Geophysical Research Letters

RESEARCH LETTER

10.1029/2020GL087008

Key Points:

- Evolution of ENSO forecasts initialized in late spring is too strongly tied to the observed evolution that precedes their initialization
- This state-dependent model bias reduces the reliability of ENSO forecasts made in late-spring
- Coupled models failed to capture the correct direction of ENSO evolution in half of last eight springs (2011 2018) including 2014

Supporting Information:

- Supporting Information S1

Correspondence to:
M. K. Tippett,
mkt14@columbia.edu

Citation:
Tippett, M. K., L'Heureux, M. L., Becker, E. J., & Kumar, A. (2020). Excessive momentum and false alarms in late-spring ENSO forecasts. *Geophysical Research Letters*, *47*, e2020GL087008. https://doi.org/10.1029/2020GL087008

Received 7 JAN 2020
Accepted 24 MAR 2020
Accepted article online 28 MAR 2020
Corrected 1 AUG 2020

This article was corrected on 1 AUG 2020. See the end of the full text for details.

Excessive Momentum and False Alarms in Late-Spring ENSO Forecasts

Michael K. Tippett[1], Michelle L. L'Heureux[2], Emily J. Becker[3], and Arun Kumar[2]

[1]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA, [2]NOAA/NWS/NCEP/Climate Prediction Center, College Park, MD, USA, [3]Cooperative Institute for Marine & Atmospheric Studies (CIMAS), Rosenstiel School of Marine & Atmospheric Science, University of Miami, Coral Gables, FL, USA

Abstract The unanticipated stalled El Niño-Southern Oscillation (ENSO) evolution of 2014 raises questions about the reliability of the coupled models that were used for forecast guidance. Here we have analyzed the skill and reliability of forecasts of the Niño 3.4 tendency (3-month change) in the North American multimodel ensemble (1982–2018). We found that forecasts initialized April–June (AMJ) have “excessive momentum” in the sense that the forecast Niño 3.4 tendency is more likely to be a continuation of the prior observed conditions than it should be. Models tend to predict warming when initialized after observed warming conditions and cooling when initialized after observed cooling conditions. Excessive momentum appears in AMJ forecast busts and false alarms including the 2014 one. In some models, excessive momentum appears to be related to model formulation rather than initialization. A concerning trend is that four of the nine years with AMJ forecast busts occurred in the last decade.

1. Introduction

El Niño-Southern Oscillation (ENSO) is the leading source of climate predictability on seasonal timescales, and knowing its phase and amplitude in advance is important for managing its impacts. Seasonality plays an important role in ENSO prediction. ENSO events, which peak in boreal winter, often start to develop in boreal spring when the tropical Pacific begins to transition from one ENSO state to another. Therefore, forecasts issued in late spring and early summer for the expected ENSO state of the coming winter are of particular interest and value. However, spring is also the time of year when the skill of ENSO forecasts is lowest, due to the so-called spring predictability barrier, which refers to the fact that ENSO forecasts passing through the months of April–June have relatively low skill. As a result, even capturing the correct direction of ENSO evolution (i.e., toward warmer or cooler conditions) is a forecasting challenge in late spring and early summer.

In any forecasting endeavor, analysis of “wrong” forecasts can be informative since forecast failures can reveal factors that limit skill and are barriers to improved prediction (Rodwell et al., 2013). Currently most operational ENSO forecasts are provided in probabilistic formats that give the probabilities of the tropical Pacific being in El Niño or La Niña conditions or that give the probabilities of ranges of values of the Niño 3.4 index (Barnston et al., 2015; L'Heureux et al., 2019, 2020). Strictly speaking, a probabilistic forecast cannot be considered to be wrong, even when the outcome that occurs was predicted to be unlikely because probabilistic forecasts explicitly contain uncertainty. However, over time, a forecast system can prove itself to be unreliable if the events to which it gives low probability occur too often (Weisheimer & Palmer, 2014). A long record of forecast performance over many cases is required to determine if probabilities from a forecast system are skillful and reliable.

Despite the impossibility of assessing reliability from a single ENSO event, the ENSO evolution of 2014 raises questions about the reliability of the coupled models that were used for forecast guidance. As late as June 2014, coupled model forecasts were confidently predicting strong El Niño conditions to come that winter (Carrington et al., 2014). Those warnings of a strong El Niño turned out to be a false alarm, and forecasts that failed to include the eventual outcome as a possibility were judged to be “busts.” The unexpected ENSO evolution in 2014–2015 has been examined from observational, modeling, and forecast perspectives (Ineson et al., 2018; Hu & Fedorov, 2016; Larson & Kirtman, 2015; Levine & McPhaden, 2016; Menkes et al., 2014; Puy et al., 2019; Wang & Hendon, 2017; Zhu et al., 2016). Despite the breadth of studies and the wide range of approaches, some practical questions remain unaddressed, namely, whether this forecast failure is indicative of systematic forecast system deficiencies and what the impact of such deficiencies might have on other forecasts and on overall skill and reliability. The 2014 false alarm could have been a rare occurrence or a warning of unreliability in the current generation of coupled model forecast systems.

The goal of this work was to address the broader implications of the 2014 forecast bust by examining the reliability of a large number of coupled model ENSO forecasts, specifically ENSO tendency forecasts, which we defined as forecasts of the change in the Niño 3.4 index over a period of a few months. We focused on tendency forecasts because they isolate a basic feature of ENSO evolution—whether conditions are warming or cooling. We analyzed a multimodel ensemble of initialized forecasts to examine the possible range of behavior across models and to include realistic impacts of both model formulation and initialization. We focused on false alarms and forecast busts in a long set of hindcasts and forecasts to identify common features and possible origins of forecast error, which differs from prior comprehensive assessments of ENSO forecast skill. By false alarm or forecast bust, we mean specifically cases in which coupled model forecasts confidently and incorrectly indicated the direction of future Niño 3.4 evolution.

The paper is organized as follows. Section 2 describes the forecast and observational data, and the methods used to assess reliability. Section 3 examines the reliability of the forecast ensemble, identifies a set of forecast busts, and finds that they share a common feature. We term this common feature *excessive momentum*, which is the propensity of the forecast values of Niño 3.4 index to continue in the direction of the prior observational tendency. Section 4 gives a summary and discussion.

2. Data and Methods

2.1. Data

Except where noted, we used the Niño 3.4 index computed from monthly averages of daily optimum interpolation sea surface temperature (Reynolds et al., 2007) for the period January 1982–September 2019. We repeated some of the calculations using the Niño 3.4 index computed from version 5 of the Extended Reconstructed Sea Surface Temperature dataset (Huang et al., 2017).

