Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL085270

Special Section:
Bridging Weather and Climate: Subseasonal-to-Seasonal (S2S) Prediction

A Priori Identification of Skillful Extratropical Subseasonal Forecasts

John R. Albers[1,2] and Matthew Newman[1,2]

[1]Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA, [2]Physical Sciences Division, NOAA Earth System Research Laboratory, Boulder, CO, USA

Key Points:

- Subseasonal predictability is largely due to shifts in the ensemble-mean signal, at least for the Northern Hemisphere extratropical winter
- While some individual subseasonal forecasts have high skill, so can random chance, underlining the need to determine skill a priori
- A linear inverse model identifies the high skill subset (10%-30%) of operational model forecasts at Weeks 3-6, and both have similar skill

Supporting Information:

- Supporting Information S1

Correspondence to:
J. R. Albers,
john.albers@noaa.gov

Citation:
Albers, J. R., & Newman, M. (2019). A priori identification of skillful extratropical subseasonal forecasts. *Geophysical Research Letters*, *46*, 12,527–12,536. https://doi.org/10.1029/2019GL085270

Received 5 SEP 2019
Accepted 13 OCT 2019
Published online 6 NOV 2019

Abstract

The current generation of subseasonal operational model forecasts has, on average, low skill for leads beyond 3 weeks. This is likely a fundamental property of the climate system, due to the relative high amplitude of unpredictable weather variability compared to potentially predictable, but generally weaker, climate signals. Thus, for subseasonal forecasts to be useful, their high versus low skill events should be identified at time of forecast. We show that a linear inverse model (LIM), an empirical-dynamical model constructed from covariability statistics of wintertime (December–March) weekly averaged observational analyses, can be used to identify, a priori, the expected extratropical subseasonal surface and midtropospheric forecast skill. The LIM's predicted signal-to-noise ratio identifies the subset (10%–30%) of Weeks 3–6 forecasts—of the LIM and two operational models from the National Centers for Environmental Prediction and the European Centre for Medium-Range Weather Forecasts—with relatively higher skill versus the much larger remainder of forecasts whose skill cannot be distinguished from random chance.

Plain Language Summary

Our current understanding of weather prediction is that usable daily forecasts cannot be made more than 15 days in advance. This is a consequence of chaos: Any small initial uncertainty in our picture of the atmosphere (e.g., wind, temperature, and pressure) when the forecast is made will lead to errors growing to become as large as the weather we are trying to predict. Recently, however, focus has turned to “subseasonal” forecasts, predictions of weekly averaged weather made 3 to 6 weeks ahead, because climate phenomena (e.g., El Niño) sometimes produce persistent weather patterns that might be predicted even though individual storms within them cannot be. To identify when these “forecasts of opportunity” will occur, we developed a statistical subseasonal forecast model capable of predicting when its own forecasts—and those of sophisticated physical models from U.S. and European operational centers—will be usable. Our model successfully identifies the 20%–30% of forecasts at Weeks 3 and 4 and 10% of forecasts at Weeks 5 and 6 that are usable. Our results show a path forward to develop techniques for identifying usable subseasonal forecasts beforehand, so that practical forecast guidance may be given to end users in a variety of societal contexts.

1. Introduction

Subseasonal climate prediction is aimed at forecast leads between about 3–6 weeks, past the expected predictability limit of daily weather variations (e.g., Bauer et al., 2015; Simmons & Hollingsworth, 2002; Zhang et al., 2019). In principle, slowly evolving climate phenomena provide subseasonal skill (Butler et al., 2019; Vitart et al., 2017, and references therein) despite the loss of predictability due to error growth associated with chaotic nonlinearities (Lorenz, 1963, 1969; Weber & Mass, 2017). Unfortunately, in the extratropics, unpredictable daily weather has such high amplitude that subseasonal forecast skill is, on average, quite low, and can be highly variable even when climate anomalies are large (e.g., El Niño–Southern Oscillation; Zhang et al., 2018). This has led to a focus on identifying “forecasts of opportunity,” those subseasonal forecasts with sufficiently high skill to be useful (e.g., National Academies of Sciences, Engineering, and Medicine, 2016).

If we view subseasonal forecast skill as inherently limited by largely unpredictable weather noise, then subseasonal predictability (i.e., the potential for skill) can be assessed from the signal-to-noise ratio, which may be observationally determined (e.g., Chen & van den Dool, 1999; Feng et al., 2013; Madden, 1976; Shukla & Gutzler, 1983). This approach often implicitly assumes that variability can be separated into “slow” (e.g., climate) and “fast” (e.g., weather) time scales. In this case, especially when predicting only the slow subseasonal component, nonlinear dynamics may be well approximated by stochastically forced linear dynamics, which in turn may be empirically captured with a linear inverse model (LIM; Penland & Sardeshmukh, 1995; Winkler et al., 2001). The LIM has been shown to produce extratropical subseasonal forecasts with skill comparable to comprehensive numerical models, allowing it to quantify the forecast signal-to-noise ratio and thereby to estimate both overall limits of potential skill (Pegion & Sardeshmukh, 2011) and, more importantly, the expected skill of each individual forecast (Newman et al., 2003; hereafter N2003). The LIM's potential to "forecast the forecast skill" (Tennekes et al., 1986) is particularly promising, especially since estimating perfect model predictability from physical models currently offers relatively poor guidance for identifying high skill forecasts (Pegion et al., 2017).

For the LIM to assess subseasonal forecast skill of an ensemble prediction system, however, predictability needs to be primarily due to variations in the forecast ensemble mean, which a LIM can capture, rather than to variations in the forecast ensemble spread, which it cannot. Previous LIM studies have suggested this might be the case, but they focused on shorter forecast leads of 2–3 weeks and did not evaluate skill of operational forecast models. In this paper, we test this hypothesis by developing a new LIM to examine skill of the subseasonal (Weeks 3–6) surface and midtropospheric forecasts from two current generation operational models, the European Centre for Medium Range Forecast Integrated Forecasting System (ECMWF IFS-CY43) and the National Centers for Environmental Prediction Climate Forecast System version 2 (NCEP CFSv2). For all three models, we find that while median forecast skill at leads beyond Week 3 is quite low, a small subset of forecasts has notably higher skill occurring more frequently than expected from random chance. To isolate these forecasts, we use the LIM's signal-to-noise ratio to compute its expected perfect model ensemble mean skill, which can identify skillful forecasts a priori for the Pacific and Atlantic basins for both the LIM and the operational models.

