Evaluating Three Decades of NMME Hindcasts to Assess Model Performance in Predicting ENSO Onset

Authors:

Sungjoon Park[1], Muhammad Azhar Ehsan[2,3,4]

1) Department of Computer Science, Columbia College at Columbia University, New York, USA.
2) Center for Climate Systems Research, Columbia Climate School, Columbia University, New York, NY 10025, USA.
3) NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA.
4) International Research Institute for Climate and Society, Columbia Climate School, Columbia University, New York, NY 10968, USA.

Submitted to

Scientific Reports

Corresponding author's address: Muhammad Azhar Ehsan, Center for Climate Systems Research, Columbia Climate School, Columbia University, New York, USA.
Email: mae2171@columbia.edu

Abstract

El-Niño Southern Oscillation (ENSO) phase transitions—shifts between neutral, La Niña, and El Niño states—reshape global climate patterns. Accurate predictions of these transitions are critical for agriculture, water management, and disaster preparedness. This study evaluates North American Multi-Model Ensemble (NMME) operational models to assess their performance in predicting Niño3.4 region sea surface temperature anomalies during ENSO onset seasons. A 30-year analysis (1991-2020) identified 18 ENSO episodes—nine El Niño and nine La Niña—exceeding the Niño 3.4 region's ±0.5°C threshold for ≥5 consecutive seasons. ENSO onsets predominantly occurred in non-winter seasons, with El Niño typically initiating in boreal spring-summer and La Niña emerging in boreal summer-fall. NMME models generally captured observed mean and variance well, but some models (e.g., NASA-GEOSS2S, COLA-CESM1) diverged significantly at longer lead times. Skill assessments reveal a seasonal pattern: correlation analysis and squared error skill scores both indicate lower prediction accuracy for late boreal spring to early summer targets, but notably improved performance for boreal fall and winter targets. However, some ENSO events are inherently difficult to predict, as seen with the poorly forecasted 2017 La Niña and 1994 and 1997 El Niño onsets. This underscores the event-specific and model-dependent nature of ENSO onset prediction.

Keywords

ENSO, Onset, El Niño, La Niña, NMME

1. Introduction

The El Niño/Southern Oscillation (ENSO) is a fundamental climate pattern that has been shown to cause interannual variability in precipitation and temperature in the Pacific Ocean and beyond through atmospheric teleconnection (Ropelewski and Halpert 1986, Wyrtki 1973). As a major driver of global climate variability, ENSO affects weather patterns worldwide. The ENSO phase transitions—shifts between neutral, La Niña, and El Niño states—reshape global climate patterns, affecting crop yields across major agricultural belts in the Americas, Asia, Africa, Australia, and China (Anderson et al. 2018, Wang et al. 2020, Shuai et al. 2013, Hansen et al. 1998). Understanding and predicting the onset of ENSO phase transitions is crucial for climate forecasting and planning across various sectors, and this research focuses specifically on this critical aspect.

ENSO transitions involve complex ocean-atmosphere interactions, including recharge-discharge processes and equatorial heat content changes. While observational insights and theoretical frameworks have advanced our understanding of ENSO transitions, representing transition processes accurately in coupled ocean-atmosphere models continues to be a significant challenge due to the underlying complexity of ocean-atmosphere interactions and the randomness of atmospheric disturbances which constitute the development of ENSO (Fedorov et al. 2003). Furthermore, model-specific variabilities in regard to seasonal dependencies make predicting these transitions with climate models a constant challenge.

Since 2011, the North American Multi-Model Ensemble (NMME) project has increased the predictability of ENSO by combining member model projections into probabilistic predictions of ENSO phase (Becker et al. 2022). When comparing the prediction skill of NMME1 (2011-2012) to NMME4 (2019-), Becker et al. reported a measured improvement (9%) in lead-1 sea surface temperature (SST) correlation (Becker et al. 2020). Predictions generated from the array of dynamical models constituting the NMME are widely used throughout meteorological agencies and sets the standard for ENSO prediction. .

Comparisons between dynamical models and statistical models in their ability to predict ENSO phase have continued throughout the NMME project. Barnston et al. assessed model performance using real-time ENSO model predictions during 2002-2011 and found a statistically significant edge in dynamical models over statistical models in the earlier half of the year (Barnston et al. 2012). This result is significant because uncertainty in predictions made across

spring to early summer seasons remains the largest barrier to extending accurate predictions of ENSO beyond the monthly time scale.

More recently, event-based comparisons of real-time ENSO forecast skill between dynamical and statistical models in the IRI ENSO Predictions Plume have continued to demonstrate the performance advantage of dynamical models over statistical models, especially for forecasts issued for spring seasons. Ehsan et al., using real-time forecasts of Niño 3.4 index across two decades (2001-2023) demonstrated that dynamical models, with their higher Squared Error Skill Score and lower mean absolute errors, provide more valuable predictions than statistical models do, especially during spring to summer months (Ehsan et al. 2024). Wang et al., meanwhile, reported that both dynamical and statistical models exhibited comparable forecasting skills except in the boreal spring season, where dynamical models generally outperformed statistical models in Spearman correlation between observed and forecasted SST anomalies (Wang et al. 2024).

While there have been a few studies on the real-time forecast performance of dynamical models and statistical models as groups, no up-to-date studies exist on the predictive skill of individual dynamical models at predicting the onset of all ENSO events in the last three decades. This gap in the literature is further highlighted by the fact that studies on the predictability of ENSO onset often focus exclusively on the forecast of El Niño onset and magnitude, rather than a comprehensive view of both cold and warm phase of ENSO (Ludescher et al. 2014, Lundescher et al. 2023). This study aims to address this gap in the literature by specifically focusing on the transition of ENSO, both in cold and warm phase onset. We separately evaluate and show the performance of each operational dynamical model within NMME to highlight the event and model-dependence of ENSO onset prediction. Furthermore, this study is unique in terms that it evaluates the skill of NMME models through hindcasts rather than real-time forecasts, which expands the study period of ENSO phase onset to all events that happened in the three decades between 1991 and 2020.

The core objective of this paper is to evaluate the performance of operational NMME models in predicting ENSO transitions, revealing each model's ability to (1) accurately hindcast SST anomalies (SSTAs) in the east-central equatorial Pacific and thereby (2) capture the onset of both cold and warm phases of ENSO at monthly lead times. We suggest that NMME models' ability to predict ENSO onset differ significantly based on the specific initialization conditions

preceding the onset and that models each have differing degrees of accuracy at predicting ENSO onset.

2. Data and Methodology

2.1 Observations and Model Data

The collection of seven dynamical models currently in operation, listed in Table 1, formally known as the North American Multi-Model Ensemble (NMME: Kirtman et al. 2014) project, represents the most recent efforts by North American institutions to model ocean-atmosphere interactions on a global scale. The period of NMME model predictions studied here is based on hindcasts from 1991 to 2020, covering a total of 30 years, which serves as a common period among all models. The maximum lead time of the models varies from 9 months for the NASA model to 10 months for the NCEP model, and up to 12 months for the CanESM, GFDL, NEMO, COLA, and CanSIPS models. The number of unique predictions made by ensemble models ranges from 4 for the NASA model hindcast dataset to 40 for the CanSIPS-IC4 model, which combines predictions from the CanESM and NEMO models. Further details are provided in Table 1. Area-averaged SST predictions and observations for the Niño 3.4 region in the central equatorial Pacific (5°S to 5°N, 170°W to 120°W) were analyzed. Monthly data were converted into seasonal data by calculating 3-month running averages for both models' hindcast predictions and observations. NMME model hindcast data were used to establish a 30-year climatology base period for each model, season, and start time. Predicted seasonal anomalies for each model were then calculated by subtracting each model's predictions from its respective 30-year climatology mean, based on start month and lead time, to obtain lead-dependent anomalies.

2.2 Identification of ENSO Events and Onset Timing

Using the observation data, we identified ENSO phases that occurred from 1991-2020. This study follows the methodology used by the Climate Prediction Center (CPC) to identify ENSO events: the 3-month running mean of SSTAs in the Niño 3.4 region must meet or exceed the threshold of +/- 0.5ºC for a minimum of 5 consecutive seasons. The first season where the Niño 3.4 anomaly meets or exceeds the threshold is defined as the onset, the beginning of each ENSO phase. Using this method, we identified 18 ENSO events – 9 warm (El Niño) and 9 cold (La Niña) – during the 1991 to 2020 period. The duration of each ENSO phase, along with the SST anomaly values at the time of onset, is detailed in Table 2. A notable tendency for cold ENSO phases, typically La Niña episodes, to persist for multiple years is observed, often manifesting as

double or triple dipping. This behavior is largely absent in warm El Niño episodes. An event is considered double or triple dipping (Ehsan et al. 2024) if the SST anomalies remain below -0.5 for two or three consecutive boreal winter seasons respectively (December to February: DJF). During the 1991-2020 period, we identified one double-dip event (2010-2012, marked with a single asterisk) and two triple-dip events (1998-2002 and 2020-2022, marked with double asterisks), as shown in Table 2. Notably, during these double and triple dips, the SST anomalies sometimes did not remain within the threshold for a couple of seasons, but we still considered them part of a single event and did not count them as new onsets.

2.3 Statistical Methods

This study utilizes a range of statistical techniques to compare model predictions with observational data, with a focus on seasonal mean and variance analyses across all twelve seasons (January-February-March: JFM to December-January-February: DJF), anomaly correlations, and skill evaluation of specific events using the Squared Error Skill Score (SESS). The following section provides a detailed explanation of the methodologies employed.

In initialized forecasts, lead time is a crucial concept. For example, in this study, a prediction is considered a 1-month lead time (Lead 1-month) if it is initialized in August for the target season of August-September-October (ASO). Following this convention, the GFDL, CanESM, NEMO, COLA, and CanSIPS models provide predictions with a maximum seasonal lead of up to 10 months (Lead 10-month). In contrast, the NASA model offers predictions with a seasonal lead of up to 7 months (Lead 7-month), and the NCEP model provides predictions with a seasonal lead of up to 8 months (Lead 8-month).

