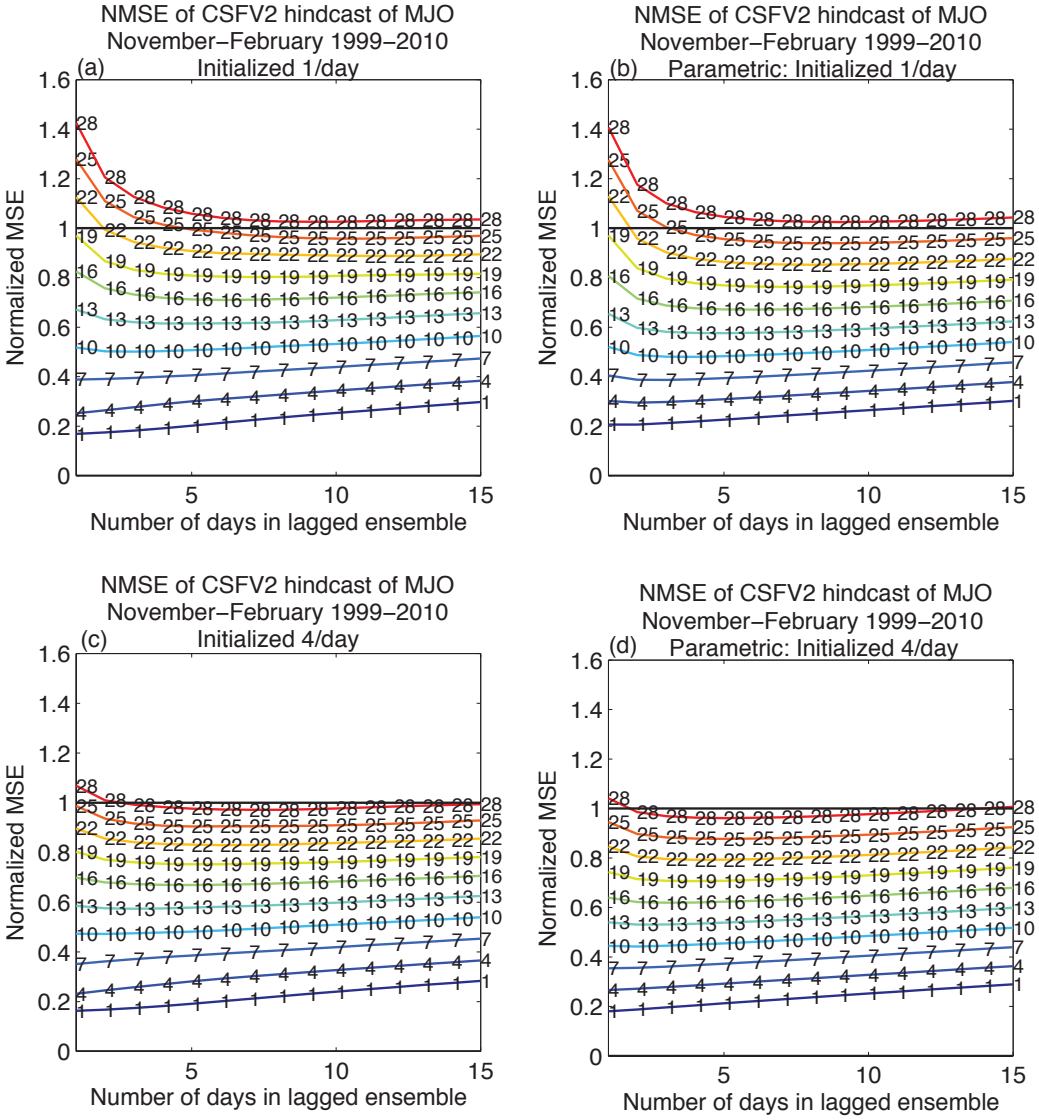

Figure 5: Normalized MSE of the boreal winter (Nov. 1 - Feb. 28, 1999-2010) MJO forecast as a function of lagged ensemble size (horizontal axis) and lead (colored curves - the number denotes forecast lead in days). The MSE is computed in terms of the standard Wheeler and Hendon (2004) RMM1/RMM2 indices. (a) normalized MSE for MJO forecast from the 0Z initialization of CFSv2. (b) Empirically derived normalized MSE computed using the fit shown in Fig 4. (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 Fig. 4 interpolated to include 4 separate initializations.