2. Models and Metrics

2.1. Linear Inverse Model

The nonlinear dynamics of an anomalous climate (coarse grained) state vector $\boldsymbol{x}$ may be approximated with the stochastically forced, stable linear dynamical system:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{L}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\eta} \qquad (1)$$

(e.g., Hasselmann, 1976; Just et al., 2001; Penland, 1996), where the linear operator ($\boldsymbol{L}$) represents the sum of both the linearized and the linearly parameterizable nonlinear physics of the climate system and $\boldsymbol{B\eta}$ represents forcing by the unpredictable (i.e., rapidly decorrelating) nonlinear remainder, with $\boldsymbol{B}$ a noise amplitude matrix and $\boldsymbol{\eta}$ a vector of unit variance white noise. Then, as in N2003, $\boldsymbol{B}$ is assumed state independent and constant, so that (on average) there is no covariance between the noise and atmospheric state. The infinite-member ensemble mean forecast at lead τ is $\widehat{\boldsymbol{x}}(t+\tau) = \exp[\boldsymbol{L}\tau]\boldsymbol{x}(t)$, with individual ensemble members derived from different noise realizations and integration of (1) over the interval $[t, t+\tau]$.

Following N2003, we estimate (1) from observations by constructing a LIM. We define the state vector using Japanese Meteorological Agency 55-year Reanalysis (JRA-55; Kobayashi et al., 2015) data, coarse-grained in time with a 7-day running mean filter and in space by area averaging onto a 5° grid, to obtain anomalies from 1979 to 2015 climatologies (corresponding to relevant cross-validation periods; section 2.2) of mean-sea level pressure (MSLP, 0°–90°N); tropospheric stream function (0°–90°N at 750 hPa); 500 hPa geopotential height (0°–90°N); stratospheric stream function (30°–90°N, at 100 and 10 hPa); and a measure of column integrated diabatic heating (30°S–30°N), for the extended winters (December–March) of 1979–2015. Additional details on LIM construction and discussion of LIM variable choice are in supporting information Text S1; none of these choices materially affect our results. This LIM is improved relative to N2003 because it uses a 4DVar-based reanalysis and represents both surface and vertically deep stratospheric variability (Albers & Birner, 2014; Birner & Albers, 2017; Hitchcock & Haynes, 2016; O'Neill & Pope, 1988), allowing future diagnostic studies of dynamical processes to go beyond earlier LIM analyses (e.g., Newman & Sardeshmukh, 2008; Winkler et al., 2001).

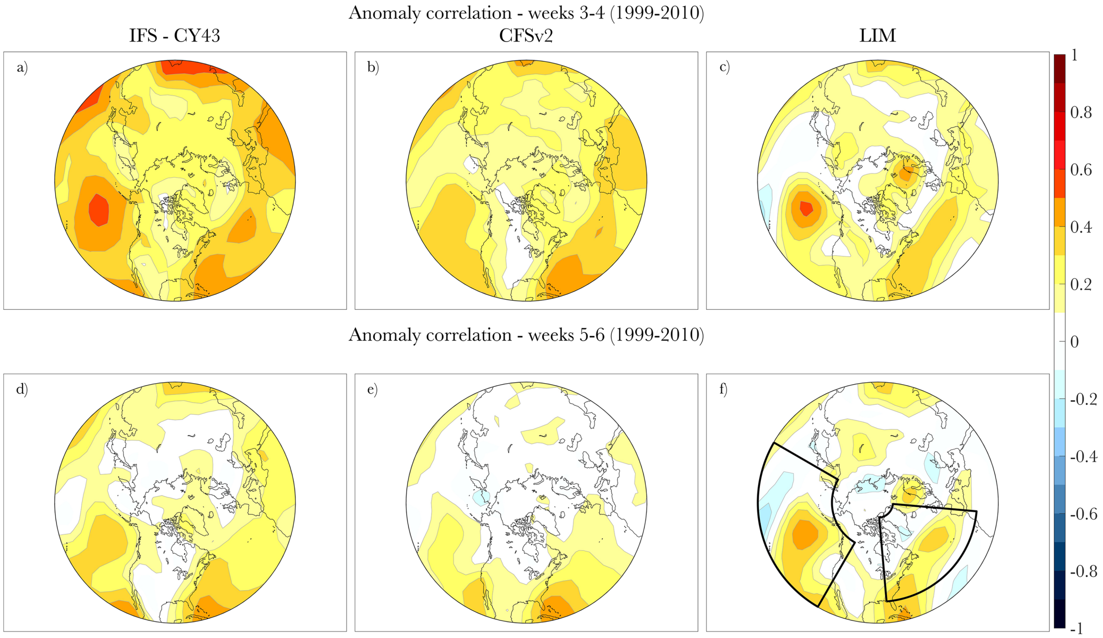
Anomaly correlation - weeks 3-4 (1999-2010)
IFS - CY43
CFSv2
LIM
a)
b)
c)
Anomaly correlation - weeks 5-6 (1999-2010)
d)
e)
f)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 1. IFS-CY43 (a, d), CFSv2 (b, e), and LIM (c, f) ACs for Weeks 3/4 (top panels) and Weeks 5/6 (bottom panels) MSLP hindcasts, for 1999–2010 common model period only (Weeks 3/4 include 536 forecasts and Weeks 5/6 include 440 forecasts). The two boxes in the bottom LIM panel show the Pacific and Atlantic regions used for all APC calculations.

2.2. Hindcasts

We examine hindcasts of two operational models, both obtained from the Subseasonal-to-Seasonal Prediction Project database (Vitart et al., 2017): ECMWF's IFS CY43R1/R3, operational in 2017, and NCEP's CFSv2, operational starting in 2011. Common initialization dates are used whenever comparing hindcast skill between models. The “three-model common period” is therefore 1999–2010, sampled at the twice weekly frequency of the IFS-CY43 hindcasts. We also use the full IFS-CY43 period (twice weekly) or the full LIM period (daily). See supporting information Text S2 for details of these models, including their hindcast sample sizes. To focus on boreal winter, we only include forecast initialization dates on or after 1 December that also have verification dates on or before 15 March.

IFS-CY43 and CFSv2 hindcasts are coarse-grained (section 2.1) to match the LIM hindcasts and verification data sets, and their anomalies are computed by removing the lead dependent and model specific climatologies, which also serves as a mean bias correction. Tenfold cross-validated LIM hindcasts are calculated for ~4-year periods where $\boldsymbol{L}$ is trained on the JRA-55 data remaining after removal of the relevant hindcast period. When constructing the cross-validated $\boldsymbol{L}$, the time series is recentered, ensuring that as τ increases the LIM forecasts asymptote to climatology.