To analyze seasonal mean and member-based variance, we first grouped observational data by season, resulting in 30 data points per season. These data points were used to calculate both the mean and variance of the observed Niño 3.4 index. Similarly, model hindcast data were grouped by target season (JFM to DJF) and lead time (which varied by model), allowing us to calculate the model mean and variance for each season and lead time. For model variance, we computed the mean of the 30-year variance of individual ensemble members at each lead and season. This approach ensures that we capture the member-based variance for each model, which can be compared directly to the observational variance. To examine the model's performance in capturing deviations from the mean, we performed error and correlation analyses on the seasonal Niño 3.4

SSTAs. Observed anomalies ($O_T{}'$) at each target season were calculated by subtracting the seasonal observation mean (μ_T) from the observed value (O_T).

$$O_T{}' = O_T - \mu_T$$

Similarly, model hindcast anomalies ($F'_{S,L}$) for each model prediction were computed by subtracting the model's own seasonal forecast mean ($\mu_{M,T}$) from the forecasted value ($F_{M,S,L}$). This process ensures that predictions are assessed relative to each model's baseline seasonal climatology.

$$F'_{M,S,L} = F_{M,S,L} - \mu_{M,T}$$

Using these model hindcast anomalies and observed anomalies, the Anomaly Correlation Coefficient ($r_{M,S,L}$) was calculated at each start month and lead time across the study period, for each model:

$$r_{M,S,L} = \frac{Cov(F'_{M,S,L}, O'_T)}{\sigma_{F'_{M,S,L}} \sigma_{O'_T}}$$

Finally, event-specific hindcast anomaly errors ($FE_{M,S,L}{}'$) for each model were calculated by subtracting the observed anomaly from the model hindcast anomaly at each lead time (LT).

$$FE_{M,S,L}{}' = F'_{M,S,L} - O_T{}'$$

The Squared Error Skill Score (SESS) was applied to quantify the accuracy of the model forecasts with respect to the observed seasonal variance of each target season. The SESS for each prediction was calculated using the following formula, following the approach outlined by Barnston et al. (2012):

$$SESS = 1 - \frac{(FE_{M,S,L}{}')^2}{(\sigma_T)^2}$$

where $FE_{M,S,L}{}'$ represents the hindcast anomaly error given model (M), start (S), and lead (L) time, $O_T{}'$ represents the observed anomaly at the target season, and $(\sigma_T)^2$ represents the observed seasonal variance at the target season. An SESS of 1 indicates that the forecasted Niño 3.4 index values match the observed values perfectly, resulting in a zero-error term and signifying excellent forecast performance. An SESS of 0 means that the forecast error is equivalent to the observational variability, indicating no additional skill beyond what could be achieved by simply using the

observed variability as a forecast. A negative SESS value indicates that the forecast performance is worse than using the observational variability as a benchmark, reflecting poor forecasting.

3. Results

3.1 Model Performance during Three Decades (1991-2020): Climatology, Variance, Error, and Anomaly Correlation

This section begins with the results, examining the NMME models' performance during the last three decades using different evaluation metrics. Over the past three decades, the number of warm and cold ENSO events was balanced, with nine warm and nine cold events (see Table 2 for reference). However, the occurrence of one double-dip and two triple-dip La Niña events, where La Niña conditions persisted for two to three consecutive boreal winters, led to a greater overall duration of cold phase seasons (Table 2).

Figure 1 (left column) illustrates the seasonal mean values of the Niño3.4 index across all twelve seasons, ranging from JFM to DJF. The figure compares observational data with predictions from various NMME models at lead times of one month (Fig. 1a), four months (Fig. 1b), and seven months (Fig. 1c), respectively. The x-axis represents seasonal periods, while the y-axis shows the average Niño3.4 index in degrees Celsius. The observed Niño3.4 index peaks during late spring to early summer (AMJ) and reaches its lowest point in late fall to early winter (NDJ). This pattern aligns with previous studies on the annual cycle of background SSTs in the tropical Pacific, which have attributed the warm sea surface temperatures in April–June and cold SSTs in November–January to the asymmetry in the climate system's response to the annual cycle of solar radiation (Mitchell and Wallace, 1992; Jiang et al., 2025; Li and Philander, 1996). Replicating the annual cycle of the average Niño3.4 index is crucial for demonstrating a model's ability to accurately simulate the climate system of the equatorial Pacific. At LT-1, most models mimic the seasonal mean Niño3.4 index values as observed. However, at longer lead times (LT-4 and LT-7), the accuracy of model predictions generally decreases, with some models diverging further from observations than others. COLA-CESM1 consistently predicts a pronounced dip in the Niño3.4 index during the fall seasons (ASO to OND) at LT-4 and LT-7, overestimating the cooling trend observed in these months. The NCEP-CFSv2 model also shows a similar overestimation of post-summer cooling at LT-4 but recovers some accuracy at LT-7. The mean Niño3.4 index values of CanSIPS-IC4 sister models (CanESM5 and GEM5.2-NEMO) diverge significantly as lead time

increases, with GEM5.2-NEMO consistently predicting values over 2°C below observations at LT-7 due to substantial underestimation during summer months (JJA). NASA-GEOSS2S diverges from observed seasonal Niño3.4 values at longer lead times, underestimating index values during FMA to MJJ and overestimating them during SON to NDJ. With peak Niño3.4 index values predicted in AMJ, NASA's seasonal pattern at LT-7 is nearly the reverse of what is observed. GFDL-SPEAR closely replicates the seasonal mean variations in the Niño3.4 index and performs well even at longer lead times.

Figure 2 shows the 30-year member-based variance of NMME hindcast predictions and observations of the Niño3.4 index. Observed variance reaches its lowest point during the boreal spring to early summer seasons (MAM-MJJ) and rises to its peak in the late boreal fall to winter seasons (OND-DJF), displaying an opposite distribution compared to the mean Niño3.4 index shown in Figure 1. This observation aligns with findings by Ehsan et al. (2024) regarding the seasonal change in Niño3.4 SST variance (refer to the black curve in Fig. 4 of Ehsan et al., 2024). At LT-1, most models closely replicate the observed seasonal values in variance. However, as lead time increases, models vary in their ability to track the observed seasonal variance. NASA-GEOSS2S predicted nearly double the observed variance during boreal fall to spring (OND-MAM) seasons, with this overestimation worsening at longer lead times. COLA-CESM1 broadly overestimates variance at longer lead times (LT-4 and beyond), particularly during boreal summer to winter (JAS-DJF) seasons, though its error is generally smaller than that of NASA-GEOSS2S. NCEP-CFSv2 also overestimates variance, but this error is less pronounced compared to NASA-GEOSS2S and COLA-CESM1 and is primarily limited to boreal fall to winter (SON-NDJ) seasons. The Canadian models (CanESM5, CanSIPS-IC4, and GEM5.2-NEMO) capture the seasonal values in variance reasonably well, with GEM5.2-NEMO consistently predicting higher variance than CanESM5. GFDL-SPEAR performs most consistently across all lead times, closely tracking observed variance with an error of less than 0.5°C. Comparing Figures 1 and 2, NMME models generally track the mean and variance of SSTs well at shorter lead times. However, at longer lead times, some models show significant drawbacks, including overestimation or underestimation of the mean and variability of the Niño3.4 index, which can impact ENSO prediction during transition phases or the onset of new events.

Supplemental Figure 1 illustrates the 30-year member-based variance for each model from boreal spring to fall (AMJ-OND), the seasons during which ENSO onset occurred in the study

period (Table 2). Canadian models (CanESM5, CanSIPS-IC4, GEM5.2-NEMO) and GFDL-SPEAR show the lowest variance for predictions made during boreal spring and the highest variance for predictions made after spring. This pattern is most evident in CanESM5 (Fig. S1a) and GFDL-SPEAR (Fig. S1b), where predictions after spring exhibit higher variance compared to those made during spring. In contrast, COLA-CESM1 (Fig. S1d) and NASA-GEOSS2S (Fig. S1f) predict increasing variance toward longer lead times, with particularly high variance in long-lead predictions made during boreal spring (e.g., LT-7 and LT-8 for SON and OND). NCEP-CFSv2 (Fig. S1g) consistently predicts variance with low errors across lead times, performing best at LT-1. COLA-CESM1 and NASA-GEOSS2S overestimate variance during and after boreal spring by wide margins, unlike Canadian models and GFDL-SPEAR, which tend to underestimate variance for boreal fall target seasons at long lead times.

Figure 3 illustrates the anomaly correlation for all lead times, focusing on ENSO onset seasons across NMME models. Onset seasons, identified in Table 2 as AMJ to OND, represent periods when at least one ENSO event began during the base period (1991 to 2020). This analysis highlights these critical transition seasons to evaluate the skill of NMME models through anomaly correlation coefficient (ACC), which is essential for understanding and forecasting ENSO development during onset periods. The shaded cells represent model predictions with an anomaly correlation of 0.7 (or higher), indicating that the models can explain at least 50% of the seasonal variance, reflecting significant predictive skill in capturing the relationship between observed and predicted anomalies at each lead time and target season. The CanESM5 (Fig. 3a) model shows strong predictive skill for later target seasons, particularly OND, where ACC values exceed 0.9 even at longer lead times, though its performance is weaker for earlier seasons like AMJ, with ACC values below 0.5 at longer (beyond LT-4) lead times. GFDL-SPEAR (Fig. 3b) demonstrates moderate performance overall, with ACC values improving closer to the target season; OND correlations exceed 0.9 at shorter lead times. For the earlier seasons from AMJ to JAS, the model exhibits weaker predictive skill, with low correlation coefficients even at shorter leads, and these values drop significantly as the lead time increases (e.g., Almazroui et al., 2022). GEM5.2-NEMO (Fig. 3c) performs well overall during all onset seasons from AMJ to OND. Notably, it maintains ACC values above 0.8 for AMJ and JJA seasons up to four leads, demonstrating strong performance by this model during springtime—a period when models typically show lower skill (e.g., Barnston et al. 2015[36], Tippett et al. 2020[37], Levine et al. 2025[38]). Additionally, it continues

to perform well for later seasons like SON and OND (Fig. 3c), achieving high ACC values. The high ACC values seen here can be interpreted as a measure of potential skill rather than actual skill, as studies such as Murphy & Epstein (1988)[39] and Feddersen et al. (1999)[40] have shown that ACC ignores variance and bias, meaning that even if a model has high ACC, climatology miscalibration can lower prediction skill in practice. The COLA-CESM1 model (Fig. 3d) demonstrates high predictive skill (ACC > 0.7, shaded yellow to brown) for shorter lead times and late-year seasons like SON and OND, where ACC values exceed 0.9 at lead times 1–3. However, it exhibits lower ACC values for earlier seasons such as AMJ, MJJ and JJA, with ACC dropping below 0.7 at longer leads for these specific seasons. CanSIPS-IC4 (Fig. 3e) demonstrates robust performance during the AMJ to OND seasons, similar to the GEM5.2-NMEO model, although slightly less than the GEM model. This difference may be attributed to the fact that GEM5.2-NMEO is a sister model of CanSIPS-IC4, alongside the CanESM5 model. Nonetheless, the CanSIPS-IC4 model shows promising ACC values during the ENSO onset seasons. NASA-GEOSS2S (Fig. 3f) exhibits gradually higher ACC values from JAS to OND seasons, while it shows weaker performance in earlier seasons particularly from AMJ to JJA. Notably, there is a noticeable decrease in skill at LT-1 from 0.91 to 0.79 at LT-2 for the JJA season in this model. The NCEP-CFSv2 (Fig. 3g) model ACC demonstrates lower predictive skill during the AMJ to JJA seasons, even at shorter lead times, with values dropping to approximately 0.5 at LT-4. The model's performance improves gradually from JAS to OND seasons. In summary, the models generally exhibit stronger predictive skill for later ENSO onset seasons (SON and OND), while their performance varies significantly for earlier seasons like AMJ and JJA depending on the model used. This pattern aligns with earlier research, which has consistently shown that models typically display lower skill (e.g., Barnston et al. 2015[36], Tippett et al. 2020[37], Levine et al. 2025[38], Kumar et al. 2017[40]) during boreal spring and early summer but higher skill during late summer, fall, and winter (e.g. Zhang et al. 2025[41], Tippett & Becker 2024[42]). In the next section, we will delve into the predictive capabilities of these NMME models for individual ENSO events, examining their ability to accurately forecast specific events onsets.