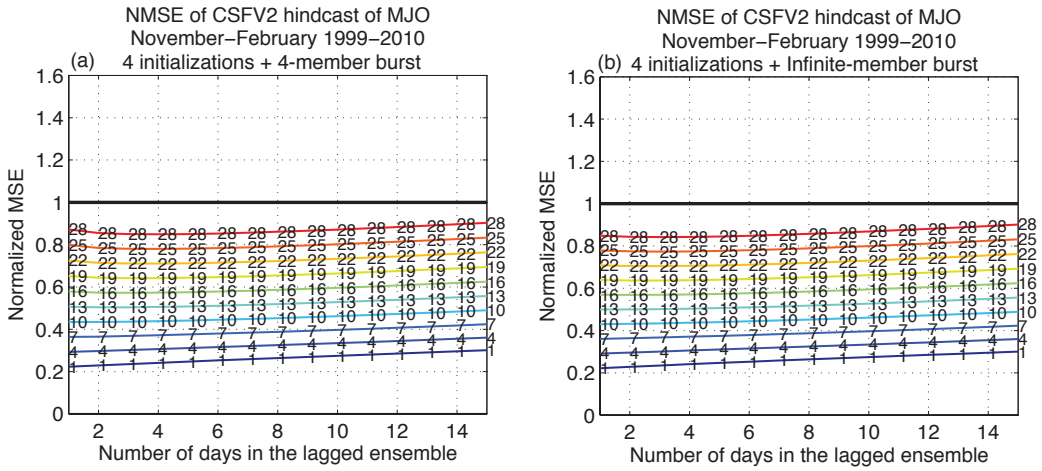

Figure 6: Empirically derived normalized MSE for CFSv2 hindcasts of MJO during boreal winter (1 November to 28 February) as a function of lagged ensemble size (horizontal axis) and lead (colored curves-the number denotes forecast lead in days). MSE is found using the parametric model fitted to error 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.