2.3. Predictability and Skill Metrics

Assessing predictability within the LIM is straightforward since (1) has distinctly predictable and unpredictable dynamics. The *expected* deterministic skill of any perfect model infinite-member ensemble-mean forecast for lead τ at forecast initialization time t is

$$\rho_\infty(t;\tau) = \frac{S^2(t;\tau)}{\left[\left(S^2(t;\tau)+1\right)S^2(t;\tau)\right]^{1/2}} \tag{2}$$

(Sardeshmukh et al., 2000), where the forecast signal-to-noise ratio, $S^2(t;\tau)$, is determined in the LIM from the state-dependent forecast signal $\widehat{\boldsymbol{x}}(t+\tau)$ and the expected state-independent, forecast lead-dependent error variance (N2003). That is, for a given forecast lead, since each LIM forecast distribution has identical shape (e.g., spread) and differs *only* in its mean displacement from zero, ρ_∞ is sufficient to describe the predictability of (1) (N2003; Chang et al., 2004).

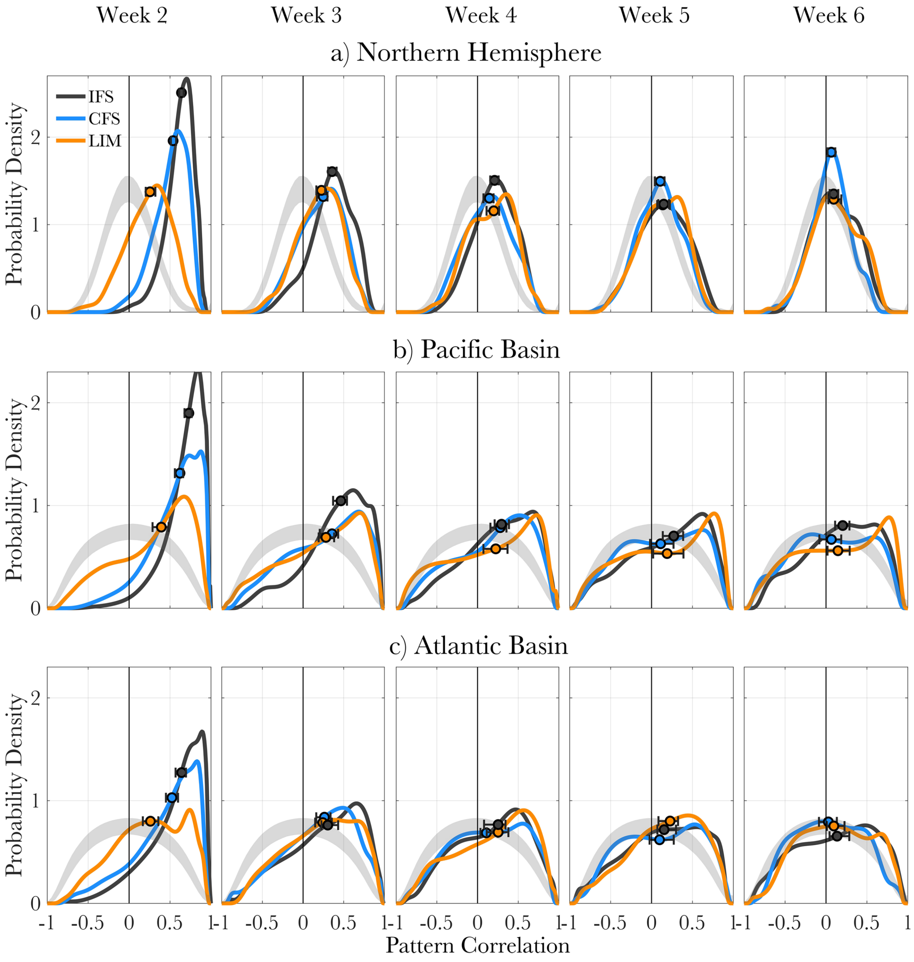

Figure 2. Weeks 2–6 APC distributions for NH (a), Pacific (b), and Atlantic (c) MSLP hindcast skill for the common model period. Circles denote median values; whiskers denote bootstrap confidence intervals. The 2.5th to 97.5th percentile range of the forecast distributions from the Monte Carlo set of random-skill realizations (section 2.3) is shown with gray shading. See supporting information Figure S2 for sample sizes at each forecast lead.

Hindcast skill can be similarly measured by correlating ensemble-mean forecast anomalies (f') to observed verification anomalies (v'). Following other subseasonal prediction studies (e.g., N2003; Robertson & Vitart, 2018, and references therein), we use local anomaly correlation (AC) and uncentered, area-weighted anomaly pattern correlation (APC), written as

$$[AC(x,y), APC(t)] = \left[\frac{\sum_t \left(f_i' - \overline{f'}\right)\left(v_i' - \overline{v'}\right)}{\left[\sum_t \left(f_i' - \overline{f'}\right)^2 \sum_t \left(v_i' - \overline{v_i'}\right)^2\right]^{1/2}}, \frac{\sum_{x,y}\left(f_{i,j}'\right)\left(v_{i,j}'\right)}{\left[\sum_{x,y}\left(f_{i,j}'\right)^2 \sum_{x,y}\left(v_{i,j}'\right)^2\right]^{1/2}} \right],$$

where (x,y) denote longitude and latitude and t denotes time. For AC, each time series is additionally centered about its mean, indicated by overbars. APC is computed for the NH (20°–90°N), Pacific basin (20°–60°N and 150°–240°E), and Atlantic basin (30°–80°N and 275°–355°E; see Figure 1f). Other related metrics yield qualitatively consistent results (see supporting information). We define minimum "useful" skill as correlations >0.5–0.6, following the criterion outlined by Arpe et al. (1985) and Murphy and Epstein (1989). Bootstrapping is used to estimate APC confidence intervals (see supporting information).