3.2 Event-Specific Analysis of Niño 3.4 Index Prediction Errors and Skill Scores: Performance of NMME Models During ENSO Onset

Figure 4 illustrates the anomaly errors (Model - OBS) for Niño3.4 index predictions of cold phase onsets across multiple models as a function of lead time, increasing toward the upper

part of each panel. The values are organized by the season of onset, from earliest (MJJ) to latest (OND), to better reflect seasonal variations. Errors closer to zero indicate better alignment between predicted and observed Niño3.4 index values, while larger deviations highlight prediction failures. Notably, positive errors exceeding 1°C (highlighted in deep red) signify significant inaccuracies in capturing the onset of cold events. For instance, CanESM5 (Fig. 4a) consistently shows positive errors across nearly all cold events, indicating a tendency to underestimate cooling trends. The 2010 and 2017 cold events exhibit substantial errors at longer lead times, demonstrating poor predictive accuracy. GFDL-SPEAR (Fig. 4b) exhibits mixed performance, with both positive and negative errors. It also underestimates cooling during certain events, such as 2010 and 2017, with significant errors persisting across lead times, highlighting challenges in predicting these transitions. GEM5.2-NEMO (Fig. 4c) predominantly underestimates cooling trends, as reflected by positive anomaly errors. The model struggles with key events like 2010 and 2017, with errors exceeding 1°C at longer lead times. COLA-CESM1 (Fig. 4d) displays a mix of overestimation (negative errors) and underestimation (positive errors). While it overestimates cooling for the 1998 event, it significantly underestimates the intensity of the 2010 and 2017 cold events, similar to other models. Notably, errors for the 1998 cold event also remain high (Fig. 4d). CanSIPS-IC4 (Fig. 4e) shows consistently large positive errors across most events and lead times. Its performance during the 2010 and 2017 cold events is particularly poor, with errors exceeding 1°C across all lead times except shorter leads, indicating major prediction failures. NASA-GEOSS2S (Fig. 4f) and NCEP-CFSv2 (Fig. 4g) demonstrate relatively moderate performance compared to other models but still struggle with key events like 2010 and 2017. Notably, NASA-GEOSS2S (Fig. 4f) performs well for the 2017 event at shorter leads up to LT-3, showing better accuracy compared to other models, which exhibit large and increasing errors beyond lead 1. In summary, the failure to accurately predict critical cold events, such as those in 2010 and 2017, highlights significant limitations in models during these periods. This underscores that some specific cold event onsets are particularly challenging to predict.

On the flipside, Figure 5 shows the anomaly errors (Model - OBS) for Niño3.4 index predictions of warm phase onsets across multiple models as a function of lead time, increasing

toward the upper part of each panel. The 1991 MJJ onset columns are blank at Lead 6 and beyond because data collection begins with the 1991 January initialization. As with Figure 4, accurate predictions would present an anomaly error close to zero. In this case, a strong negative (dark blue) anomaly error is more inaccurate than a positive (red) anomaly error because the former would indicate a prediction failure (prediction of a cold event when in reality, a warm event occurred). Similar to the challenges in capturing the Niño 3.4 index during the onset of cold events, the NMME models exhibit notable difficulties in predicting warm event onsets. For the 1994 warm event, all models struggled significantly, showing large errors across nearly all lead times, indicating a consistent challenge in capturing this event's dynamics. The onset of the 1997 warm event also posed forecasting challenges, with models like GFDL-SPEAR and GEM5.2-NEMO displaying substantial errors, particularly at longer lead times, reflecting difficulties in accurately predicting this major El Niño event well in advance. In contrast, the 2014 warm event saw the COLA-CESM1 model stand out due to extremely large positive anomaly errors (up to +1.16°C) at shorter lead times, suggesting a significant overestimation of warming compared to observations. Other models generally showed smaller errors for this event, although NASA-GEOSS2S also displayed notable positive biases at some leads. Across all events, models such as CanSIPS-IC4 and CanESM5 consistently demonstrated negative errors, suggesting a bias toward underestimating warming, while NASA-GEOSS2S and NCEP-CFSv2 exhibited mixed performance with errors varying across events and lead times. Overall, while certain models performed better for specific events or lead times, none consistently outperformed others across all cases. The struggles in forecasting the 1994 event at all leads and the 1997 event at longer leads highlight limitations in capturing complex El Niño dynamics, while the overestimation of the 2014 event by COLA-CESM1 underscores variability in model biases.

Figures 6 and 7 show the Squared Error Skill Score of each model at all LT for cold and warm onsets, respectively. The SESS of each prediction considers the natural year-to-year variability of each season and evaluates the anomaly error on equal footing across seasons. In other words, if a model predicts, with the same anomaly error, two onsets at the same LT but situated at different target seasons, the prediction made for the target month with a higher variance will be given a higher SESS score.

The analysis of cold onset predictions in Figure 6 shows that the skill of onset prediction often depends on the model and the onset event. For instance, the 2017 SON onset is poorly

predicted by all models at long LT. In this case, it is suspected that the initialization conditions beyond LT-5 were indicative of an imminent warm event, leading to a prediction failure across the board.

The analysis of warm onset predictions in Figure 7 presents a similar narrative of onset prediction being largely dependent on the model and the onset event. While some columns show the expected behavior of SESS decreasing as LT increases (see CanSIPS 1997 AMJ and GFDL 2018 ASO), the skill scores of most other predictions appear to be dependent on the model's overall grasp of each onset event. For instance, CanSIPS predictions for the 2014 SON onset stays above 0.75 at all leads, and the model predicts the onset with the same skill score in Lead 10 as it does in Lead 1. In another example, COLA predictions for the 2002 MJJ onset shows an oscillation between moderate to strong prediction success across LT. These examples show that the accuracy of model predictions of the SSTAs of the Niño 3.4 region are not necessarily dependent on the LT.

Together, Figures 6 and 7 indicate a general increase in SESS for onsets that happen in fall target seasons over those in spring target seasons, though this trend is not without exceptions. The most notable exceptions are 2017 SON and 1994 ASO onsets, where most models reported a prediction failure at long LT despite their onset season being situated in the fall. It is not immediately clear how these prediction failures fit into the wider narrative of an increase in prediction skill for fall onsets – whether to attribute these challenges to the initialization conditions of the onsets or other sources of bias is a matter that will be further explored in the discussion.

4. Discussion

One large-scale trend observed across the onset-specific and seasonal analyses is the general increase in prediction skill outside spring target seasons. The anomaly correlation results show that models struggle to predict Niño 3.4 anomaly observations when predicting through the spring. The SESS results show models generally struggling with predicting MJJ season onsets, which would require the models to predict through the spring season. Previous studies have labeled this model behavior as the "spring predictability barrier," attributing the reduced predictability of ENSO through the spring to causes such as weak ENSO SSTAs (Xue et al. 1994), unstable air-sea interactions in the springtime eastern equatorial Pacific (Philander et al. 1984), and the decaying phase of many ENSO events (Torrence and Webster 2018).

While this study does not delve into details regarding ENSO dynamics through the spring, the 30-year mean and variance figures offer a glimpse into the behavior of observed and model

prediction SSTs over the base period. The observed mean SSTs peaking during the AMJ season. At the same time, the variance of the observed SSTs reaches its very bottom. This means two things for the Niño 3.4 region of the AMJ season: 1) SSTs are at the highest energy level of the year, conducive to unstable air-sea interactions as demonstrated in Philander et al. 1984, and 2) SSTs do not vary much year-to-year, giving weak ENSO SSTAs as illustrated in Xue et al. 1994.

In this context, models use initialization conditions for each LT to predict the SST anomaly at the time of onset. Recall the 2017 SON cold onset that models failed to predict at long LT. Looking at the 3-month running average SST anomaly values leading up to the SON season, we see a warming trend through the spring of 2017 (+0.06 in 2017 FMA to +0.32 in 2017 MJJ). However, after the MJJ season, we see a reversal into a cooling trend (+0.32 in 2017 MJJ to -0.65 at 2017 SON). Through the spring, the warming trend coming out of a La Niña reversed into a cooling trend into another La Niña event. The lower SESS at LT-4 and LT-5, where most models lose their prediction accuracy, correspond to initialization conditions set in June and May, respectively.

Models attempt to make accurate predictions based on what limited information they have during the spring. Models such as CanESM, GFDL, NEMO, and CanSIPS limit the variance of predictions made at this time, increasing the consistency of the predictions but missing the signal of ENSO onset on numerous occasions. On the flipside, NASA and COLA sacrifice consistency for high-variance forecasts during the spring, aggressively predicting the ENSO signal but recording large margins of error for mispredictions. On some occasions, the risk-taking pays off: the skillful prediction of the 1994 ASO warm onset by NASA and the high anomaly correlation of COLA and NASA are two testaments to the payoff. At the same time, large margin forecast failures such as 2010 MJJ, 2017 SON, and 1994 ASO onsets by the COLA model show the risks of pursuing such a strategy.

Potential explanations for model biases can be found in the literature, where studies attribute a cold bias in pre-1999 NCEP model hindcasts of Niño 3.4 SSTAs to the assimilation of new satellite observations starting from October 1998 (Xue et al. 2011) and a spurious warming trend in equatorial Pacific SST hindcasts (Shin and Huang 2017). The latter study is particularly interesting because it reports an exaggerated decade-to-decade warming in equatorial Pacific SSTs seen across NMME member models. If modern NMME models still carry this exaggerated decade-

to-decade warming trend, then it will provide insights into interpreting the hindcast analysis results shown in this study.