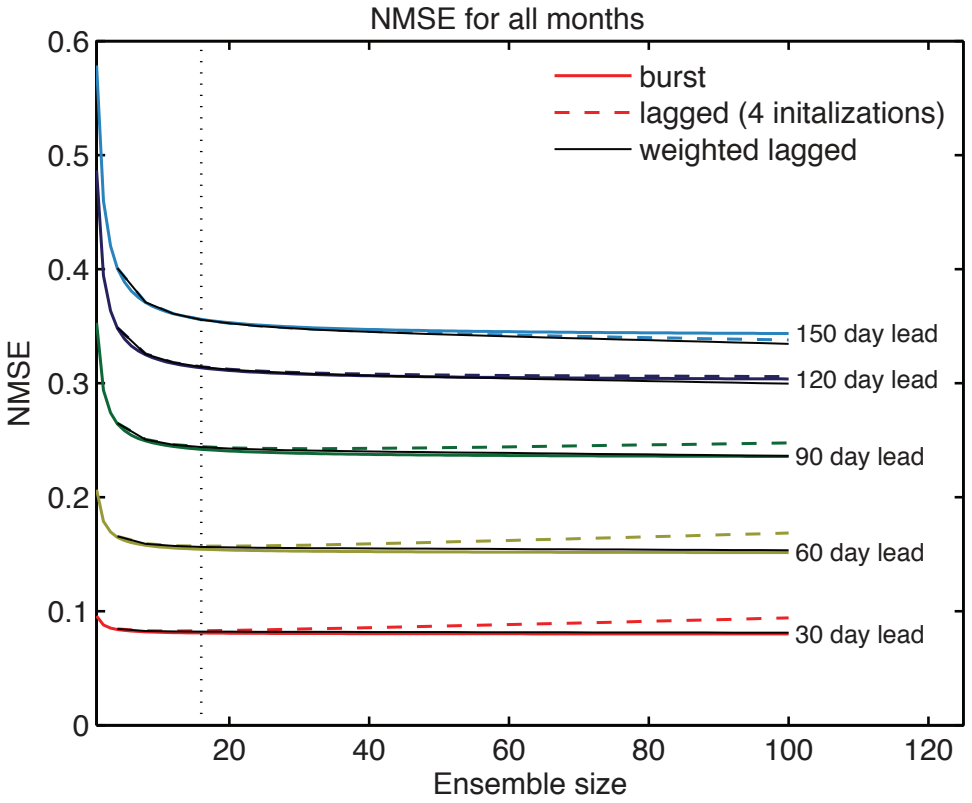

Figure 7: Parametrically derived normalized MSE for CFSv2 forecasts of Niño3.4 as a function of ensemble size for a "burst" ensemble (solid color curves), lagged ensemble with four initializations per day (color dashed curves), and optimally weighted lagged ensemble for forecasts initialized 4 times per day(black curves). Each set of color curves corresponds to MSE estimates for the specified lead time. Estimates for the "burst" ensemble are computed assuming the ensemble members are initialized an infinitesimal time step apart. The dotted vertical line denotes the location of the 8 day lagged ensemble when four separate initializations are included.