For a perfect model forecast ensemble evolving from some initial state, ρ_∞ is the expectation value of skill taken over all forecast outcomes weighted by their relative probabilities. However, for a given forecast only one outcome occurs, whose actual skill may be above or below ρ_∞. Also, case-to-case variations in

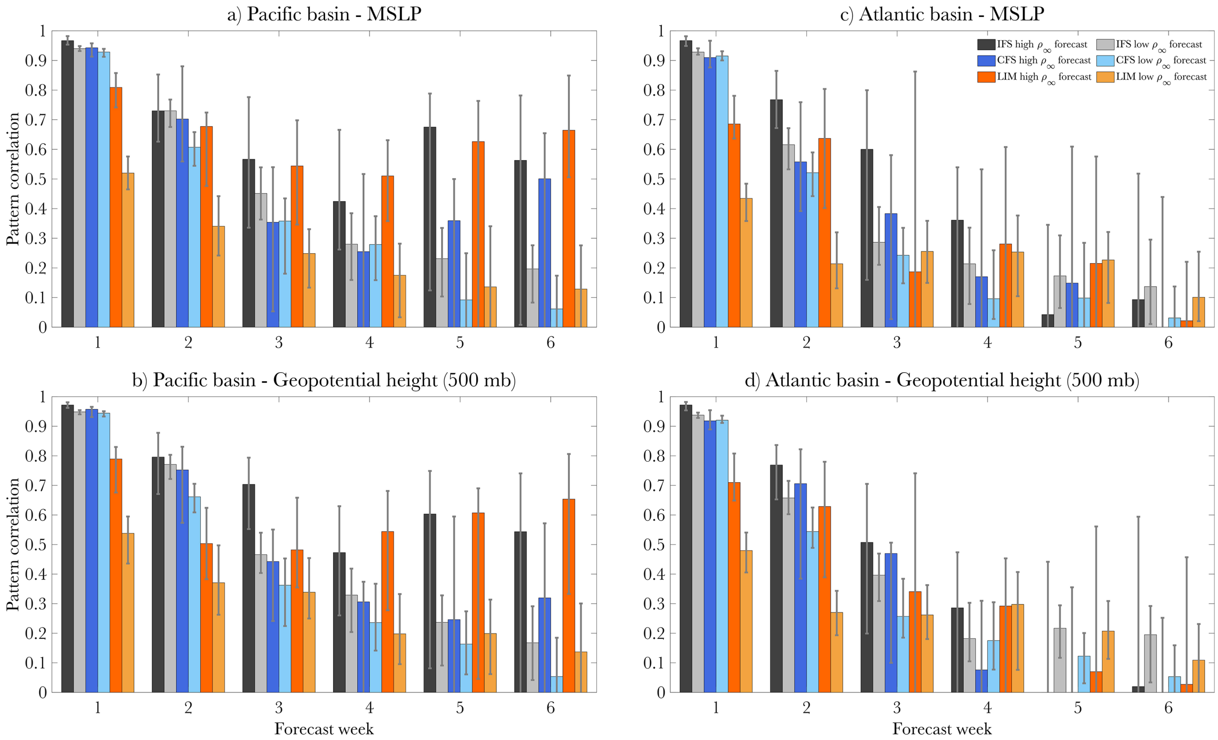

Figure 3. Pacific (a, b) and Atlantic (c, d) basin MSLP (top row) and geopotential height (bottom row) median APCs for hindcasts (over the common model period) corresponding to the 90th–100th percentile of ρ_∞ values (darker bars) versus the remaining 90% of hindcasts (lighter bars). Whiskers denote bootstrap confidence intervals. See supporting information Figure S2 for sample sizes at each forecast lead.

predictability (ρ_∞) give rise to variations in actual skill. To help isolate predictable from random skill variations, we use as our null hypothesis forecasts that are "randomly chosen states of the system" (Lorenz, 1969). In this "random skill" model, every forecast is a random draw from the climatological distribution of observed anomalies, with a Monte Carlo approach taken to address the limited observational sample (see supporting information for details) and determine confidence intervals. This model is perfect (and reliable) since all its forecast states are drawn from nature's attractor, with no bias, but its ρ_∞ is zero. Its individual forecasts, however, have a range of skill values whose distribution serves as a baseline for model hindcast skill distributions.

3. Results

3.1. Overall Skill

For all three models (IFS-CY43, CFSv2, and the LIM), median weeks 3–6 skill is generally low. The LIM and operational models have broadly comparable extratropical skill, with similar spatial features for both MSLP (Figure 1; see Figure S2 for MSESS) and 500-hPa geopotential height (Figures S3 and S4). Skill of MSLP hindcasts is largest over the Pacific and Atlantic basins but few locations at weeks 3 and 4, and no locations at weeks 5 and 6 have median AC values representing useful (see section 2.3) skill.

We assess how often individual hindcasts have useful skill from the distribution of hindcast APC skill values for each lead. Figure 2 shows MSLP distributions, evaluated separately for the NH, Pacific basin, and Atlantic basin (Figure S5 shows week 1; Figure S6 shows similar results for geopotential height). Distributions are derived via kernel density estimation (see supporting information and Figure S7 for details). Not surprisingly, for increasing forecast lead time, the median skill of all three models decreases toward zero and the distributions become notably less skewed, with high skill hindcasts occurring less frequently and low skill hindcasts occurring more frequently, especially for the larger NH domain.