This study aimed to provide a holistic assessment of NMME member models' skill at predicting the onset of ENSO phases. To this end, this study focused on the models' month-to-month performance to predict SSTAs at the Niño 3.4 region based on a singular 30-year base period. With this in mind, additional analyses into the decade-to-decade warming trends using the hindcast data are recommended. Spatial maps of decade-to-decade warming trends in the NMME hindcasts of the equatorial Pacific is crucial to revealing the source of bias of current generation NMME models. Follow-up studies using a split climatological period for models with strong decade-to-decade warming bias are recommended to assess the performance of models in the vacuum of spurious warming trends.

Acknowledgements

The data analysis portion of this work has been undertaken as part of the Lamont-Doherty Earth Observatory (LDEO) summer internship program.

The NMME project and data dissemination is supported by NOAA, NSF, NASA and DOE.

Author Contributions

MA Ehsan commenced the study and provided directions throughout the study period.

Corresponding author

Correspondence to MA Ehsan (mae2171@columbia.edu)

Availability Statement

The NMME hindcast data used in this study are publicly accessible on the International Research Institute for Climate and Society (IRI) Data Library at the following link: http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME.

The Niño 3.4 Region SST observations used here are the NOAA Extended Reconstruction SSTs Version 5 (ERSSTv5) publicly accessible at https://www.cpc.ncep.noaa.gov/data/indices/.

References

1. Ropelewski, C. F. & Halpert, M. S. North American precipitation and temperature patterns associated with the El Niño/southern oscillation (ENSO). *Mon. Weather Rev.* 114, 2352–2362 (1986).
2. Wyrtki, K. (1973). Teleconnections in the Equatorial Pacific Ocean. *Science*, *180*(4081), 66–68. http://www.jstor.org/stable/1735305
3. Anderson, W., Seager, R., Baethgen, W. & Cane, M. Trans-pacific ENSO teleconnections pose a correlated risk to agriculture. *Agric. For. Meteorol.*262, 298–309 (2018).
4. Bin Wang, Puyu Feng, Cathy Waters, James Cleverly, De Li Liu, Qiang Yu, Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia,Agricultural and Forest Meteorology, Volume 291, 2020, 108043, ISSN 0168-1923, https://doi.org/10.1016/j.agrformet.2020.108043.
5. Shuai, J., Zhang, Z., Sun, D.-Z., Tao, F., & Shi, P. (2013). ENSO, climate variability and crop yields in China. *Climate Research, 58*(2), 133–148. https://doi.org/10.3354/cr01194
6. Hansen, J. W., Hodges, A. W., & Jones, J. W. (1998). ENSO Influences on Agriculture in the Southeastern United States. *Journal of Climate*, *11*(3), 404-411. https://doi.org/10.1175/1520-0442(1998)011<0404:EIOAIT>2.0.CO;2
7. Fedorov, A. V., et al. "HOW PREDICTABLE IS EL NIÑO?" Bulletin of the American Meteorological Society, vol. 84, no. 7, 2003, pp. 911–19. JSTOR, http://www.jstor.org/stable/26216859. Accessed 24 June 2024.
8. Becker, E. J., B. P. Kirtman, M. L'Heureux, Á. G. Muñoz, and K. Pegion, 2022: A Decade of the North American Multimodel Ensemble (NMME): Research, Application, and Future Directions. Bull. Amer. Meteor. Soc., 103, E973–E995, https://doi.org/10.1175/BAMS-D-20-0327.1.
9. Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. Geophysical Research Letters, 47, e2020GL087408. https://doi.org/10.1029/2020GL087408
10. Barnston, Anthony & Tippett, Michael & Ranganathan, Meghana & L'Heureux, Michelle. (2019). Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. Climate Dynamics. 53. 10.1007/s00382-017-3603-3.
11. Duan, W., & Wei, C. (2013). The 'spring predictability barrier' for ENSO predictions and

its possible mechanism: Results from a fully coupled model. *International Journal of Climatology*, **33**(5), 1280–1292. https://doi.org/10.1002/joc.3513

12. Pang, Y., Jin, Y., Zhao, Y., Chen, X., Li, X., Liu, T., & Hu, J. (2023). Sea surface salinity strongly weakens ENSO spring predictability barrier. *Geophysical Research Letters*, 50, e2023GL106673. https://doi.org/10.1029/2023GL106673
13. Shin, CS., Huang, B. A spurious warming trend in the NMME equatorial Pacific SST hindcasts. Clim Dyn 53, 7287–7303 (2019). https://doi.org/10.1007/s00382-017-3777-8
14. Diro, G. T., W. J. Merryfield, H. Lin, W.-S. Lee, R. Muncaster, V. V. Kharin, R. Parent, et al. "The Canadian Seasonal to Interannual Prediction System Version 3.0 (CanSIPSv3.0) Technical Note." Canadian Center for Meteorological and Environmental Prediction, 10 June 2024.
15. Song Yang, Zhenning Li, Jin-Yi Yu, Xiaoming Hu, Wenjie Dong, Shan He, El Niño–Southern Oscillation and its impact in the changing climate, *National Science Review*, Volume 5, Issue 6, November 2018, Pages 840–857, https://doi.org/10.1093/nsr/nwy046
16. Cai, Wenju & Wang, Guojian & Santoso, Agus & McPhaden, Michael & Wu, Lixin & Jin, Fei-Fei & Timmermann, Axel & Collins, M. & Vecchi, Gabriel & Lengaigne, Matthieu & England, Matthew & Dommenget, Dietmar & Takahashi, Ken & Guilyardi, Eric. (2015). Increased frequency of extreme La Niña events under greenhouse warming. Nature Climate Change. 5. 132-137. 10.1038/nclimate2492.
17. Tao Tang, Jing-Jia Luo, Ke Peng, Li Qi, Shaolei Tang, Over-projected Pacific warming and extreme El Niño frequency due to CMIP5 common biases, *National Science Review*, Volume 8, Issue 10, October 2021, nwab056, https://doi.org/10.1093/nsr/nwab056
18. Glantz, M.H., Ramirez, I.J. Reviewing the Oceanic Niño Index (ONI) to Enhance Societal Readiness for El Niño's Impacts. *Int J Disaster Risk Sci* **11**, 394–403 (2020). https://doi.org/10.1007/s13753-020-00275-w
19. Wang-Chun Lai, A., M. Herzog, and H. Graf, 2018: ENSO Forecasts near the Spring Predictability Barrier and Possible Reasons for the Recently Reduced Predictability. *J. Climate*, **31**, 815–838, https://doi.org/10.1175/JCLI-D-17-0180.1.
20. Xue, Y., M. A. Cane, S. E. Zebiak, and M. B. Blumenthal, 1994: On the prediction of ENSO: A study with a low-order Markov model. Tellus, 46A, 512–528, https://doi.org/10.3402/ tellusa.v46i4.15641.

21. Philander, S. G. H., T. Yamagata, and R. C. Pacanowski, 1984: Unstable air–sea interactions in the tropics. J. Atmos. Sci., 41, 604–613, https://doi.org/10.1175/1520-0469(1984)041,0604: UASIIT.2.0.CO;2.
22. Torrence, C., and P. J. Webster, 1998: The annual cycle of persistence in the El Niño/Southern Oscillation. Quart. J. Roy. Meteor. Soc., 124, 1985–2004, https://doi.org/10.1002/ qj.49712455010.
23. Xue Y, Huang B, Hu ZZ, Kumar A, Wen C, Behringer D, Nadiga S (2011) An assessment of oceanic variability in the NCEP climate forecast system reanalysis. Clim Dyn 37:2511–2539, DOI:10.1007/s00382-010-0954-4
24. Muhammad Azhar Ehsan, Michelle L'Heureux, Michael Tippett et al. Real-Time ENSO Forecast Skill Evaluated Over the Last Two Decades, with Focus on Onset of ENSO Events, 15 November 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3588191/v1]
25. Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S., & DeWitt, D. G. (2012). Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society, 93*(5), 631–651. https://doi.org/10.1175/BAMS-D-11-00111.1
26. Ehsan, M.A., L'Heureux, M.L., Tippett, M.K. *et al.* Real-time ENSO forecast skill evaluated over the last two decades, with focus on the onset of ENSO events. *npj Clim Atmos Sci* **7**, 301 (2024). https://doi.org/10.1038/s41612-024-00845-5
27. Wang, H., Asefa, T., & Duncan, J. (2024). Event-based evaluation of operational ENSO forecasting models in 2002–2020: Implications for seasonal water resources management. *Journal of Hydrology, 636*, 131295. https://doi.org/10.1016/j.jhydrol.2024.131295
28. Ludescher, J., Bunde, A. & Schellnhuber, H.J. Forecasting the El Niño type well before the spring predictability barrier. *npj Clim Atmos Sci* **6**, 196 (2023). https://doi.org/10.1038/s41612-023-00519-8
29. J. Ludescher, A. Gozolchiani, M.I. Bogachev, A. Bunde, S. Havlin, H.J. Schellnhuber, Very early warning of next El Niño, *Proc. Natl. Acad. Sci. U.S.A.* 111 (6) 2064-2066, https://doi.org/10.1073/pnas.1323058111 (2014).
30. Mitchell, T. P., and J. M. Wallace, 1992: The Annual Cycle in Equatorial Convection and Sea Surface Temperature. *J. Climate*, **5**, 1140–1156, https://doi.org/10.1175/1520-

0442(1992)005<1140:TACIEC>2.0.CO;2.