Niño 3.4 forecasts were taken from the North American multimodel ensemble (NMME) project database and have nominal monthly start dates of 1 January 1982–1 December 2018 (Kirtman et al., 2014). Reforecasts (hindcasts) and real-time forecasts from eight prediction systems were analyzed. The deterministic and probabilistic skill of the same models and of the multimodel ensemble has been analyzed previously over the shorter period 1982–2015 (Barnston et al., 2019; Tippett et al., 2019). Most of the forecasts extend 12 months past their start date. CFSv2 and NASA-GMAO forecasts are shorter and extend 10 and 9 months from their initializations, respectively. Initialization methods and ensemble sizes vary by prediction system (see, e.g., Kirtman et al., 2014 for details). Over most of the period, there are approximately 100 ensemble members at the shortest lead (lead-0), and 64 members at the last lead (lead-11). The NASA-GMAO model was replaced, and its final forecast was made in January 2018. The multimodel ensemble here is smaller during the remainder of 2018. Observed Niño 3.4 anomalies are computed with respect to the period 1982–2010. Forecast anomalies for most models are computed with respect their 1982–2010 forecast climatology. Two forecast climatologies (1982–1998 and 1999–2010) are used to compute the CFSv2 and CCSM4 forecast anomalies to account for a discontinuity in their hindcast initializations (Barnston & Tippett, 2013; Kumar et al., 2012; Xue et al., 2011).

For a Niño 3.4 forecast with nominal start date of 1 May, the 3-month forecast tendency is defined to be the August forecast value (lead-3) minus the May forecast value (lead-0). The corresponding prior 1-month observed tendency is defined to be the observed April value minus the March value. The same pattern is used for other start months and tendency lengths. We refer to the correlation between the prior tendency and the forecast tendency as *momentum* since it is a measure of the extent to which Niño 3.4 forecasts tend to continue in the direction of the prior observations. The observed ENSO momentum is defined in the same way except that observed values are used instead of forecast ones. In the above example, the observed ENSO momentum is the correlation between the observed March–April and observed May–August tendencies. To address the question of whether excessive momentum is due to model formulation or initialization, momentum can also be computed using forecast trajectories alone. For instance, the correlation between the forecast March–April tendency and the forecast May–August tendency for a forecast initialized in January is a measure of momentum that is more directly related to the dynamics of the forecast model and less to its initialization.

Niño 3.4 was also computed in a short set of ECMWF SEAS5 hindcasts and forecasts (25 members; January 1993–December 2018 starts Johnson et al., 2019) and UKMO GloSea5-GC2 System13 hindcasts (28 members; January 1995–December 2016 starts MacLachlan et al., 2015) which extend 6 months.

2.2. Methods

We used two methods to assess forecast reliability. Reliability, broadly speaking, indicates that issued forecast probabilities are in line with the outcomes that occur. Although a single probabilistic forecast cannot be judged to be either right or wrong, forecast reliability can be assessed from a set of forecasts and observations.

The first method for assessing reliability is based on what the weather and climate communities refer to as rank histograms or Talagrand diagrams (Hamill, 2001) and what is known in the wider forecasting community as the probability integral transform (Diebold et al., 1998; Gneiting et al., 2007). This method is suitable for forecasts of continuous scalar quantities such as the Niño 3.4 index and its tendency. The key idea is that if the verifying observation and ensemble members were drawn from the same probability distribution in each forecast, then the observations would be uniformly dispersed among the ensemble members and would not be preferentially found in the middle or extremes of the ensemble. In other words, the rank of the observation relative to the ensemble members would be uniformly distributed between 1 and $N + 1$, where N is the number of ensemble members. Equivalently, the probability integral transform formulation says that if O is the verifying observation and F is the cumulative distribution function (CDF) of the forecast ensemble, then $F(O)$ is uniformly distributed on the interval 0 to 1. Since $F(O)$ is defined as the fraction of ensemble members less than the verifying observation O, $100 \times F(O)$ is the percentile rank of the observation. $F(O)$ being uniformly distributed means that $\mathrm{Prob}\{F(O) \leq X\} = X$ for any X between 0 and 1. Therefore, for reliable forecasts, we expect that for any X the percentile rank of the observation will be less than $(100 \times X)$ in $(100 \times X)\%$ of the forecasts. For instance, taking $X = 0.3$, we expect that observations will be ranked in the bottom 30% of the ensemble in 30% of the forecasts.

We used the Kolmogorov-Smirnov test to check whether the distribution of the observation ranks is statistically significantly different from a uniform distribution. The Kolmogorov-Smirnov test is based on the CDF and thus avoids binning of ranks, which is a greater concern here than in weather applications in which the number of forecasts is larger. We also computed rank histograms (probability density functions rather than CDFs) which provide a familiar and easily interpreted visual depiction of reliability. The probability density function of a uniform distribution is flat.

The second method, reliability diagrams, is used to assess the reliability of probability forecasts of categorical variables such as the sign of the Niño 3.4 tendency. The forecast probability P of the occurrence of a condition C is reliable when $\mathrm{E}(C|P) = P$, where E means expectation (average). In other words, reliability means that the condition C occurs with frequency P when the forecast probability is P. The graph of the conditional average $\mathrm{E}(C|P)$ versus P is called a reliability diagram and is computed here by grouping forecast probabilities into five equally spaced bins and computing the frequency of occurrence for each bin. Here the condition C being forecast is that the ENSO tendency is positive. Ninety percent confidence intervals for the frequency of occurrence are based on the binomial distribution.

3. Results

ENSO forecasts that are initialized in spring have notoriously low skill compared to forecasts initialized at other times of the year (Barnston et al., 2019; Jin et al., 2008). However, these forecasts are of particular interest to forecasters since spring is the time of year when ENSO typically transitions from one state to another (Torrence & Webster, 1998). With this in mind, we considered forecasts of 3-month ENSO tendencies, which we defined as the lead-3 value of Niño 3.4 minus the lead-0 value. The correlation skill of 3-month tendency forecasts is highest for starts during January–March (Figure S1) when most of the tendency variability is due to the decay of mature ENSO events. Skill drops for subsequent start months, with a local minimum for June starts. October and November starts, when the 3-month tendency depends sensitively on details of ENSO peak behavior, have the lowest skill.