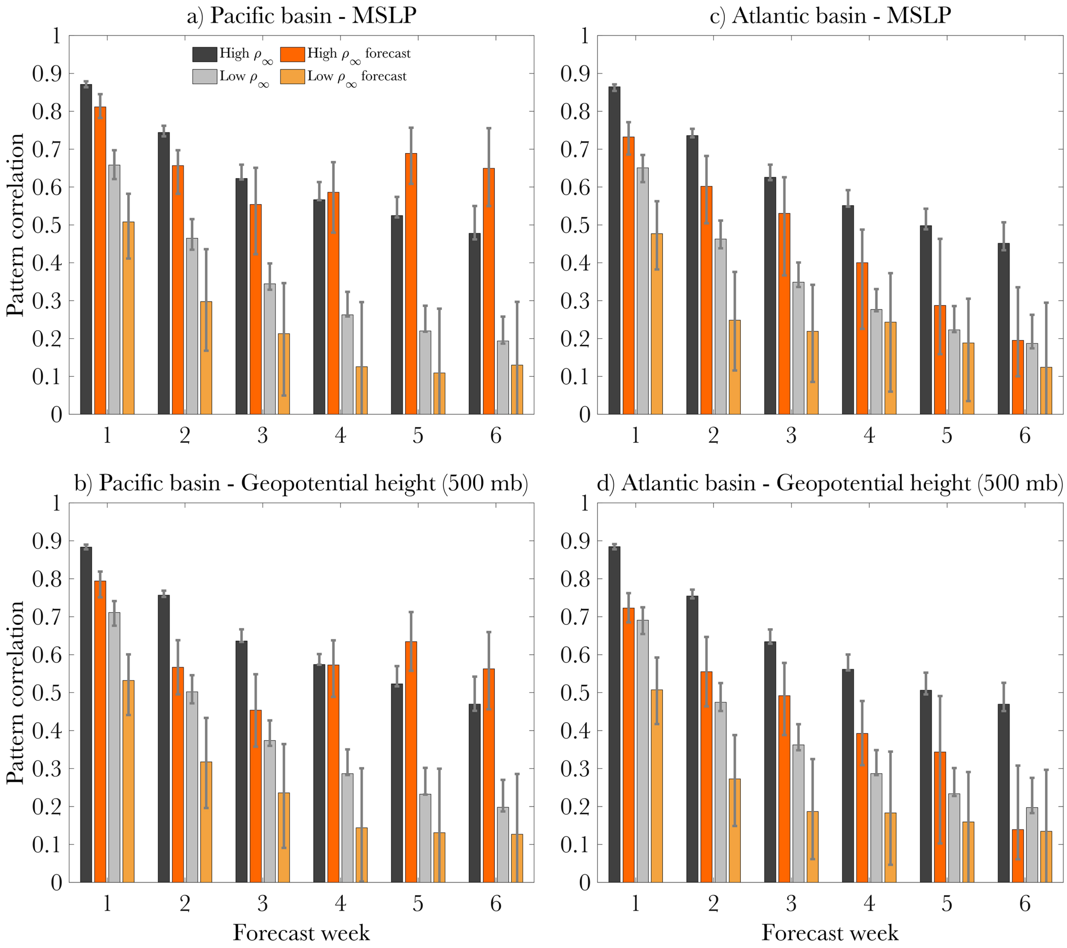

Figure 4. Pacific (a, b) and Atlantic (c, d) basin MSLP (top row) and geopotential height (bottom row) median APCs (1979–2015) for hindcasts corresponding to the 90th–100th percentile of ρ_∞ values (dark orange bars), actual 90th–100th percentile of ρ_∞ values (dark gray bars), the remaining 90% of hindcasts corresponding to the 90th percentile of ρ_∞ (light orange bars), and the actual ρ_∞ values in the 90th percentile (light gray bars). Whiskers denote bootstrap confidence intervals. See supporting information Figure S2 for sample sizes at each forecast lead.

Given how many hindcasts have low (even *negative*) skill, do all hindcasts with relatively high skill reflect true predictability? To answer this, Figure 2 also shows distributions from the random skill model, whose 95% confidence interval is shaded, determined separately for each region. These distributions illustrate the difference between a skillful model (e.g., at Week 2, the IFS-CY43 and CFSv2 distributions are largely disjoint from the random skill distributions) versus a low-skill model whose skill distribution is not significantly different from random chance (e.g., at Week 6, the CFSv2 for the Atlantic basin). At Weeks 4–6, the models all have median skill < 0.2, statistically indistinguishable (i.e., the error bars of the median values overlap) from the random-skill model (Figure 2), yet their skill distributions have enough negative skew—particularly for the two ocean basins—to be significantly different, even at Week 6. The key question then becomes: How can those forecasts with predictable high skill be identified a priori?

3.2. A Priori Skill

In this section, for each lead time τ we relate actual hindcast skill to the LIM's expected perfect model ensemble-mean skill $\rho_\infty(\tau)$, which varies for each initialization time t. With increasing lead, ρ_∞ typically decreases monotonically as the signal-to-noise ratio decreases. However, for some initial states, ρ_∞ will be larger and decrease more slowly with lead time; on average, these higher ρ_∞ events should be more predictable and correspond to higher skill forecasts.

First, for each model we separated hindcasts whose predicted skill (ρ_∞) lay in the upper decile from the remaining 90% of hindcasts. For the Pacific region, the upper ρ_∞ decile indeed corresponds to higher

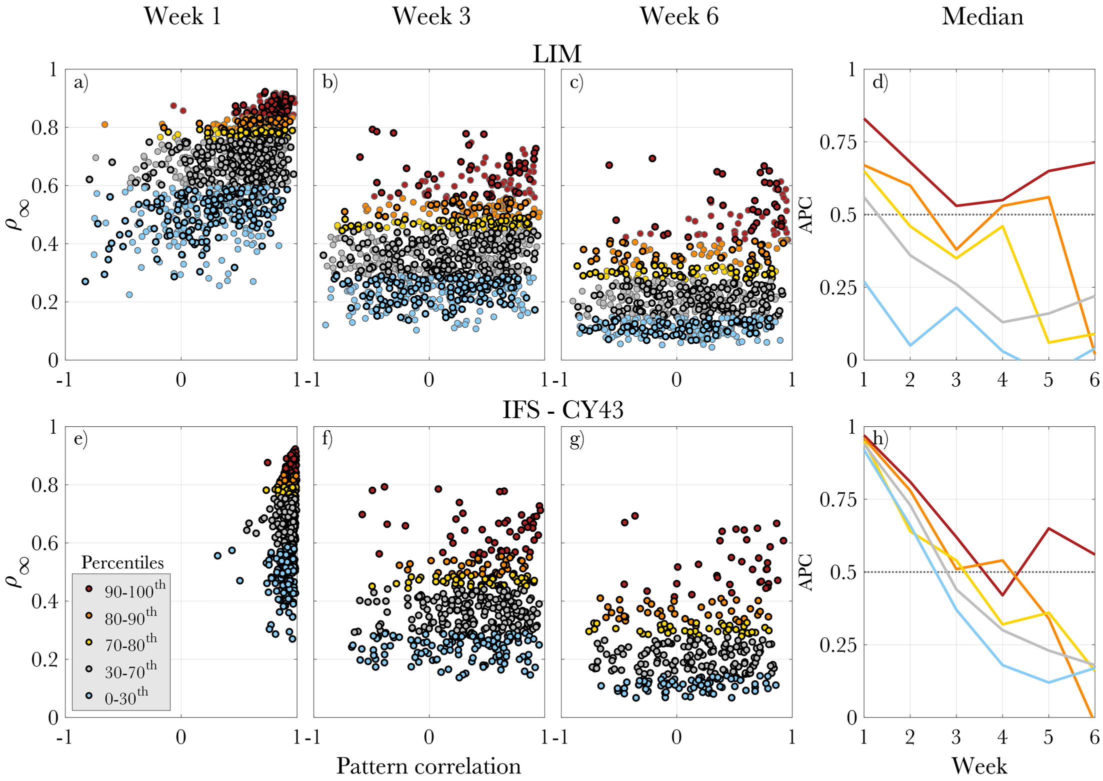

Figure 5. Pacific basin APCs for the LIM (top row; 1979–2015 as gray outlined dots; 1996–2015 IFS-CY43 sampling as bold black outlined dots) and IFS-CY43 (bottom row; 1996–2015 all forecasts) versus corresponding ρ_∞ values for MSLP hindcast Weeks 1 (a, e), 3 (b, f), and 6 (c, g). Colors indicate the upper tercile of ρ_∞ hindcasts in 10% increments (red, orange, and yellow), the lower tercile of ρ_∞ hindcasts (blue), and the remaining middle 40% of the data (gray). Median APCs for each percentile category at all hindcast leads are also shown (d, h). See supporting information Figure S2 for sample sizes at each forecast lead.

median skill of both MSLP and geopotential hindcasts (Figures 3a and 3b), though the difference between high and low ρ_∞ conditioned skill is larger for the LIM than for the IFS-CY43 and CFSv2. For the Atlantic region (Figures 3c and 3d), conditioning on ρ_∞ identifies higher skill through Week 4 for MSLP and Weeks 3 or 4 for geopotential, depending on the model, though these differences are within the bootstrap confidence intervals. For all three models the Atlantic relationship fails for Weeks 5 and 6.