31. Jiang, S., Zhu, C., & Jiang, N. (2025). Impacts of the annual cycle of background SST in the tropical Pacific on the phase and amplitude of ENSO. *Atmospheric and Oceanic Science Letters, 18*(1), 100496. https://doi.org/10.1016/j.aosl.2024.100496
32. Xing, X., Fang, X., Pang, D. et al. Seasonal Variation of the Sea Surface Temperature Growth Rate of ENSO. Adv. Atmos. Sci. 41, 465–477 (2024). https://doi.org/10.1007/s00376-023-3005-x
33. Tebaldi, C., Dorheim, K., Wehner, M., & Leung, R. (2021). Extreme metrics from large ensembles: Investigating the effects of ensemble size on their estimates. Earth System Dynamics, 12(4), 1427–1501. https://doi.org/10.5194/esd-12-1427-2021
34. Almazroui, M., Ehsan, M.A., Tippett, M.K. et al. Skill of the Saudi-KAU CGCM in Forecasting ENSO and its Comparison with NMME and C3S Models. Earth Syst Environ 6, 327–341 (2022). https://doi.org/10.1007/s41748-022-00311-3
35. Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., III, Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., & Wood, E. F. (2014). The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585-601. https://doi.org/10.1175/BAMS-D-12-00050.1
36. Barnston, A. G., Tippett, M. K., Van den Dool, H. M., & Unger, D. A. (2015). Toward an improved multimodel ENSO prediction. *Journal of Climate, 28*(14), 1579–1594. https://doi.org/10.1175/JCLI-D-14-00376.1
37. 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
38. Levine, A. F. Z., L'Heureux, M. L., & Wen, C. (2025). *Understanding spring forecast El Niño false alarms in the North American Multi-Model Ensemble*. NPJ Climate and Atmospheric Science, 8(1), 94.
39. Murphy, A. H., & Epstein, E. S. (1989). Skill scores and correlation coefficients in model

verification. *Monthly Weather Review, 117*(3), 572–581. https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

40. Kumar, A., Z.-Z. Hu, B. Jha,and P. Peng, 2017: Estimating ENSO predictability based on multi-model hindcasts. *Climate Dyn.*, ***48***, 39–51, doi:10.1007/s00382-016-3060-4.
41. Zhang, W., Villarini, G., Slater, L., Vecchi, G. A., & Bradley, A. A. (2017). Improved ENSO forecasting using Bayesian updating and the North American Multimodel Ensemble (NMME). *Journal of Climate, 30*(22), 9007–9025. https://doi.org/10.1175/JCLI-D-17-0073.1
42. Tippett, M. K., & Becker, E. J. (2024). Trends, skill, and sources of skill in initialized climate forecasts of global mean temperature. *Geophysical Research Letters, 51*(16), e2024GL110703. https://doi.org/10.1029/2024GL110703

Tables and Figures

Table 1: List of 7 NMME models examined in this study

Model	Development organization	Hindcast period	No. of hindcast ensemble members	Max lead (months)
CanESM5*	Canadian Centre for Climate Modelling and Analysis (CCCma)	CanESM5	20*	12
GFDL-SPEAR	Geophysical Fluid Dynamics Laboratory (GFDL)	GFDL-SPEAR	15	12
GEM5.2-NEMO*	Canadian Centre for Climate Modelling and Analysis (CCCma)	NEMO	20*	12
COLA-CESM1	National Center for Atmospheric Research (NCAR)	COLA	10	12
CanSIPS-IC4	Canadian Centre for Climate Modelling and Analysis (CCCma)	CanSIPS	40	12
NASA-GEOSS2S	NASA Global Modeling and Assimilation Office (GMAO)	NASA	4	9
NCEP-CFSv2	NOAA National Centers for Environmental Prediction (NCEP)	NCEP	24	10

**CanESM5 and GEM5.2-NEMO are sister models for CanSIPS-IC4.*

Table 2 List of El Niño/cold (9 events) and La Niña/warm (9 events) that occurred during study period of 1991-2020

#	Year	Season of onset	Value of Niño 3.4 index at onset season (°C)	Duration from onset to demise (months)
Warm events				
1	1991	MJJ	0.51	13
2	1994	ASO	0.59	7
3	1997	AMJ	0.63	12
4	2002	MJJ	0.59	9
5	2004	JJA	0.46	8
6	2006	ASO	0.54	5
7	2009	JJA	0.45	9
8	2014	SON	0.49	19*
9	2018	ASO	0.49	10
Cold events				
1	1995	JAS	-0.56	8
2	1998	JJA	-0.89	33**
3	2005	OND	-0.52	5
4	2007	MJJ	-0.47	13
5	2008	OND	-0.56	5
6	2010	MJJ	-0.66	23*
7	2016	JAS	-0.55	5
8	2017	SON	-0.65	7
9	2020	JAS	-0.57	30**

**Indicates a double dip La Niña. ** indicates a triple dip La Niña.*

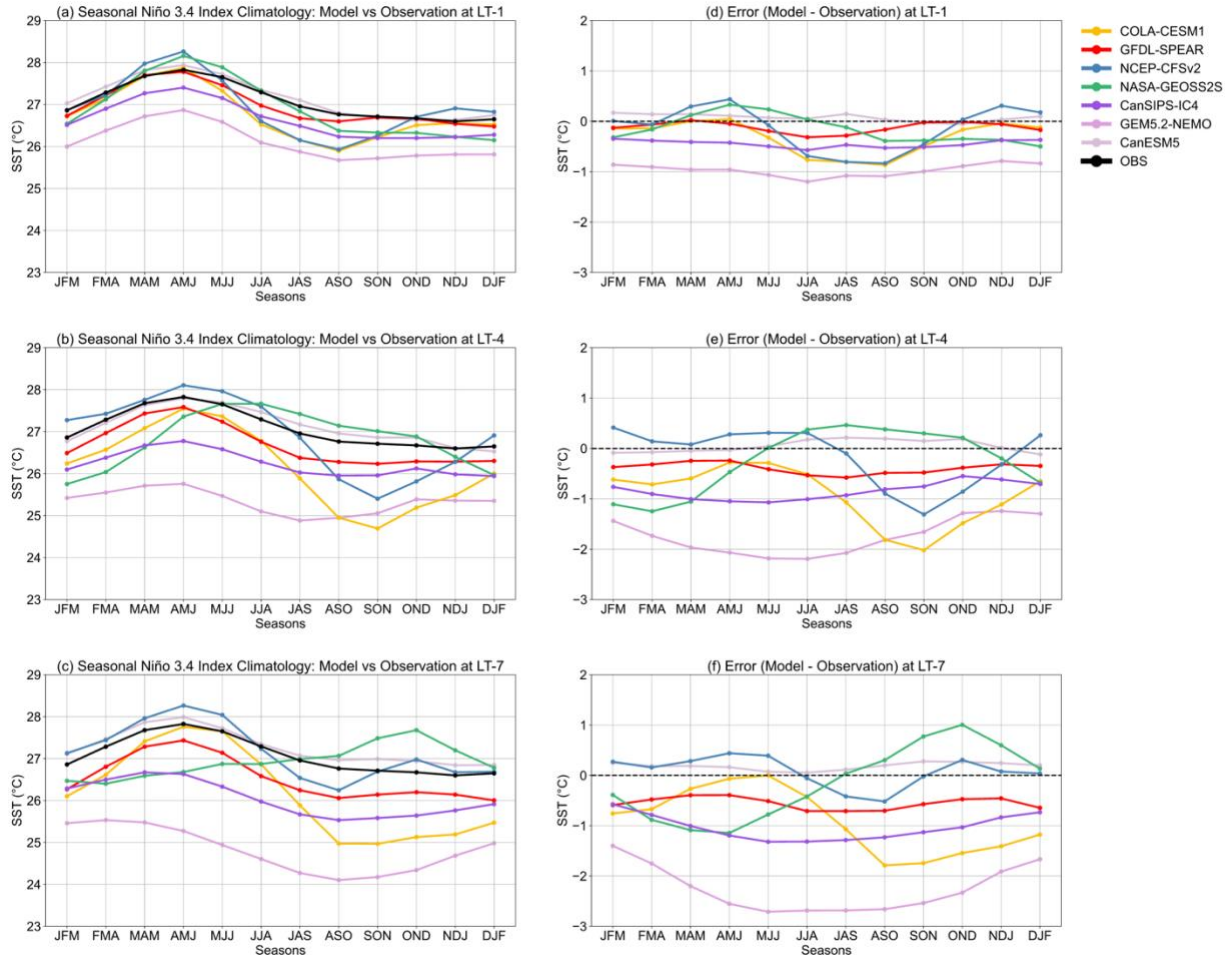

Figure 1: (**Left Column**) Seasonal mean Niño 3.4 index observation (black) and NMME hindcasts (each model represented by a specific color) for the 1991-2020 period at lead times of 1, 4, and 7 months (panels a, b, c). (**Right Column**) Seasonal mean Niño 3.4 index errors (model minus observation) for the same lead times (panels d, e, f), with each model represented by the same colors as in the left column. The observed data used for verification is from the ERSSTv5 dataset.

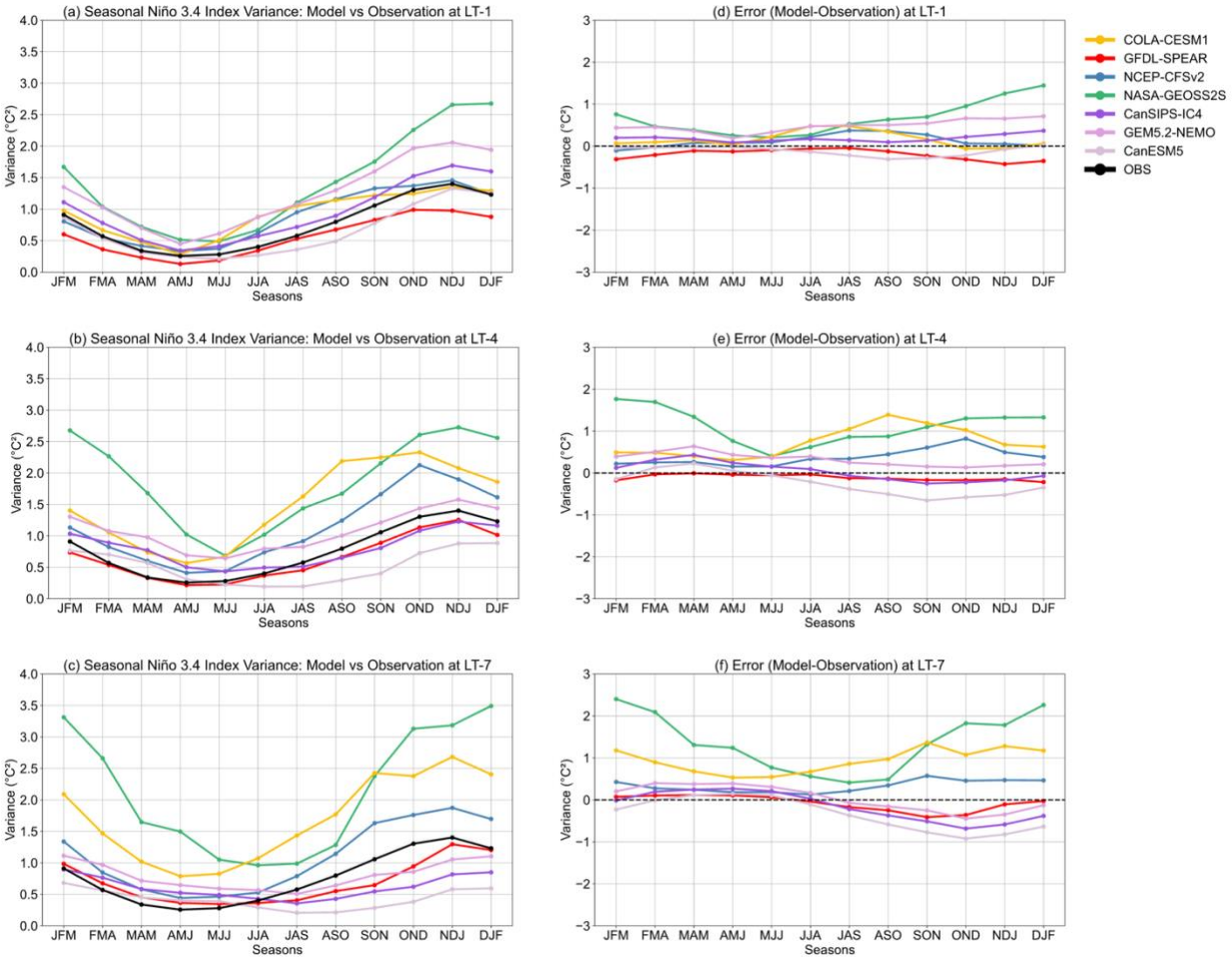

Figure 2: Similar to Figure 1, but variance is shown instead. The model variance is calculated based on the variance of individual ensemble members.

