Spring onset predictability in the North American Multi-Model Ensemble

Carlos M. Carrillo, Toby R. Ault, Daniel S. Wilks

Nov 28, 2017

ABSTRACT

This study assesses the predictability of spring onset using an index of its interannual variability. We use the North American Multi-Model Ensemble (NMME) experiment to assess this predictability. The input dataset to compute spring onset index, SI-x, were treated with a daily joint bias correction (JBC) approach, and the SI-x outputs were post-processed using an ensemble model output statistic (EMOS) approach—non-homogeneous Gaussian regression. This EMOS approach quantifies the effect of training period length and ensemble size on forecast skill. The highest range of predictability for the timing spring onset is from 10 to 60 days, and it is located along a narrow band between 35° to 45°N in the US. Using rank probability scores based on quantiles (q), a forecast threshold (q) of 0.5 provides a range of predictability that falls into two categories 10-40 and 40-60 days, which seems to represent the effect of the intra-seasonal scale. Using higher thresholds (q=0.6 and 0.7) predictability shows lower range with values around 10-30 days. The post-processing work using JBC improves the predictability skill by 13% from uncorrected results. Using EMOS, a significant positive change in the skill score is noted in regions where the skill with JBC shows evidence of improvement. The consensus of these techniques shows that regions of better predictability can be expanded.
1. Introduction

Variations in the timing of spring onset affect ecosystems, fire forests, drought, pollen, and agriculture (Westerling et al., 2006; Ault et al., 2013). Given its importance to human and ecological health, there is a pressing need to characterize the potential predictability of spring onset on seasonal time horizons. In principle, such forecasts could be issued alongside seasonal predictions of more traditional variables like precipitation and temperature (Kirtman et al., 2014; Saha et al. 2014; Mo and Lettenmaier et al., 2014). However, the predictability of such seasonal transitions has not yet been widely explored.

Forecasting seasonal transitions can extend the usability of forecasts on seasonal time horizons. For example, dry conditions in central U.S. during early spring have been related to unusual wet late spring in the eastern half of the country (Koster et al., 2014; Schubert et al., 2016). Characterizing such transitions requires systematic indices that are consistent through space and time, such as the “extended” spring indices (SI-x) at Schwartz et al. (2013) and Ault et al. (2015). Development of this particular index relied on previous efforts that established a strong relationship between blooming of plants and the spring onset (Schwartz and Marotz, 1986; Cayan et al., 2001; Schwartz et al., 2006), and also linked the interannual variability of spring onset to large-scale atmospheric patterns and ocean forcing as noted in sea surface temperature (Ault et al., 2011).

Here, we evaluate the potential predictability of spring onset as determined by the SI-x (Schwartz et al., 2013; Ault et al., 2015). We focus on SI-x because it integrates temporal and spatial atmospheric patterns of variability across synoptic to intra-seasonal scales. As such, the SI-x serve as a proxy for spring onset across North America, and predicting the timing of this
seasonal transition may be critical for anticipating warm-season events at long lead times. That is, an early spring would lead to different ecological and agricultural risks in summer than a late spring because, for example, an early start to the growing season could favor invasive species or certain plant and human pathogens (Monahan et al., 2016). Specifically, we are interested in quantifying the lead times on which SI-x can be predicted. In addition, a state-of-the-art post processing technique—the ensemble model output statistics (EMOS)—is used to answer how the multi model-ensemble outperforms ensembles from individual models, and also whether longer reforecast training periods improve post-processing capacity by enhancing prediction skill.

2. Data and Methodology

2.1 Observational data and SI-x

The “extended spring index” (SI-x) used in this study was originally developed in Schwartz and Marotz (1986) and Schwartz et al. (2006), then updated for continental scale coverage in Schwartz et al. (2013). Briefly, it is a temperature-based index that identifies the day of year (DOY) when key early-spring phenological events are likely to occur. Its only time-varying inputs are minimum and maximum temperature, meaning that it can be applied over a wide range of temperate climates to yield a consistent metric of the start of spring at each location across space and over many years. Additional details on the assumptions and limitations are documented elsewhere (e.g., Ault et al., 2015) and the code for computing the SI-x is widely available through GitHub (https://github.com/cornell-eas/SI-X). We calculated SI-x from the Berkeley Earth surface temperature dataset (BEST, Rohde et al., 2013), which includes daily maximum (Tmax) and minimum (Tmin) temperatures at 1° lat/lon spatial resolution, obtained
from http://www.BerkleyEarth.lbl.org/data/. We use observational data over the period from 1981 to 2012 for comparison with the NMME reforecast data. Two variables are evaluated in this study: the “leaf” and “bloom” indices. However, the interannual variability in both indices is similar (Ault et al., 2011), and here we only show results for the leaf index as a proxy for the start of spring.

2.2 The NMME forecasted data

Forecasts of maximum and minimum temperature are obtained from the North American Multi-Model Ensemble (NMME) Phase 2 dataset (Kirtman et al., 2014), which includes multiple models and multiple ensemble members from individual models over the period from 1981 to present. All model fields were regridded to a uniform 1° lat/lon grid. As we are interested in spring onset, we only used forecasts initialized from January through April. Five models were used to assess SI-x predictability (Table I). Ten members per each model ensemble were used to be consistent with the weighting among models.