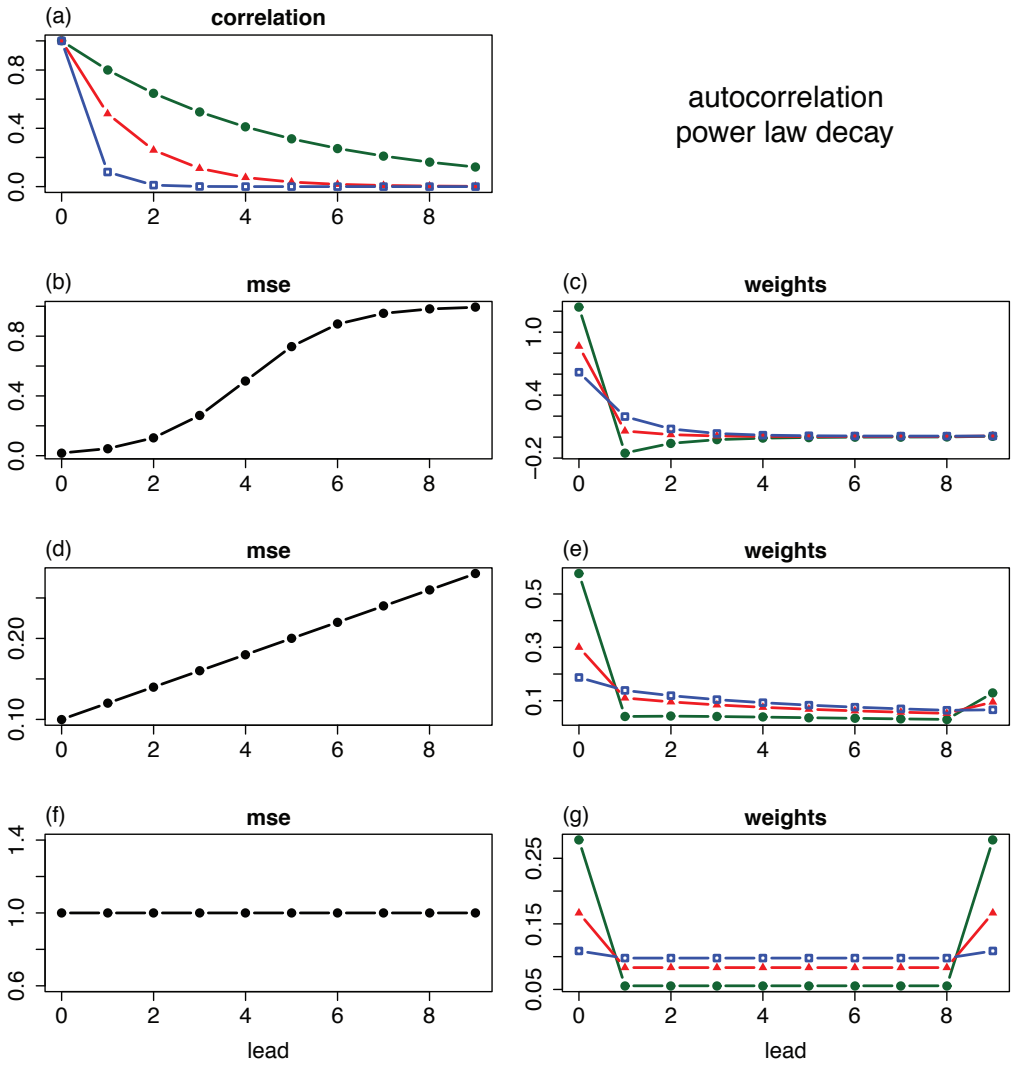

Figure 8: Illustration of the cross-lead error covariance matrix and corresponding weights of an optimal lagged ensemble. The covariance matrix can be decomposed in terms of a correlation matrix (**R**) and a diagonal matrix (**D**). Assuming the correlation matrix is given by power law decay, with correlations shown in (a), parameterized using the values ρ = (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 3a used to define the covariance matrix.