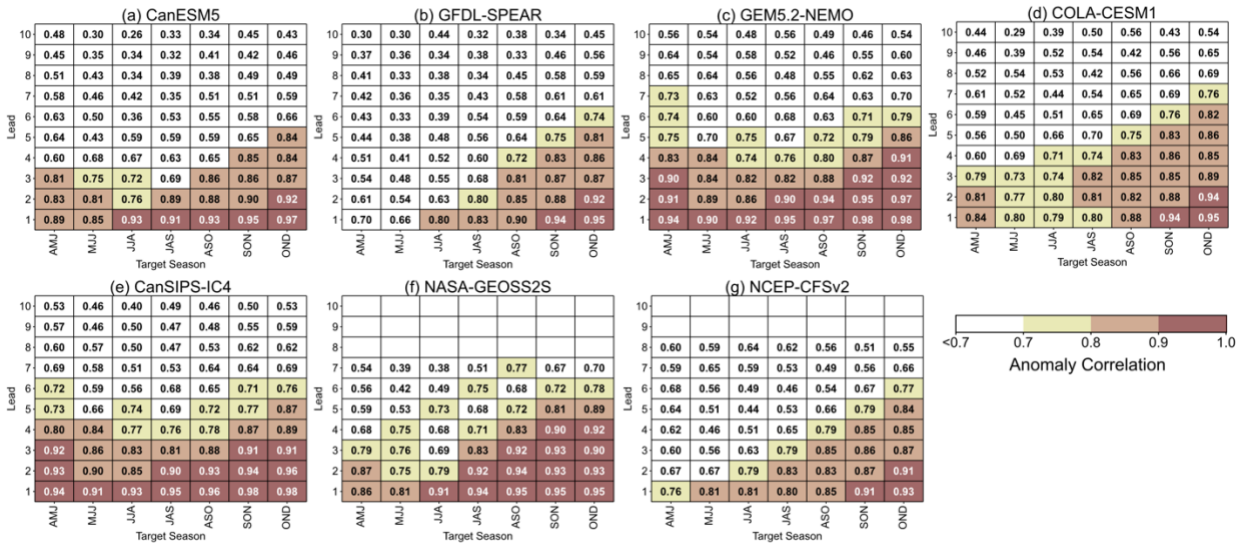

Figure 3: NMME anomaly correlation as a function of lead time (increasing toward the upper part of each panel) during ENSO onset seasons (AMJ to OND). Correlations above 0.7 are shaded, as this threshold explains at least 50% of the seasonal variance.

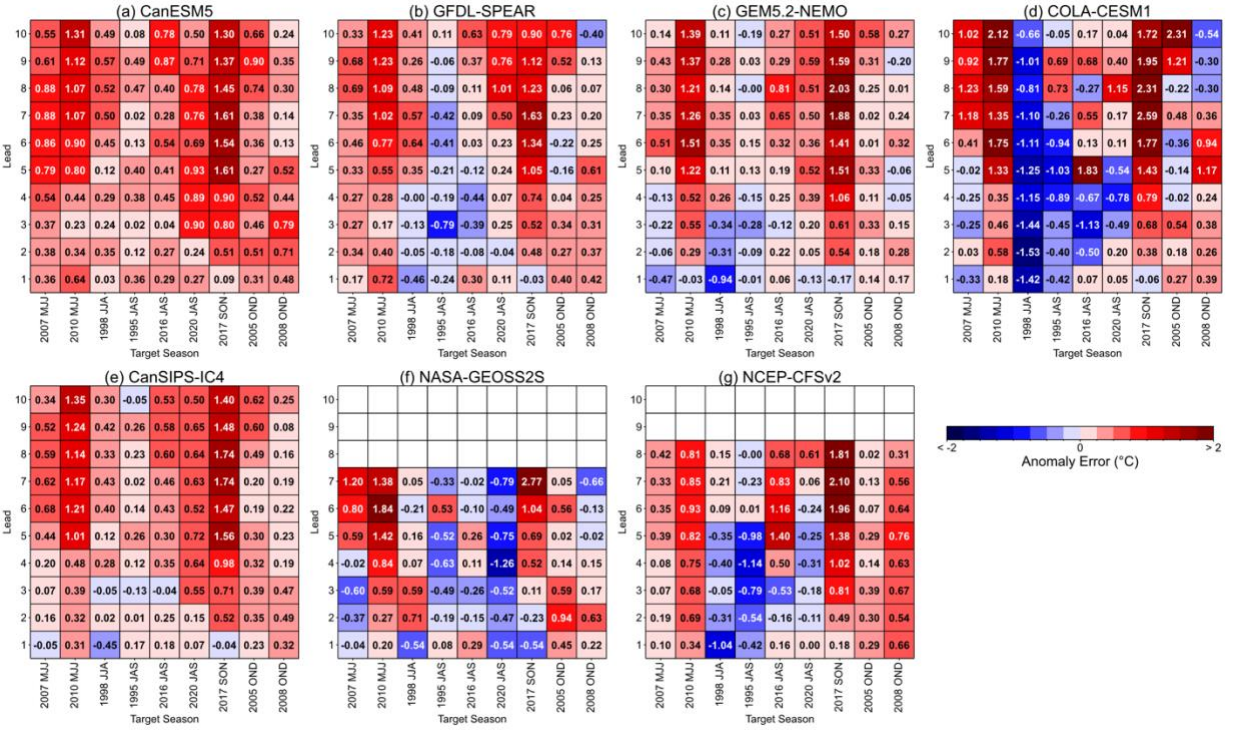

Figure 4: NMME anomaly errors, by cold phase onset, as a function of lead time (increasing toward the upper part of each panel). Positive values indicate underestimation of cooling trend, and positive values above 1.0 indicate prediction failure.

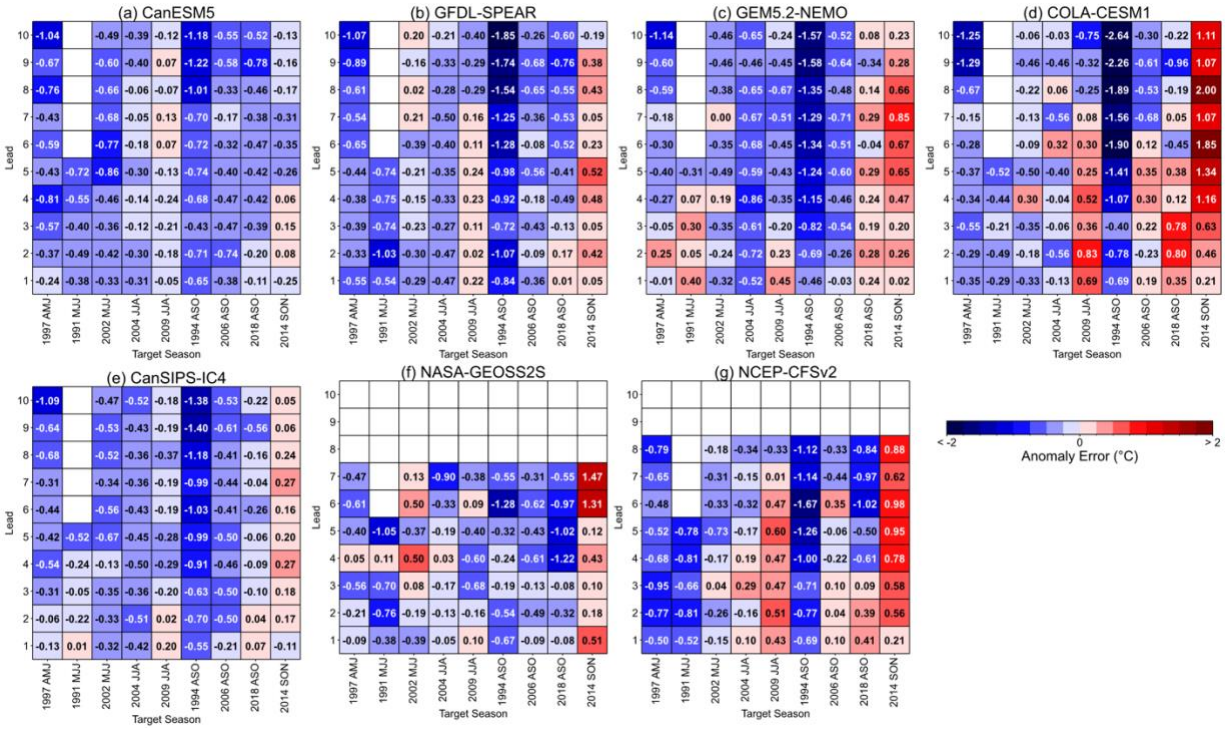

Figure 5: NMME anomaly errors, by *warm* phase onset, as a function of lead time (increasing toward the upper part of panel). Negative values indicate underestimation of the warming trend, and negative values under -1.0 indicate prediction failure.