The poorer results in the Atlantic might mean that the three-model common period has an inadequate sample size to quantify the potential for high skill. While the full IFS-CY43 period provides little improvement (not shown), the longer 1979–2015 period does allow the upper decile of ρ_∞ to generally identify higher LIM hindcast skill for all leads (cf. orange bars for Figures 3 and 4). Additionally, Figure 4 shows that in both basins, bootstrap confidence intervals for the high ρ_∞ cases are narrowed for the longer period, while those for the low ρ_∞ cases remain wide, reflecting the inherently greater forecast uncertainty typifying cases with low signal-to-noise ratio.

To differentiate between predictable and random high skill, we compared ρ_∞ and actual skill for every Pacific basin MSLP hindcast (Figure 5 for selected leads and Figure S8 for all leads). Red, orange, and yellow shading indicates whether ρ_∞ lies in the top, second, or third decile, respectively, with remaining hindcasts shaded gray (next 40%) or blue (bottom 30%). For each ρ_∞ category, as forecast lead increases both the median skill decreases (Figures 5d and 5h) and the range of skill increases, until eventually the distribution of skill cannot be distinguished from the random-skill model. However, this transition (from mostly predictable to mostly random skill) occurs at later forecast lead times for the top ρ_∞ categories than for the bottom ρ_∞ categories. For example, the IFS-CY43 skill distribution at Weeks 1 and 2 has almost uniformly high skill and is largely disjoint from the random-skill distribution (Figures S5 and 2b), with useful median skill for every ρ_∞ percentile (Figures 5e, and 5h). In contrast, by Week 6 the small portion of the IFS-CY43 skill distribution that significantly extends outside the random-skill distribution (Figure 2b) is largely due to the top ρ_∞ decile (denoted by the red dots and red line in Figures 5g and 5h, respectively), while the remaining (~90%) hindcast skill values appear largely random.

At Week 1, all IFS-CY43 hindcast categories have high skill, while the LIM reaches useful skill only for the ρ_∞ upper tercile (Figure 5a). This is entirely unsurprising, since the IFS-CY43 also has daily skill while the LIM, approximating daily variability as white noise, does not, although it is interesting that both LIM and IFS-CY43 Week 1 skill are stratified by ρ_∞. By Week 3, however, IFS-CY43 forecasts have useful skill for only the upper tercile of ρ_∞. By Week 4, only ~20% of the forecasts for either model are useful, decreasing to ~10% by Weeks 5 and 6. Overall, Figure 5 shows that ρ_∞ successfully sorts high from low skill forecasts for all percentile categories, at all leads, for both the LIM and IFS-CY43.

Finally, recall that LIM assesses predictability by assuming that skill results from changes in the mean of the forecast PDF, rather than from changes in its shape. To test this assumption, we repeated Figure 5 but compared hindcast skill to a standard measure of ensemble spread (Whitaker & Loughe, 1998) determined from each IFS ensemble for each initialization. Ensemble spread was a significantly weaker and less consistent predictor of skill than ρ_∞ (Figure S9; also cf. Tables S3–S5), supporting our use of ρ_∞ to identify a priori ensemble mean skill.

4. Concluding Remarks

While weather predictability is typically related to the spread of the forecast ensemble (e.g., Grimit & Mass, 2007; Hopson, 2014; van Schaeybroeck & Vannitsem, 2016), seasonal predictability often appears better related to variations in the forecast ensemble mean (e.g., Doblas-Reyes et al., 2000; Tang et al., 2008, Chen & Kumar, 2015, Pegion et al., 2017; Newman & Sardeshmukh, 2017). The nature of predictability on the intermediate subseasonal time scales has been less clear, although studies have found the relationship between predictability and spread to considerably weaken by Week 2 (e.g., Barker, 1991; Whitaker & Loughe, 1998; Hopson, 2014). In our study, a stochastically forced, linear dynamical model (namely, a LIM derived from observations) is used to demonstrate what earlier analyses (N2003; Pegion & Sardeshmukh, 2011) have suggested: Subseasonal predictability is largely due to shifts in the ensemble-mean signal, at least for the Northern Hemisphere extratropical wintertime variables analyzed here. Specifically, (1) the LIM's deterministic skill is competitive with ensemble-mean skill from two operational models, NCEP's CVSv2 and ECMWF's IFS-CY43, and (2) the LIM's expected skill ρ_∞ can identify the small subset (~10%–30%, depending upon forecast lead) of subseasonal MSLP and 500-hPa geopotential height forecasts with predictable "useful" skill (correlations >0.5–0.6) *for both itself and the operational models*, with skill of the remaining forecasts statistically indistinguishable from chance (i.e., climatology). Even over the short 1999–2010 three-model common period, the LIM can predict hindcast skill in the Pacific basin, correctly stratifying IFS-CY43 and LIM forecast skill for all ρ_∞ percentiles at all leads. In the Atlantic, the relationship is weaker so that statistical significance may require more than the 20 years of twice weekly IFS-CY43 hindcast data available, but the 36 years of LIM hindcast data appears sufficient.