Reliability is a key aspect of forecast quality that measures the extent to which forecast uncertainty is accurately captured. We checked the reliability of NMME forecasts of 3-month ENSO tendencies by examining

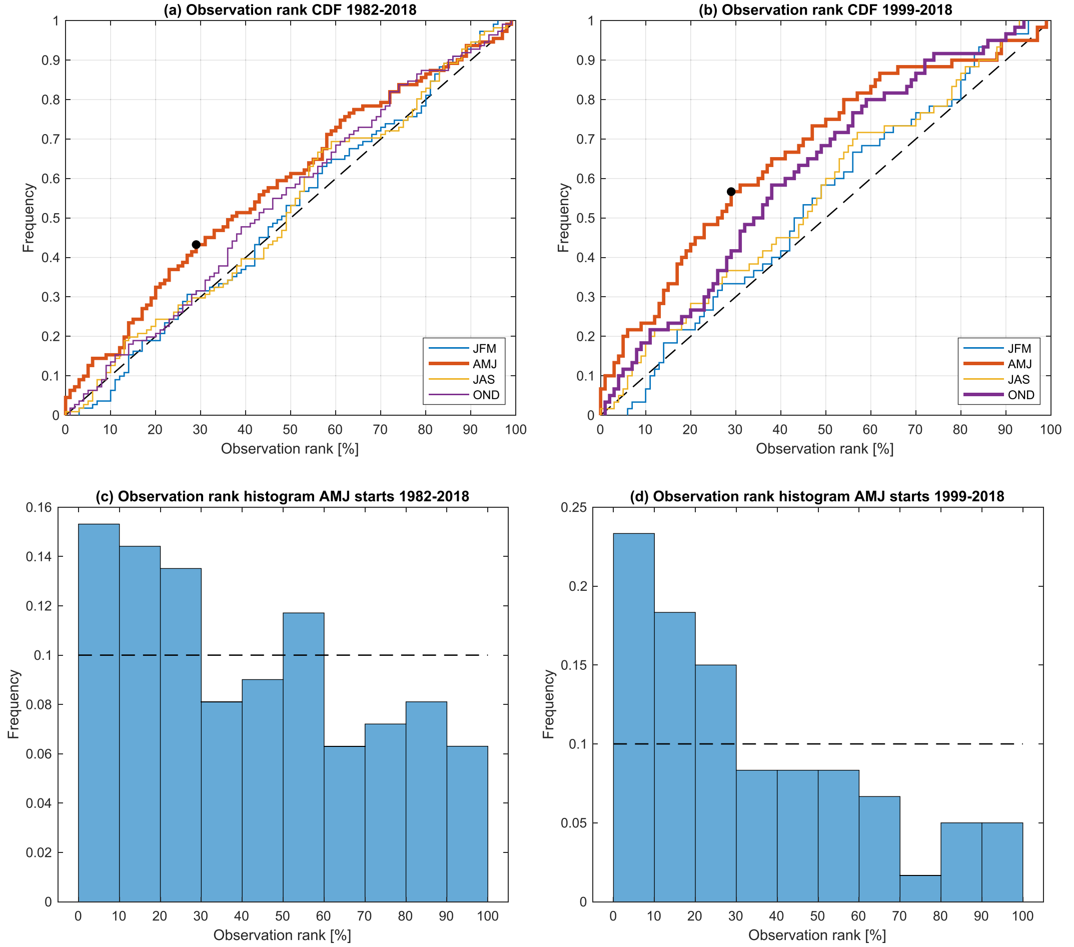

Figure 1. Cumulative distribution functions (CDFs) of observation ranks for 3-month tendency forecasts starting in January–March (JFM), April–June (AMJ), July–September (JAS), and October–December (OND) during (a) 1982–2018 and (b) 1999–2018. Thick lines indicate differences from the uniform distribution (black dashed line) that are statistically significant at the 5% level. Black dots in panels (a) and (b) indicate that 43% and 57% of the verifying observations were ranked in the bottom 30% of forecast ensembles 1982–2018 and 1999–2018, respectively. Observation rank histograms of 3-month tendency forecasts starting in AMJ during (c) 1982–2018 and (d) 1999–2018. The black dashed line indicates the uniform distribution.

the rank distributions of the verifying observations. We pooled the monthly starts into four 3-month groups and computed the rank distributions for each group. Forecasts of 3-month ENSO tendencies starting April–June (AMJ) are unreliable in the sense that the rank distribution differs statistically significantly from the uniform distribution (Figure 1a). The CDF of the ranks for AMJ starts lies above the one-to-one line, which indicates that too many ensemble members have stronger tendencies than were observed. The AMJ rank CDF shows that the verifying observed tendency was ranked in the bottom 30% of the ensemble in about 43% of the forecasts. The miscalibration of the AMJ forecast ensembles is also apparent in the rank histogram (Figure 1c) which shows that the verifying observations ranked much more often in the bottom 30% of the ensemble than in the top 30%. The unreliability of the forecast ensemble for AMJ starts is worse during the recent period 1999–2018 when the rank CDF (Figure 1b) shows that the verifying observed tendency was ranked in the bottom 30% of the ensemble in about 57% of the forecasts, and the rank histogram (Figure 1d) shows that the verifying observations were in the bottom 10% of the ensemble in 23% of the forecasts. NMME forecasts of 3-month tendencies initialized in October–December, when tendency forecasts have low skill (Figure S1), also show unreliability of same form—stronger tendencies forecast than actually occurred—during the period 1999–2018.

In principle, predicting the sign of the 3-month Niño 3.4 tendency should be a less demanding task than predicting the tendency itself (sign and amplitude). We examined NMME probability forecasts of the 3-month