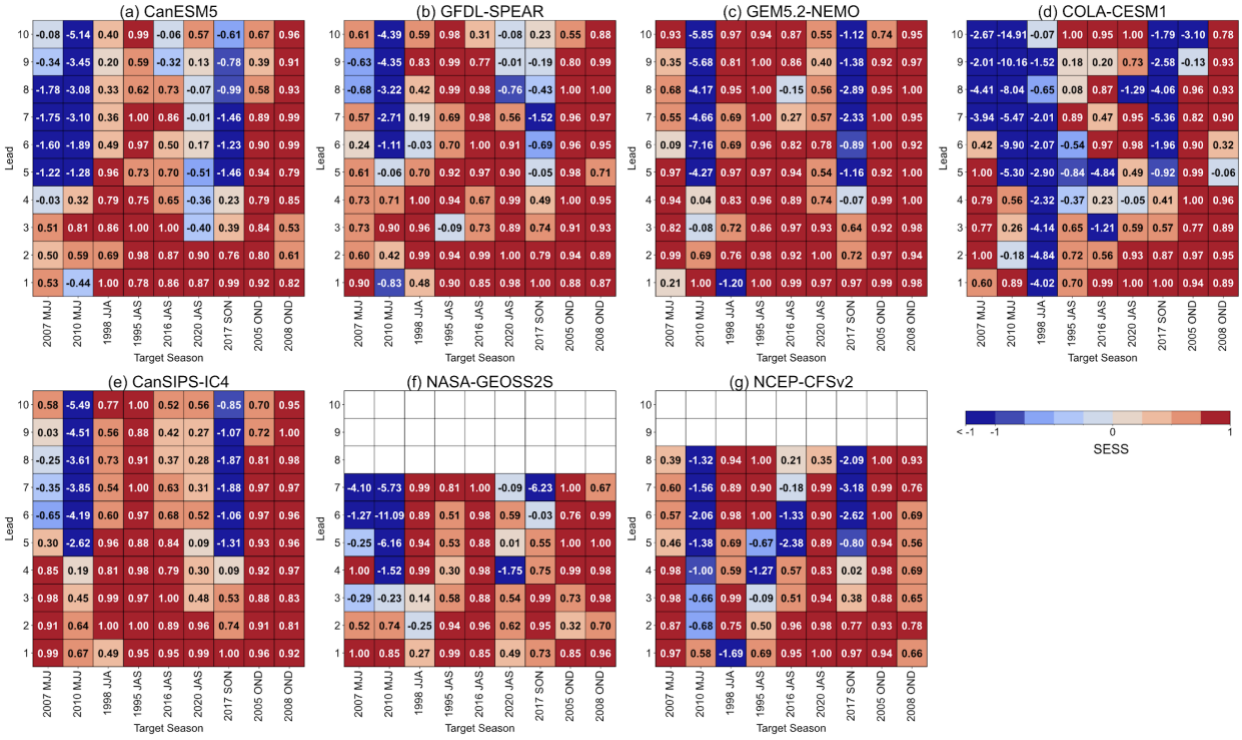

Figure 6: Squared Error Skill Score of NMME member models at *cold* onsets as a function of target season and lead in months. SESS close to 1 (*red*) indicates high skill while a negative SESS (*blue*) indicates low skill

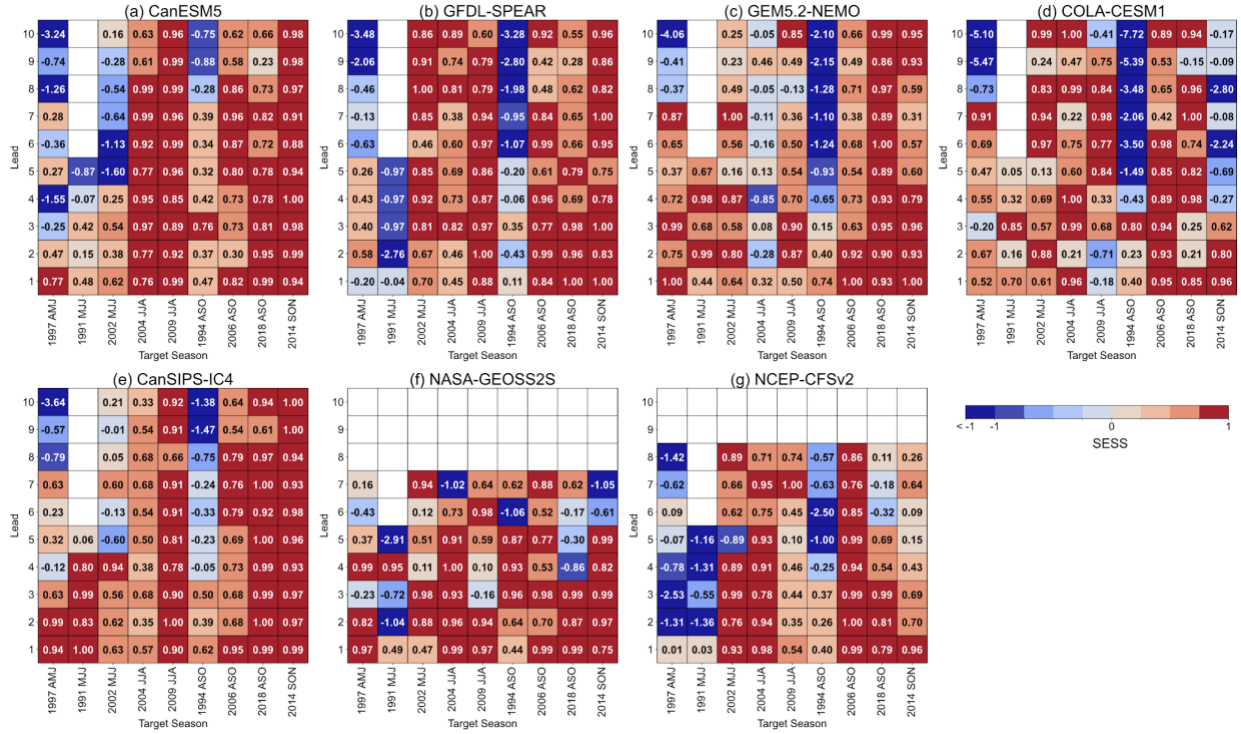

Figure 7: Squared Error Skill Score of NMME member models at *warm* onsets as a function of target season and lead in months. SESS close to 1 (*red*) indicates high skill while a negative SESS (*blue*) indicates low skill

Supplemental Figures

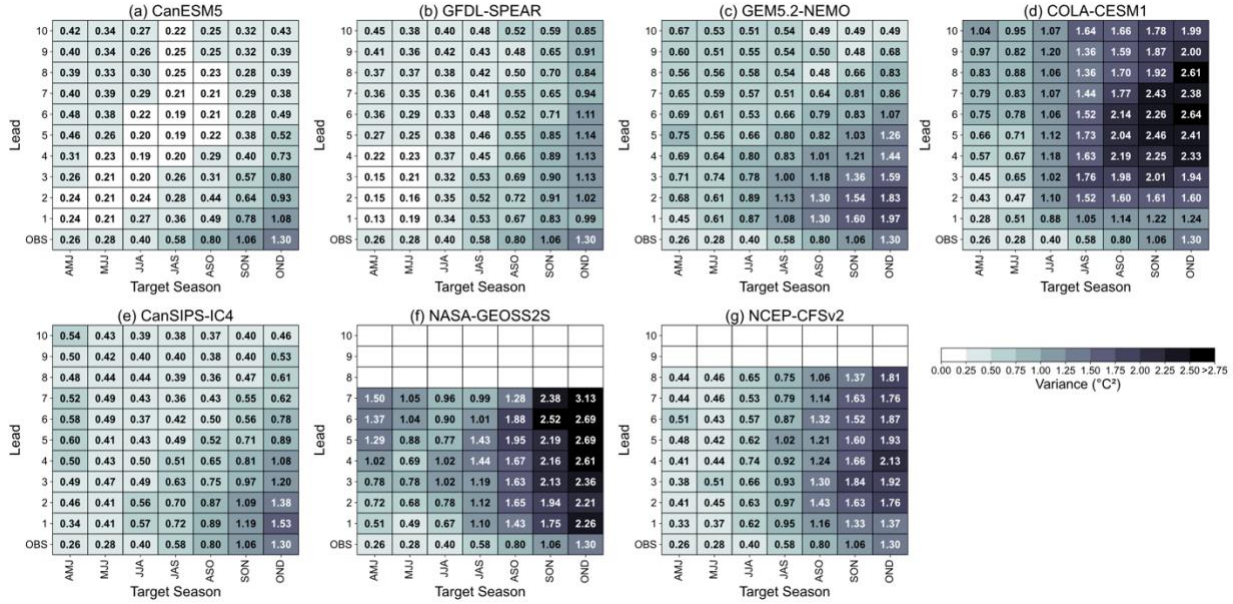

Figure S1: Lead-dependent member variance averages of NMME member models at all leads as a function of onset season. Higher values (darker cell) over observations (OBS) indicate overestimation of variance while lower values (lighter cell) over observations indicate underestimation of variance.

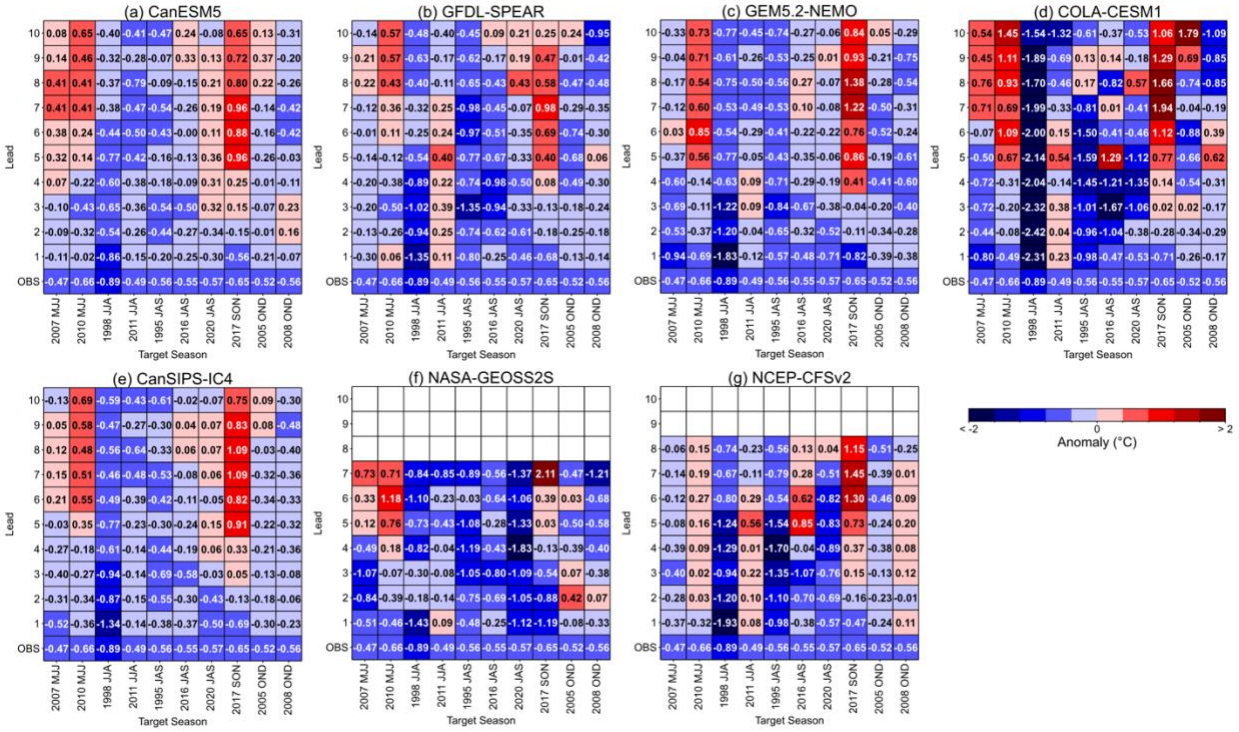

Figure S2: NMME Anomalies, by *cold* phase onset, as a function of lead time (increasing toward the upper part of panel). The observations are shown at the bottom row of each panel.