Ours is not the only study to identify forecasts of opportunity using deterministic metrics of hindcast skill, despite the inherently probabilistic nature of the subseasonal forecast problem. For example, many studies have related state dependence of hindcast skill to specific climate phenomena such as El Niño–Southern Oscillation, the Madden-Julian oscillation, and sudden stratospheric warmings (e.g., Barnston et al., 2017; DelSole et al., 2017; Kim et al., 2018; Tripathi et al., 2015; Vitart & Molteni, 2010). The LIM has also been used in this manner (Winkler et al., 2001; N2003), but here we exploited its ability to estimate ρ_∞ and thereby quantify the predictable signal from the combination of *all* such sources of predictability, which is different at every forecast time and forecast lead. Other studies have explored hindcast skill dependence on different weather regimes, often suggested to result from enhanced nonlinear predictability resulting from locally reduced forecast uncertainties (e.g., Ferranti et al., 2018; Vigaud et al., 2018). However, such state-dependent skill could also be consistent with the LIM framework (N2003), especially if the LIM could be constructed to include "correlated additive-multiplicative noise" (Martinez-Villalobos et al., 2019; Sardeshmukh & Penland, 2015), which allows for non-Gaussianity (Sura et al., 2005; Sardeshmukh et al., 2015) while impacting only the unpredictable denominator of S in (2).

Future operational model skill might eventually eclipse that of any LIM-based model, although it is noteworthy that N2003's fundamental conclusions remain relevant even after nearly two decades of model development. Still, regardless of the method used to identify high skill periods, our results suggest that relatively small forecast ensembles may be enough to determine case-to-case variations in subseasonal predictability, as they are mostly due to mean shifts in the ensemble forecast distribution. In contrast, it may take very large forecast ensembles to realize significant additional predictability gains from higher order moments of the ensemble distribution. For example, Buizza and Leutbecher (2015) were unable to establish significant mean Week 4 predictability with a 51-member forecast ensemble, a size that has yet to be used for the lengthy hindcast ensembles likely required for adequate probabilistic calibration and comprehensive predictability analysis. Meanwhile, the much cheaper LIM is available to be exploited now to guide forecasts of forecast skill for every operational subseasonal forecast ensemble.

Acknowledgments
This work was partly supported by NOAA/CPO/MAPP. LIM hindcasts are available at ftp://ftp2.esrl.noaa.gov/Projects/LIM/Weekly/AlbersNewmanGRL2019/ website. S2S data was originally accessed via the ECMWF data portal: https://apps.ecmwf.int/datasets/ website. The authors would like to thank Yan Wang for preparing the IFS/CFS S2S data.

References

Albers, J. R., & Birner, T. (2014). Vortex preconditioning due to planetary and gravity waves prior to sudden stratospheric warmings. *Journal of the Atmospheric Sciences*, *71*(11), 4028–4054. https://doi.org/10.1175/JAS-D-14-0026.1

Arpe, K., Hollingsworth, A., Tracton, M., Lorenc, A., Uppala, S., & Kållberg, P. (1985). The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. *Quarterly Journal of the Royal Meteorological Society*, *111*(467), 67–101. https://doi.org/10.1002/qj.49711146703

Barker, T. (1991). The relationship between spread and forecast error in extended-range forecasts. *Journal of Climate*, *4*(7), 733–742. https://doi.org/10.1175/1520-0442(1991)004<0733:TRBSAF>2.0.CO;2

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

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Birner, T., & Albers, J. R. (2017). Sudden stratospheric warmings and anomalous upward wave activity flux. *SOLA*, *13A*(Special_Edition), 8–12. https://doi.org/10.2151/sola.13A-002

Buizza, R., & Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, *141*(693), 3366–3382. https://doi.org/10.1002/qj.2619

Butler, A., Charlton-Perez, A., Domeisen, D. I., Garfinkel, C., Gerber, E. P., Hitchcock, P., et al. (2019). Sub-seasonal predictability and the stratosphere. In *Sub-Seasonal to Seasonal Prediction*, (pp. 223–241). Elsevier.

Chang, P., Saravanan, R., DelSole, T., & Wang, F. (2004). Predictability of linear coupled systems. Part I: Theoretical analyses. *Journal of Climate*, *17*, 1474–1486. https://doi.org/10.1175/1520-0442(2004)017<1474:POLCSP>2.0.CO;2

Chen, M., & Kumar, A. (2015). Influence of ENSO SSTs on the spread of the probability density function for precipitation and land surface temperature. *Climate Dynamics*, *45*(3-4), 965–974. https://doi.org/10.1007/s00382-014-2336-9

Chen, W. Y., & van den Dool, H. M. (1999). Significant change of extratropical natural variability and potential predictability associated with the El Niño/Southern Oscillation. *Tellus*, *51*(5), 790–802. https://doi.org/10.3402/tellusa.v51i5.14493

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

Doblas-Reyes, F. J., Déqué, M., & Piedelievre, J. P. (2000). Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quarterly Journal of the Royal Meteorological Society*, *126*(567), 2069–2088. https://doi.org/10.1256/smsqj.56704

Feng, X., DelSole, T., & Houser, P. (2013). Comparison of statistical estimates of potential seasonal predictability. *Journal of Geophysical Research: Atmospheres*, *118*, 6002–6016. https://doi.org/10.1002/jgrd.50498

Ferranti, L., Magnusson, L., Vitart, F., & Richardson, D. S. (2018). How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, *144*(715), 1788–1802. https://doi.org/10.1002/qj.3341

Grimit, E. P., & Mass, C. F. (2007). Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results. *Monthly Weather Review*, *135*(1), 203–221. https://doi.org/10.1175/MWR3262.1

Hasselmann, K. (1976). Stochastic climate models part I. Theory. *Tellus*, *28*(6), 473–485.

Hitchcock, P., & Haynes, P. H. (2016). Stratospheric control of planetary waves. *Geophysical Research Letters*, *43*, 11–884. https://doi.org/10.1002/2016GL071372

Hopson, T. M. (2014). Assessing the ensemble spread-error relationship. *Monthly Weather Review*, *142*(3), 1125–1142. https://doi.org/10.1175/MWR-D-12-00111.1

Just, W., Kantz, H., Rödenbeck, C., & Helm, M. (2001). Stochastic modelling: replacing fast degrees of freedom by noise. *Journal of Physics A: Mathematical and General*, *34*(15), 3199. JRA-55 Atlas—Column heating. Retrieved from. http://ds.data.jma.go.jp/gmd/jra/atlas/en/D_HEATcol.html

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the Madden–Julian Oscillation: A review. *Journal of Climate*, *31*(23), 9425–9443. https://doi.org/10.1175/JCLI-D-18-0210.1

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*, 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Madden, R. A. (1976). Estimates of the natural variability of time-averaged sea-level pressure. *Monthly Weather Review*, *104*(7), 942–952. https://doi.org/10.1175/1520-0493(1976)104<0942:EOTNVO>2.0.CO;2

Martinez-Villalobos, C., Newman, M., Vimont, D. J., Penland, C., & Neelin, J. D. (2019). Observed El Niño-La Niña Asymmetry in a Linear Model. *Geophysical Research Letters*, *17*, 2399. https://doi.org/10.1029/2019GL082922