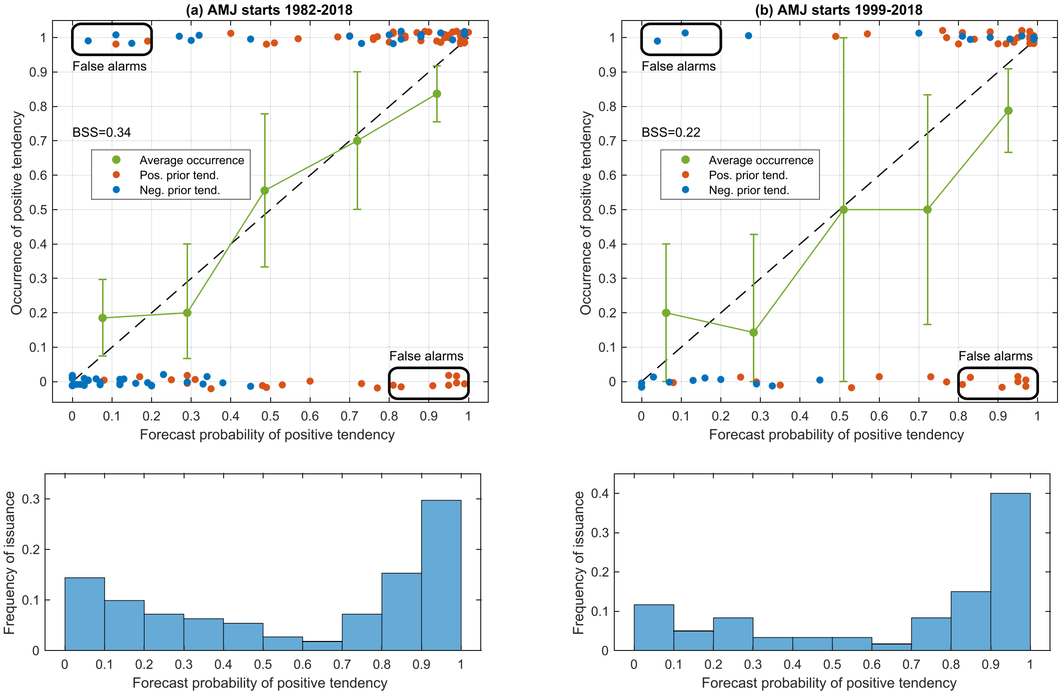

Figure 2. Reliability diagrams (top row) and forecast issuance frequency (bottom row) for April–June (AMJ) probability forecasts of the sign of 3-month Niño 3.4 tendency being positive (warming conditions) during (a) 1982–2018 and (b) 1999–2018. Each forecast issued and its verification is marked with a dot whose x-coordinate (abscissa) is the forecast probability and whose y-coordinate (ordinate) is the verifying observation, which is 0 for a negative tendency or 1 for a positive tendency, offset randomly in the vertical for legibility. The color of the dot indicates the sign of the observed tendency prior to the forecast initialization, red for positive and blue for negative. False alarm forecasts with probabilities greater than 80% for the tendency sign that did not occur are boxed. BSS = Brier skill score.

tendency being positive, that is, probability forecasts of warming conditions. The forecast probabilities were computed as the fraction of ensemble members with positive tendencies. Reliability diagrams for AMJ forecasts of the sign of the 3-month Niño 3.4 tendency show modest indications of overconfidence in the cases when the forecasts are most confident (Figure 2a). For forecast probabilities less than or equal to 20% (leftmost bin), forecast probabilities average 8%, while the occurrence frequency is 19%. For forecast probabilities greater or equal to 80% (rightmost bin), forecast probabilities average 92%, while the occurrence average is 84%. This overconfidence is slightly worse during the period 1999–2018 with forecast probabilities in the lowest bin averaging 6.2% and occurrence being 20%, and with forecast probabilities in the highest bin averaging 93% and occurrence being 79% (Figure 2b). Despite this overconfidence, the forecasts of the sign of 3-month Niño 3.4 tendency are skillful with Brier skill score (reference forecast is the base rate of the forecast period) over the full period of 0.34 and over the period 1999–2018 of 0.22. The forecast probabilities show no large unconditional biases in the sense that the frequency of positive tendencies is 56%, and the average forecast probability is 58%. There is a clear asymmetry in the distribution of forecast probabilities, with forecast probabilities of 90% or greater being the most frequent bin. There are two reasons for this asymmetry. First, there are slightly more cases (57%) during this period in which the ensemble mean tendency is positive. Second, the ensemble spread and the ensemble tendency spread are smaller for positive ensemble mean tendencies (std = 0.41°C and std = 0.39°C, respectively) than for negative ensemble mean tendencies (std = 0.55°C and std = 0.51°C, respectively). In fact, the correlation of the ensemble mean tendency with the ensemble std and ensemble tendency std is −0.64 and −0.69, respectively. Forecasts of warming conditions have less spread. There is also a negative correlation, though weaker (−0.4), between ensemble mean and ensemble mean spread.

Also shown on the reliability diagram are all the forecast probabilities for positive tendencies and their outcome (dots in Figures 2a and 2b). Dots in the lower right and upper left corners correspond to confident forecasts that failed to verify. In particular, we identified 13 forecasts starting in AMJ (nine different years) that gave probabilities greater than 80% for the wrong tendency sign (boxed off-diagonal dots in Figure 2a).