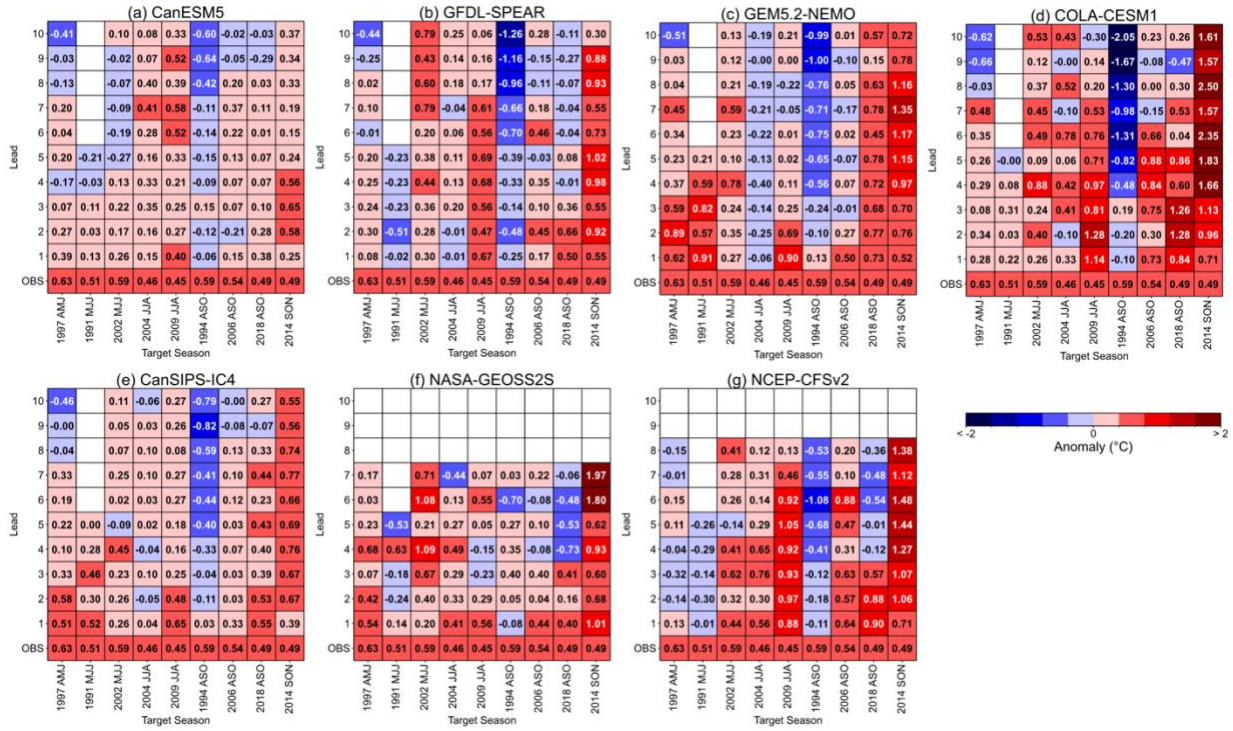

Figure S3: NMME Anomalies, by *warm* phase onset, as a function of lead time (increasing toward the upper part of each panel). The observations are shown at the bottom row of each panel.

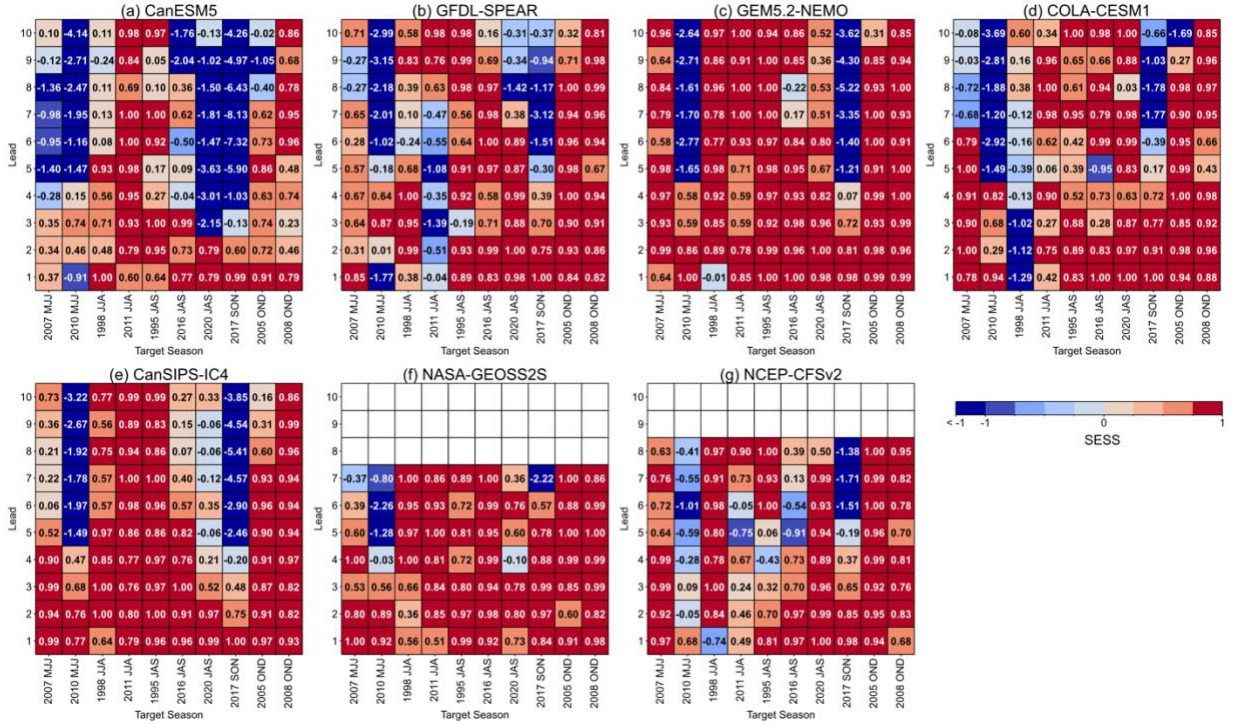

Figure S4: Squared Error Skill Score of NMME member models at *cold* onsets as a function of target season and lead in months. SESS close to 1 (*red*) indicates high skill while a negative SESS (*blue*) indicates low skill. SESS here was calculated using lead-dependent member variance from Fig. A

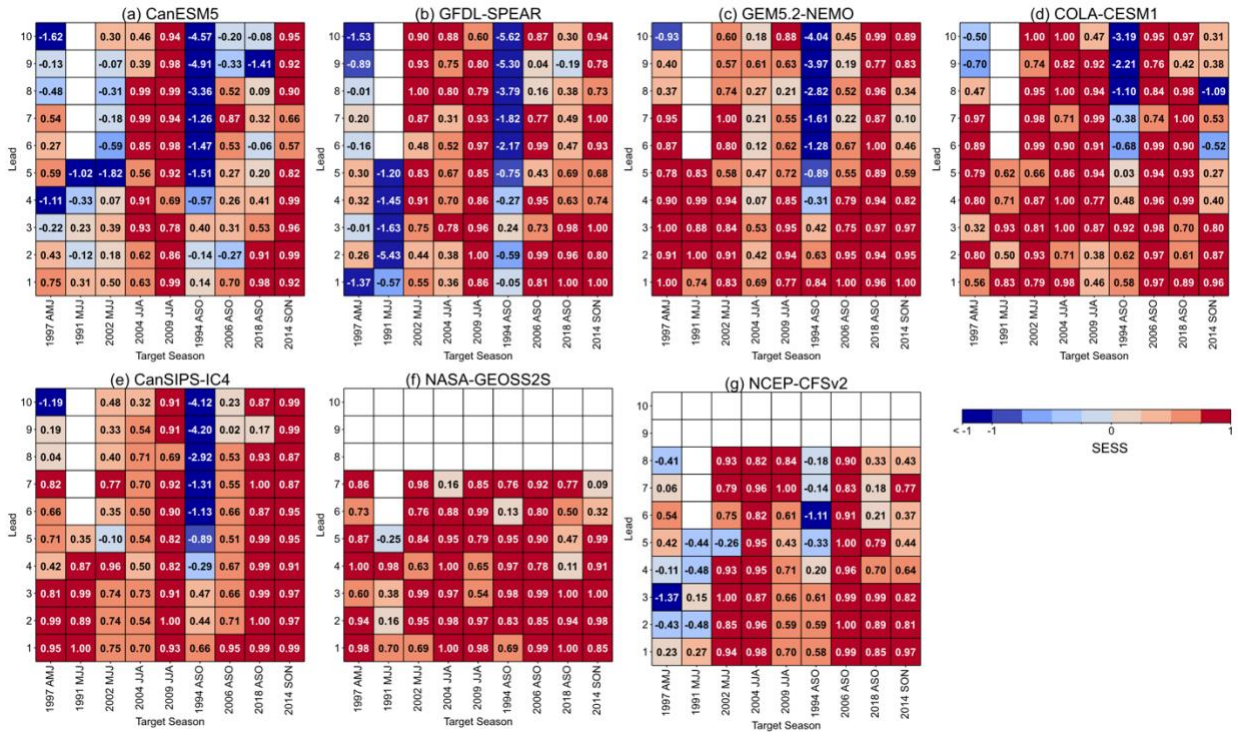

Figure S5: Squared Error Skill Score of NMME member models at *cold* onsets as a function of target season and lead in months. SESS close to 1 (*red*) indicates high skill while a negative SESS (*blue*) indicates low skill SESS was calculated using lead-dependent variance from Fig. 3