Murphy, A. H., & Epstein, E. S. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, *117*(3), 572–582. https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

National Academies of Sciences, Engineering, and Medicine (2016). *Next generation Earth System prediction: Strategies for subseasonal to seasonal forecasts*. Washington, DC: National Academies Press. https://doi.org/10.17226/21873

Newman, M., & Sardeshmukh, P. D. (2008). Tropical and stratospheric influences on extratropical short-term climate variability. *Journal of Climate*, *21*, 4326–4347. https://doi.org/10.1175/2008JCLI2118.1

Newman, M., & Sardeshmukh, P. D. (2017). Are we near the predictability limit of tropical sea surface temperatures? *Geophysical Research Letters*, *44*, 8520–8529. https://doi.org/10.1002/2017GL074088

Newman, M., Sardeshmukh, P. D., Winkler, C. R., & Whitaker, J. S. (2003). A study of subseasonal predictability. *Monthly Weather Review*, *131*(8), 1715–1732. https://doi.org/10.1175//2558.1

O'Neill, A., & Pope, V. (1988). Simulations of linear and nonlinear disturbances in the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, *114*(482), 1063–1110. https://doi.org/10.1002/qj.49711448210

Pegion, K., DelSole, T., Becker, E., & Cicerone, T. (2017). Assessing the fidelity of predictability estimates. In *Climate Dynamic*, (pp. 1–15). Berlin Heidelberg: Springer.

Pegion, K., & Sardeshmukh, P. D. (2011). Prospects for improving subseasonal predictions. *Monthly Weather Review*, *139*(11), 3648–3666. https://doi.org/10.1175/MWR-D-11-00004.1

Penland, C. (1996). A stochastic model of IndoPacific sea surface temperature anomalies. *Physica D*, *98*(2-4), 534–558. https://doi.org/10.1016/0167-2789(96)00124-8

Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, *8*(8), 1999–2024. https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2

Robertson, A., & Vitart, F. (Eds) (2018). *Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting*. Elsevier.

Sardeshmukh, P. D., Compo, G. P., & Penland, C. (2000). Changes of probability associated with El Niño. *Journal of Climate*, *13*(24), 4268–4286. https://doi.org/10.1175/1520-0442(2000)013<4268:COPAWE>2.0.CO;2

Sardeshmukh, P. D., Compo, G. P., & Penland, C. (2015). Need for caution in interpreting extreme weather statistics. *Journal of Climate*, *28*, 9166–9187. https://doi.org/10.1175/JCLI-D-15-0020.1

Sardeshmukh, P. D., & Penland, C. (2015). Understanding the distinctively skewed and heavy tailed character of atmospheric and oceanic probability distributions. *Chaos*, *25*(3), 036410. http://doi.org/10.1063/1.4914169

Shukla, J., & Gutzler, D. (1983). Interannual variability and predictability of 500 mb geopotential heights over the Northern Hemisphere. *Monthly Weather Review*, *111*(6), 1273–1279. https://doi.org/10.1175/1520-0493(1983)111<1273:IVAPOM>2.0.CO;2

Simmons, A. J., & Hollingsworth, A. (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *128*(580), 647–677. https://doi.org/10.1256/003590002321042135

Sura, P., Newman, M., Penland, C., & Sardeshmukh, P. (2005). Multiplicative noise and non-Gaussianity: A paradigm for atmospheric regimes? *Journal of the Atmospheric Sciences*, *62*(5), 1391–1409. https://doi.org/10.1175/JAS3408.1

Tang, Y., Lin, H., & Moore, A. M. (2008). Measuring the potential predictability of ensemble climate predictions. *Journal of Geophysical Research*, *113*, D04108. https://doi.org/10.1029/2007JD008804

Tennekes, H., Baede, A. P. M., & Opsteegh, J. D. (1986). Forecasting forecast skill. Workshop on Predictability in the Medium and Extended Range, 17-19 March 1986, Shinfield Park, Reading, 1986, pp. 277-302.

Tripathi, O. P., Baldwin, M., Charlton-Perez, A., Charron, M., Eckermann, S. D., Gerber, E., et al. (2015). The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 987–1003. https://doi.org/10.1002/qj.2432

van Schaeybroeck, B., & Vannitsem, S. (2016). A probabilistic approach to forecast the uncertainty with ensemble spread. *Monthly Weather Review*, *144*(1), 451–468. https://doi.org/10.1175/MWR-D-14-00312.1

Vigaud, N., Robertson, A. W., & Tippett, M. K. (2018). Predictability of recurrent weather regimes over North America during winter from submonthly reforecasts. *Monthly Weather Review*, *146*(8), 2559–2577. https://doi.org/10.1175/MWR-D-18-0058.1

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, F., & Molteni, F. (2010). Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quarterly Journal of the Royal Meteorological Society*, *136*(649), 842–855. https://doi.org/10.1002/qj.623

Weber, N. J., & Mass, C. F. (2017). Evaluating CFSv2 subseasonal forecast skill with an emphasis on tropical convection. *Monthly Weather Review*, *145*(9), 3795–3815. https://doi.org/10.1175/MWR-D-17-0109.1

Whitaker, J. S., & Loughe, A. F. (1998). The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, *126*(12), 3292–3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2

Winkler, C. R., Newman, M., & Sardeshmukh, P. D. (2001). A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill. *Journal of Climate*, *14*(24), 4474–4494. https://doi.org/10.1175/1520-0442(2001)014<4474:ALMOWL>2.0.CO;2

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S. J., Chen, J. H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*, 1077–1091. https://doi.org/10.1175/JAS-D-18-0269.1

Zhang, T., Hoerling, M. P., Wolter, K., Eischeid, J., Cheng, L., Hoell, A., et al. (2018). Predictability and prediction of Southern California rains during strong El Niño events: A focus on the failed 2016 winter rains. *Journal of Climate*, *31*(2), 555–574. https://doi.org/10.1175/JCLI-D-17-0396.1

References From the Supporting Information

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations*, (Vol. 18). New York: Oxford University Press.

Buizza, R., & Palmer, T. N. (1998). Impact of ensemble size on ensemble prediction. *Monthly Weather Review*, *126*(9), 2503–2518. https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2

Déqué, M. (1997). Ensemble size for numerical seasonal forecasts. *Tellus*, *49*(1), 74–86. https://doi.org/10.1034/j.1600-0870.1997.00005.x

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Wahrscheinlichkeitstheorie verw, Gebiete*, *57*(4), 453–476. https://doi.org/10.1007/BF01025868