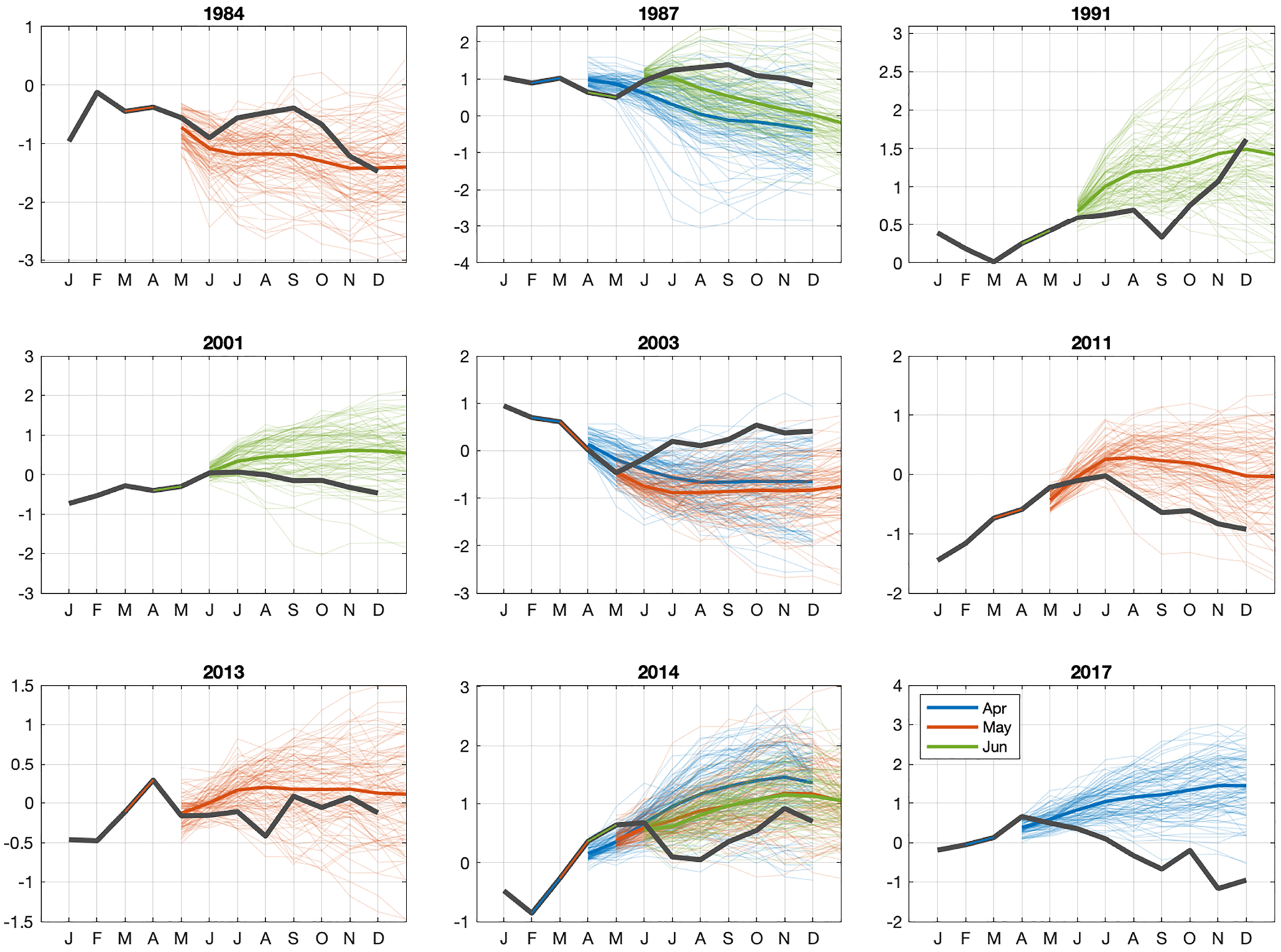

Figure 3. False alarm years (1984, 1987, 1991, 2001, 2003, 2011, 2013, 2014, and 2017) in which the forecast probability of the wrong sign of the 3-month tendency exceeded 80% for April–June starts. The black curves are observed monthly values of the Niño 3.4 index with 1-month prior tendencies highlighted in the same color as the corresponding forecast. The colored curves are forecast values with heavy lines for the North American multimodel ensemble mean and light lines for North American multimodel ensemble members. Note the differing vertical scales.

We call such forecasts "false alarms" because of the strength of their probability and their failure to verify. In one-third of the nine years (1984, 1987, and 2003), the ensemble undershot the observations, and in two-thirds of the years (1991, 2001, 2011, 2013, 2014, and 2017), the ensemble overshot the observations (Figure 3). The fact that the majority of the false alarms were in the positive direction suggests that the reduced ensemble spread for positive ensemble mean tendencies may not be justified. Year 2014 is a notable year in which all three initializations during AMJ pointed toward warming conditions which failed to occur.

The forecast and observed values of Niño 3.4 in the nine false alarm years show a strikingly consistent pattern: the forecast ensembles continued their Niño 3.4 evolution in the same direction as the observation prior to initialization, and the observations did not (Figure 3). With the exception of May 1984 and April 1987 initializations, and an argument can be made that the observations were tending downward overall, the erroneous forecasts continued in the same direction as the prior 1-month tendency. This consistency is also visible in the top row of Figure 2a where confident and incorrect warming forecasts (the forecasts in the lower right corner of the reliability diagram) are ones that follow warming conditions (red dots), and all but two of the confident and incorrect cooling forecasts (the forecasts in the upper left corner of the reliability diagram) are ones that follow cooling conditions (blue dots). Moreover, this conclusion is relatively insensitive to the choice of prior tendency length. In 1984, 1991, 2013, and 2014, observations at longer leads

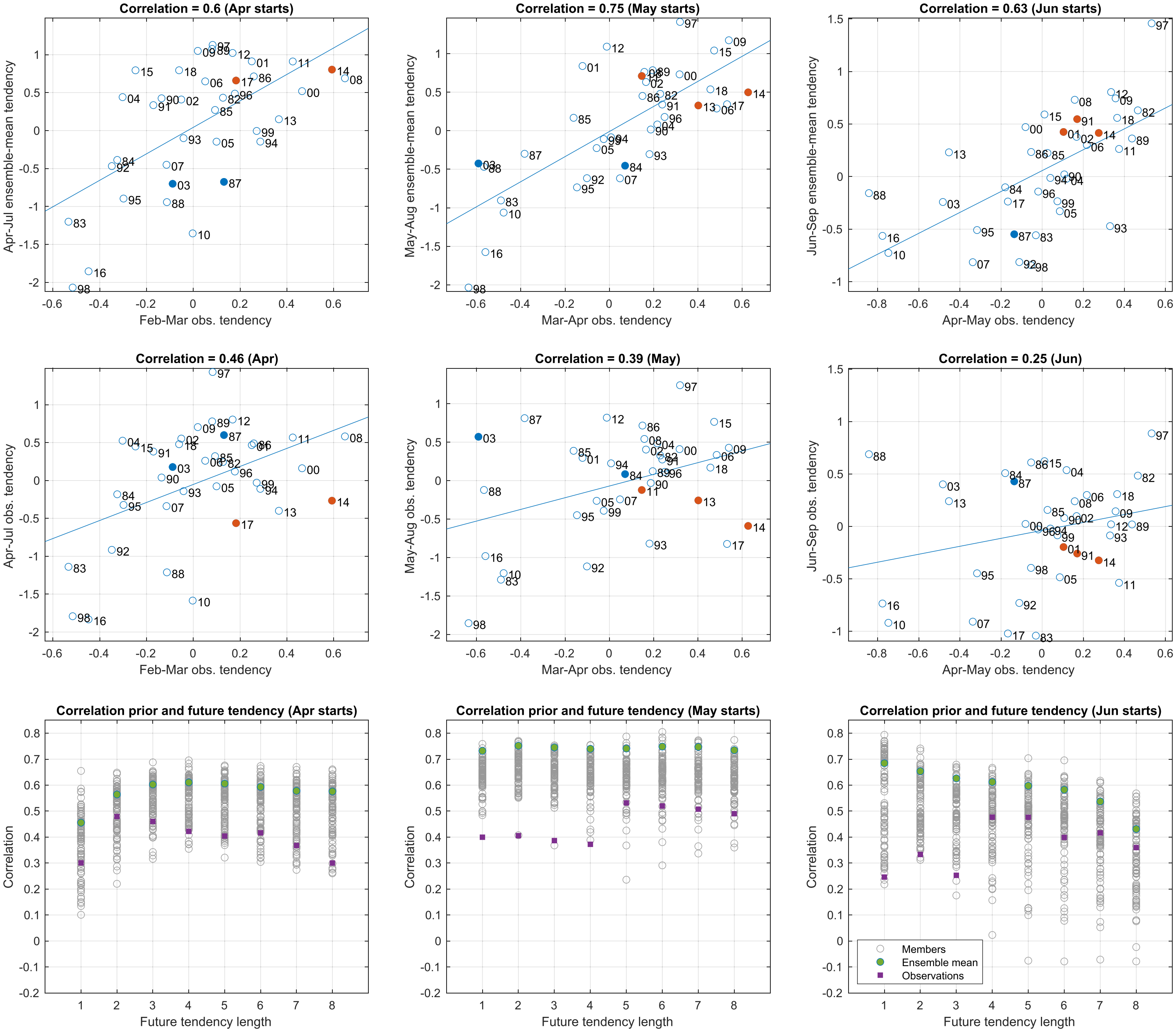

Figure 4. Scatterplots of 1-month prior Niño 3.4 tendencies and North American multimodel ensemble mean forecast 3-month Niño 3.4 tendencies for April, May, and June start (first row). Scatter plots of 1-month prior Niño 3.4 tendencies and observed future 3-month Niño 3.4 tendencies (second row). Correlation of prior Niño 3.4 tendencies with future Niño 3.4 tendencies as a function of future tendency length for each North American multimodel ensemble member (open circles), ensemble mean (green filled circles), and observations (purple squares) (third row).