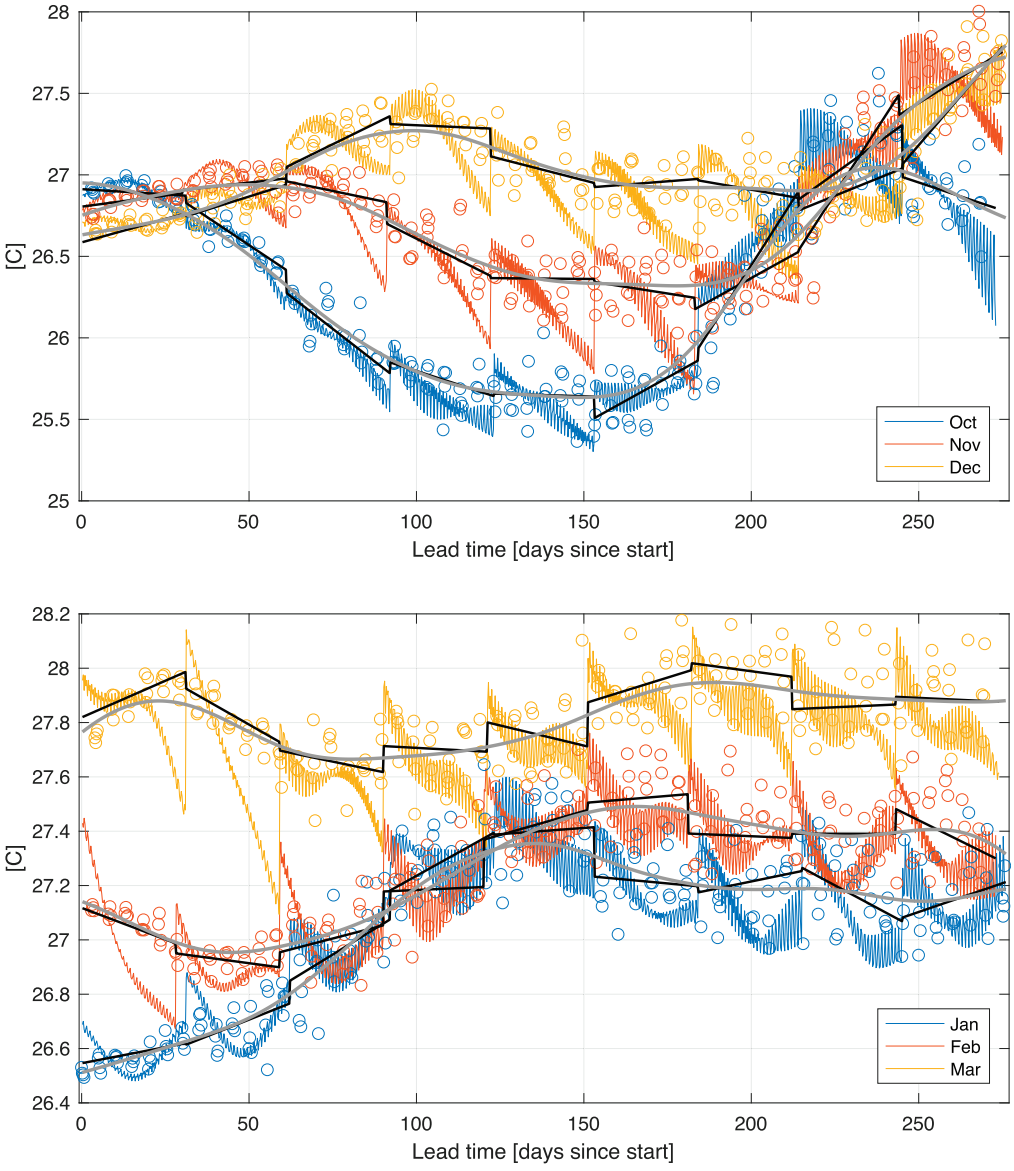

Figure 9: The CFSv2 Niño-3.4 index forecast climatology for (top) October, November, and December targets and (bottom) January, February, and March targets as a function of lead time as provided by EMC (jagged colored lines), fit to be periodic in target month T/linear in lead time L (black line segments) and estimated by local linear regression (smooth gray curve). Circles are hindcast averages.

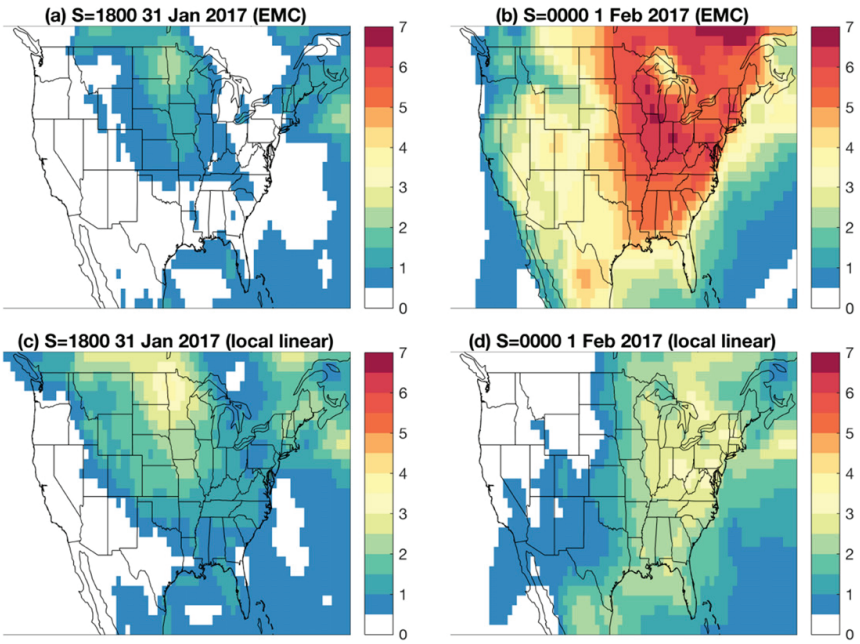

Figure 10: March 2017 2-m temperature anomalies with respect to the (a), (b) EMC climatology and the (c), (d) local linear forecast climatology for forecasts starting at (left) 1800 31 Jan 2017 and (right) 0000 1 Feb 2017. The label S denotes start time.