eventually went in the direction of the early spring tendencies and better matched the forecasts. We looked at the correlation between the 1-month prior and 3-month ensemble mean forecast tendencies in all years for AMJ starts and found that the strong relation between prior tendency and forecast tendency is present more generally and is not limited to false alarms (first row of Figure 4). We call this quantity, the correlation between prior tendency and forecast tendency, *momentum*. The observed momentum in Niño 3.4 evolution (i.e., the correlation between 1-month prior and 3-month subsequent tendencies) is notably weaker (second row of Figure 4). In addition to the observed momentum being weaker than the ensemble mean momentum, as might be expected if observations are thought of as being analogous to an ensemble member, it is also weaker than the momentum of individual ensemble members. The difference between observed and ensemble member momentum is especially striking for May starts in which the observed momentum is below that of almost all the ensemble members and for the first three leads of June starts. This excess momentum extends to longer leads (last row of Figure 4) that go into winter, although the observed momentum increases as well, consistent with the behavior previously noted above for 1984, 1991, 2013, and 2014; the forecast December minus May difference is the 7-month tendency of a May start. Spurious momentum, that is, cases when the observed 3-month momentum is poorly covered by the 3-month momentum of the ensemble members, is absent during other times of the year (Figure S2).

Excess momentum is a state-dependent forecast bias that remains despite removing the forecast climatology (Hermanson et al., 2018). When positive tendencies precede a forecast initialization, forecast ensembles will tend to overshoot observations and tendency errors (observation tendency minus forecast tendency) are more likely to be negative because the forecast tendency is larger than the observed one (e.g., 1991, 2001 2011, 2013, 2014, and 2017 in Figure 3). The correlation between the tendency error in May starts and the 1-month prior tendency (March–April) is −0.56, −0.60, and −0.58 for 2-, 3-, and 4-month tendencies, respectively. The correlation between prior tendency and forecast errors (observations minus forecast) is −0.61, −0.59, and −0.58 for lead-2, -3, and -4 forecasts, respectively.

To address the question of whether excess momentum is related to errors in initialization or model formulation, we examined the correlation of forecast March–April tendencies with forecast May–August tendencies—no observations are used. We considered March or earlier initializations so that both tendencies are part of the same forecast trajectory, and the impact of initialization shock is reduced. There is no clear dependence of momentum on start month for the majority of models, which is evidence that excessive momentum in these models reflects model deficiencies rather than initialization deficiencies (Figure S3). On the other hand, CFSv2, ECMWF, GloSea5, and to some extent CM2p1, have more realistic momentum for earlier starts, which suggests a role for initialization deficiencies or model biases that vary by lead time.

4. Summary and Discussion

We have identified previously undetected state-dependent errors in coupled model ENSO forecasts. Focusing on the NMME, we found that forecasts initialized in AMJ have “excessive momentum” in the sense that the forecast Niño 3.4 tendency is more likely to be a continuation of the prior observed conditions than is warranted. In May starts, when the excessive momentum problem is specially severe, the correlation between the forecast May–August Niño 3.4 tendency and the prior observed March–April Niño 3.4 tendency is stronger in nearly every ensemble member than in observations. As a consequence, forecasts are overconfident and reliability is degraded. Rank histograms indicate that the ensemble tendency forecasts too frequently overshoot the verifying tendency observations. Forecasts of the sign of the Niño 3.4 tendency—whether Niño 3.4 will increase or decrease—show overconfidence. Strongly confident forecasts for positive tendencies are more frequent than for negative ones, due in part to reduced ensemble spread when the ensemble mean tendency shows warming. In some models, excessive spring momentum appears to be related to model formulation rather than initialization since forecast trajectories initialized in fall and winter also have excessive spring momentum. On the other hand, the behavior of other models (CFSv2, ECMWF, and GloSea5) is more realistic in earlier initializations, which points to a potential role for initialization or perhaps lead-time dependent biases. Excessive correlation between prior observed tendencies and forecast tendencies is absent in forecasts that target other times of the year.

Signs of excessive momentum are present in nearly all AMJ forecast busts (cases in which the sign of the 3-month forecast tendency is wrong in 80% or more of the ensemble members). In two-thirds of the years with AMJ forecast busts, the forecasts overshot the observations. Year 2014 is such a case, and this particular example may provide some clues about excessive momentum in general. In 2014, Niño 3.4 trajectories were trending upward strongly in late winter and early spring. NMME forecasts in AMJ followed suit and tended upward, but observed values tended downward over the summer. High-frequency wind variability is one explanation for the failed El Niño of 2014 (Hu & Fedorov, 2016; Menkes et al., 2014). Westerly wind events (WWE) from January to April 2014 gave rise to positive sea surface temperature anomalies in the central and eastern equatorial Pacific but were followed by a dearth of WWEs in May and June and by easterly wind events in June and July (Ineson et al., 2018; Levine & McPhaden, 2016). While the impact of easterly wind events seems to be model dependent, WWE activity in model simulations can play an important role in ENSO development (Puy et al., 2019) and are themselves modulated by the large-scale equatorial Pacific sea surface temperature in a way that effectively increases the ocean-atmosphere coupling (Eisenman et al., 2005). Excessive momentum could be related to lack of diversity in the high-frequency wind variability in forecast models.

Positive March–April tendencies also occurred in 2015, and 2015–2016 went on to become one of the strongest El Niño events on record (L'Heureux et al., 2017). Ineson et al. (2018) compared ECMWF S4 and GloSea5 forecasts initialized in May of 2014 and 2015 and found that the stronger WWE activity in 2015 compared to 2014 was a key factor in the differing ENSO evolution in the two years and that this difference in wind burst activity was predictable. However, while both models captured the difference in WWE activity between 2014 and 2015, ECMWF S4 had systematically higher WWE activity and overshot the observed ENSO evolution in both years. Also the wide range of forecast ENSO states possible across ensemble members with the similar levels of WWE activity demonstrates that factors other than WWE activity are important for controlling El Niño development in forecasts (Ineson et al., 2018). The possibility of insufficient spread due to neglecting noise-driven processes (Larson & Kirtman, 2015) and differences in mean state have also been proposed as being relevant for 2014 forecasts (Wang & Hendon, 2017).

In addition to identifying a robust state-dependent model error, we have also pointed out a new ENSO precursor signal, the spring Niño 3.4 tendency, which is stronger in models than observations. Although we have not included warm water volume (WWV) in our analysis (subsurface temperatures are unavailable in the IRI NMME database), WWV is a well-established ENSO precursor that is related to the recharged and discharged ENSO states (McPhaden, 2003). Larson and Pegion (2019) found that for a subset of NMME models, initialized forecasts were more tightly tied to the springtime recharged, neutral, and discharged states than were observations, leading to an underestimate of forecast uncertainty. This forecast uncertainty may be related to the problem of excessive momentum found here since the correlation of February WWV (west; 5N–5S, 120E–155W) with the February–April Niño 3.4 tendency is 0.89.

From a pragmatic standpoint, excessive momentum in late-spring coupled model ENSO forecasts is simply another model bias and should be amenable to correction by statistical methods that adjust the forecast probabilities (van den Dool et al., 2016), or the ensemble mean and spread (Bellprat et al., 2019) based on historical model performance. For the problem of excessive momentum, we would expect that weakening of forecast probabilities or damping of the ensemble mean would result in more reliable forecast guidance. Additionally, because excessive momentum results in state-dependent errors, there may be the possibility for novel correction methods that take into account the ENSO evolution prior to the forecast. Forecasters have learned from experience that model guidance is unreliable at this time of year, and in 2017, the CPC/IRI ENSO team never issued an El Niño Watch despite the elevated El Niño probabilities from the NMME (L'Heureux, 2018). On the other hand, sampling must be accounted for carefully in statistical calibration schemes since the sample size is relatively small in seasonal applications (Tippett et al., 2014; 2005), and the differing performance of hindcasts (1982–2010) and real-time forecasts is also a concern.

We have limited our attention here to ENSO as characterized by the Niño 3.4 index, which is an incomplete description of the diversity of ENSO behavior (Takahashi et al., 2011). A question for future studies is whether excessive momentum is preferentially present in forecasts of one flavor of ENSO compared to another and whether this results in differing levels of reliability and predictability (Lee et al., 2018).

Finally, we also note that while the seasonally varying autocorrelation of Niño 3.4, including the spring persistence barrier, is a well known and widely studied aspect of ENSO evolution (Levine & McPhaden, 2015; Liu et al., 2018), the lag correlation between Niño 3.4 tendencies is a new feature and could provide insight for statistical modeling as well as a new metric for assessing model fidelity (Tian et al., 2019). Also, it is notable that four of the nine years with AMJ forecast busts occurred in the last decade with real-time forecasts as opposed to hindcasts and that reliability is worse during the recent period 1999–2018, which raises the specter of changing ENSO dynamics (Dommenget & Vijayeta, 2019).

Acknowledgments
The authors thank two anonymous reviewers for their helpful comments and suggestions. M.K.T. acknowledges a Severo Ochoa Mobility grant that supported a visit to the Barcelona Supercomputing Center. Daily OISST is available at https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.OISST/.version2/.AVHRR/.sst, and ERSST data are available at https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.ERSST/.version5/.sst. NMME forecasts are available at https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME. ECMWF and UKMO data are available at https://iridl.ldeo.columbia.edu/SOURCES/.EU/.Copernicus/.CDS/.C3S/.ECMWF/.SEAS5/ and https://iridl.ldeo.columbia.edu/SOURCES/.EU/.Copernicus/.CDS/.C3S/.UKMO/.GloSea5-GC2/, respectively. Monthly values of warm water volume (west) are available at https://www.pmel.noaa.gov/tao/wwv/data/wwv_west.dat.

References

Barnston, A. G., & Tippett, M. K. (2013). Predictions of Nino 3.4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dynamics*, *41*, 1–19. https://doi.org/10.1007/s00382-013-1845-2

Barnston, A. G., Tippett, M. K., Ranganathan, M., & L'Heureux, M. L. (2019). Deterministic skill of ENSO predictions from the North American multimodel ensemble. *Climate Dynamics*, *53*, 7215–7234. https://doi.org/10.1007/s00382-017-3603-3

Barnston, A. G., Tippett, M. K., van den Dool, H. M., & Unger, D. A. (2015). Toward an improved multi-model ENSO prediction. *Journal of Applied Meteorology and Climatology*, *54*, 1579–1595. https://doi.org/10.1175/JAMC-D-14-0188.1

Bellprat, O., Guemas, V., Doblas-Reyes, F., & Donat, M. G. (2019). Towards reliable extreme weather and climate event attribution. *Nature Communications*, *10*, 1732. https://doi.org/10.1038/s41467-019-09729-2

Carrington, D., Godenberg, S., & Redfearn, G. (2014). *How El Niño will change the world's weather in 2014*. London, UK: The Guardian. https://www.theguardian.com/environment/2014/jun/11/-sp-el-nino-weather-2014

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

Dommenget, D., & Vijayeta, A. (2019). Simulated future changes in ENSO dynamics in the framework of the linear recharge oscillator model. *Climate Dynamics*, *53*, 4233–4248. https://doi.org/10.1007/s00382-019-04780-7

Eisenman, I., Yu, L., & Tziperman, E. (2005). Westerly wind bursts: ENSO's tail rather than the dog? *Journal of Climate*, *18*, 5224–5238. https://doi.org/10.1175/JCLI3588.1

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *69*, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*, 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hermanson, L., Ren, H. L., Vellinga, M., Dunstone, N. D., Hyder, P., Ineson, S., et al. (2018). Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Climate Dynamics*, *51*, 1411–1426. https://doi.org/10.1007/s00382-017-3962-9

Hu, S., & Fedorov, A. V. (2016). Exceptionally strong easterly wind burst stalling El Niño of 2014. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 2005. https://doi.org/10.1073/pnas.1514182113

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, *30*, 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Ineson, S., Balmaseda, M. A., Davey, M. K., Decremer, D., Dunstone, N. J., Gordon, M., et al. (2018). Predicting El Niño in 2014 and 2015. *Scientific Reports*, *8*, 10733. https://doi.org/10.1038/s41598-018-29130-1

Jin, E., Kinter, J., Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B., et al. (2008). Current status of ENSO prediction skill in coupled ocean-atmosphere models. *Climate Dynamics*, *31*, 647–664.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., et al. (2019). SEAS5: The new ECMWF seasonal forecast system. *Geoscientific Model Development*, *12*, 1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Kirtman, B., Min, D., Infanti, J. M., Kinter, J. L. III., Paolino, D. A., Zhang, Q., et al. (2014). The North American multi-model ensemble (NMME): Phase-1 seasonal to interannual prediction, Phase-2 toward developing intra-seasonal prediction. *Bulletin of the American Meteorological Society*, *95*, 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Kumar, A., Chen, M., Zhang, L., Wang, W., Xue, Y., Wen, C., et al. (2012). An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Monthly Weather Review*, *140*, 3003–3016. https://doi.org/10.1175/MWR-D-11-00335.1

L'Heureux, M. L. (2018). Overview of the 2017-18 .La Niña and El Niño Watch in mid-2018. In *43rd NOAA Annual Climate Diagnostics and Prediction Workshop*. Santa Barbara, CA: NOAA Office of Science and Technology Integration. https://www.nws.noaa.gov/ost/climate/STIP/43CDPW/43cdpw-MLHeureux.pdf.

L'Heureux, M. L., Levine, A., Newman, M., Ganter, C., Luo, J.-J., Tippett, M. K., & Stockdale, T. (2020). ENSO prediction. In M. J. McPhaden, A. Santoso, W. Cai (Eds.), *El Niño-Southern Oscillation (ENSO) in a changing climate*. Washington, DC: American Geophysical Union.

L'Heureux, M. L., Takahashi, K., Watkins, A. B., Barnston, A. G., Becker, E. J., Di Liberto, T. E., et al. (2017). Observing and predicting the 2015/16 El Niño. *Bulletin of the American Meteorological Society*, *98*, 1363–1382. https://doi.org/10.1175/BAMS-D-16-0009.1

L'Heureux, M. L., Tippett, M. K., Takahashi, K., Barnston, A., Becker, E. J., Bell, G. D., et al. (2019). Strength outlooks for the El Niño-Southern Oscillation. *Weather and Forecasting*, *34*, 165–175. https://doi.org/10.1175/WAF-D-18-0126.1

Larson, S. M., & Kirtman, B. P. (2015). An alternate approach to ensemble ENSO forecast spread: Application to the 2014 forecast. *Geophysical Research Letters*, *42*, 9411–9415. https://doi.org/10.1002/2015GL066173

Larson, S. M., & Pegion, K. (2019). Do asymmetries in ENSO predictability arise from different recharged states? *Climate Dynamics*, *54*(12), 1507–1522. https://doi.org/10.1007/s00382-019-05069-5

Lee, R. W.-K., Tam, C.-Y., Sohn, S.-J., & Ahn, J.-B. (2018). Predictability of two types of El Niño and their climate impacts in boreal spring to summer in coupled models. *Climate Dynamics*, *51*, 4555–4571. https://doi.org/10.1007/s00382-017-4039-5

Levine, A. F. Z., & McPhaden, M. J. (2015). The annual cycle in ENSO growth rate as a cause of the spring predictability barrier. *Geophysical Research Letters*, *42*, 5034–5041. https://doi.org/10.1002/2015GL064309

Levine, A. F. Z., & McPhaden, M. J. (2016). How the July 2014 easterly wind burst gave the 2015–2016 El Niño a head start. *Geophysical Research Letters*, *43*, 6503–6510. https://doi.org/10.1002/2016GL069204

Liu, Z., Jin, Y., & Rong, X. (2018). A theory for the seasonal predictability barrier: Threshold, timing, and intensity. *Journal of Climate*, *32*, 423–443. https://doi.org/10.1175/JCLI-D-18-0383.1

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., et al. (2015). Global seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*, 1072–1084. https://doi.org/10.1002/qj.2396

McPhaden, M. J. (2003). Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophysical Research Letters*, *30*(9), 1480. https://doi.org/10.1029/2003GL016872

Menkes, C. E., Lengaigne, M., Vialard, J., Puy, M., Marchesiello, P., Cravatte, S., & Cambon, G. (2014). About the role of westerly wind events in the possible development of an El Niño in 2014. *Geophysical Research Letters*, *41*, 6476–6483. https://doi.org/10.1002/2014GL061186

Puy, M., Vialard, J., Lengaigne, M., Guilyardi, E., DiNezio, P. N., Voldoire, A., et al. (2019). Influence of westerly wind events stochasticity on El Niño amplitude: The case of 2014 vs. 2015. *Climate Dynamics*, *52*, 7435–7454. https://doi.org/10.1007/s00382-017-3938-9

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, *20*, 5473–5496. https://doi.org/10.1175/2007JCLI1824.1

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., et al. (2013). Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, *94*, 1393–1405. https://doi.org/10.1175/BAMS-D-12-00099.1

Takahashi, K., Montecinos, A., Goubanova, K., & Dewitte, B. (2011). ENSO regimes: Reinterpreting the canonical and Modoki El Niño. *Geophysical Research Letters*, *38*, L10704. https://doi.org/10.1029/2011GL047364

Tian, B., Ren, H.-L., Jin, F.-F., & Stuecker, M. F. (2019). Diagnosing the representation and causes of the ENSO persistence barrier in CMIP5 simulations. *Climate Dynamics*, *53*, 2147–2160. https://doi.org/10.1007/s00382-019-04810-4

Tippett, M. K., Barnston, A. G., DeWitt, D. G., & Zhang, R.-H. (2005). Statistical correction of tropical Pacific sea surface temperature forecasts. *Journal of Climate*, *18*, 5141–5162.

Tippett, M. K., DelSole, T., & Barnston, A. G. (2014). Reliability of regression-corrected climate forecasts. *Journal of Climate*, *27*, 3393–3404. https://doi.org/10.1175/JCLI-D-13-00565.1

Tippett, M. K., Ranganathan, M., L'Heureux, M. L., Barnston, A. G., & DelSole, T. (2019). Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Climate Dynamics*, *53*, 7497–7518. https://doi.org/10.1007/s00382-017-3721-y

Torrence, C., & Webster, P. J. (1998). The annual cycle of persistence in the El Niño/Southern Oscillation. *Quarterly Journal of the Royal Meteorological Society*, *124*(550), 1985–2004.

van den Dool, H., Becker, E., Chen, L.-C., & Zhang, Q. (2016). The probability anomaly correlation and calibration of probabilistic forecasts. *Weather and Forecasting*, *32*, 199–206. https://doi.org/10.1175/WAF-D-16-0115.1

Wang, G., & Hendon, H. H. (2017). Why 2015 was a strong El Niño and 2014 was not. *Geophysical Research Letters*, *44*, 8567–8575. https://doi.org/10.1002/2017GL074244

Weisheimer, A., & Palmer, T. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, *11*, 20131162.

Xue, Y., Huang, B., Hu, Z.-Z., Kumar, A., Wen, C., Behringer, D., & Nadiga, S. (2011). An assessment of oceanic variability in the NCEP climate forecast system reanalysis. *Climate Dynamics*, *37*, 2511–2539. https://doi.org/10.1007/s00382-010-0954-4

Zhu, J., Kumar, A., Huang, B., Balmaseda, M. A., Hu, Z.-Z., Marx, L., & Kinter, J. L. III. (2016). The role of off-equatorial surface temperature anomalies in the 2014 El Niño prediction. *Scientific Reports*, *6*, 19677. https://doi.org/10.1038/srep19677

Erratum

The authors accidentally overwrote the lower left panels of Figure 2 during revision of the manuscript. Figure 2 has since been corrected, and this version may be considered the version of record.