2.3 Skill Score metrics

We quantify the skill of post-processed NMME model predictions by comparing them to both climatology and uncorrected model output. To perform this evaluation, we apply two objective metrics that measure forecast skill improvement against a reference prediction: the Reduction of Variance Skill Score (SSclim) and the Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) Skill Score (SScrps). Both of these skill scores are variations of the generalized Skill Score (SS),

\[
SS = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\%, \tag{1}
\]
which measures the improvement (in units of percentages) of a given forecast \( A \) over the
departure of reference metric \( A_{\text{ref}} \) from the perfect forecast \( A_{\text{perf}} \) (Wilks, 2006).

The SSclim,

\[
SS_{\text{clim}} = \frac{MSE - MSE_{\text{clim}}}{0 - MSE_{\text{clim}}} \times 100\%,
\]

(2)
is based on the mean square error (MSE),

\[
MSE = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2,
\]

(3)
of observed \( o_k \) and forecasted \( y_k \) data. The reference metric \( A_{\text{ref}} \) is the MSE of the
climatology \( MSE_{\text{clim}} \),

\[
MSE_{\text{clim}} = \frac{1}{n} \sum_{k=1}^{n} (o_k - \bar{o})^2,
\]

(4)
where \( \bar{o} \) is the observed climatology, and the perfect forecast, \( A_{\text{perf}} \), is zero as it has zero MSE.

The SScrps is defined as:

\[
SS_{\text{crps}} = \frac{CRPS - CRPS_{\text{ref}}}{0 - CRPS_{\text{ref}}} \times 100\%,
\]

(5)
which is based on the Continuous Ranked Probability Score approach (CRPS):

\[
CRPS = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 \, dy,
\]

(6)
where \( F(y) \) is the continuous CDF of the predictand \( y \). The term \( F_o \) is the cumulative
probability step function defined by:
\[ F_o(y) = \begin{cases} 0, & y < \text{observed value} \\ 1, & y \geq \text{observed value} \end{cases} \] (7)

As SI-x follows a Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \), CRPS for a given observation \( o \) can be calculated using:

\[
\text{CRPS}(\mu, \sigma^2, o) = \sigma * \left\{ \left( \frac{o - \mu}{\sigma} \right) \left[ 2 \Phi \left( \frac{o - \mu}{\sigma} \right) - 1 \right] + 2 \phi \left( \frac{o - \mu}{\sigma} \right) - \frac{1}{\sqrt{\pi}} \right\}. \quad (8)
\]

Where \( \Phi(\ ) \) and \( \phi(\ ) \) are the CDF and PDF respectively.

## 2.4 Bias Correction

A joint bias correction technique (e.g. Thrasher et al., 2012) is applied to remove systematic model errors in both Tmax and Tmin temperature while preserving their covariance. This correction is required because SI-x is sensitive to the covariance of Tmax and Tmin, and bias correcting variables individually can generate physically unrealistic outcomes (Thrasher et al., 2012). As temperature variations tend to be normally distributed, we define the TT joint distribution to be a Gaussian variation (Wilks, 2011), which is motivated by the high correlation between daily Tmax and Tmin. After fitting the parameters of the joint distribution to gridded observations and NMME temperature, we follow a quantile remapping approach similar to the one described in Li et al. (2014; their Fig. 1). That is, first, we estimate the quantile of a Tmin value in the forecast CDF, then match this value to the same quantile in the (marginal) observational CDF; a bias corrected value for Tmin is therefore obtained by identifying the appropriate observed Tmin value for that quantile. Next, to bias correct Tmax, we condition its CDF on Tmin, then associated conditioned quantiles of simulated Tmax values with
observational ones. This procedure yields bias corrected values of Tmin and Tmax for every grid point for every day of each year, and preserves the covariance structure of Tmin and Tmax in the observations.

2.5 The Ensemble Model Output Statistics

In addition to biases, ensemble forecasts have dispersion errors from initial-condition sensitivity and model structural error, among other sources (Wilks, 2011). However, multi-model ensemble forecasts are amenable to estimating forecast-uncertainty distributions, which can be used to calibrate these ensembles probabilistically. Here, we use the non-homogeneous Gaussian regression (NGR; Gneiting et al., 2005) EMOS method to diagnose the effect of ensemble size and training period in SI-x forecast skill. Under this approach, the forecast-uncertainty distribution is assumed to be defined by a Gaussian distribution as indicated in equation 9, which describe the cumulative probability that a future observation $V$ be less than a forecast quantile $q$:

$$
\Pr(V \leq q) = \Phi \left[ \frac{q - (a + b \bar{x}_{ens})}{\sqrt{(c + d s^2_{ens})}} \right],
$$

(9)

where $\Phi[ ]$ indicates the evaluation of the cumulative distribution function, $\bar{x}_{ens}$ is the ensemble average, and $s^2_{ens}$ is the ensemble variance. The parameters $a$, $b$, $c$, and $d$ define the best mean ($\mu = a + b \bar{x}_{ens}$) and variance ($\sigma^2 = c + d s^2_{ens}$), which is used to generate the calibrated SI-x function. The mean is calculated with two parameters because the SI-x results show good agreement amount models. Therefore, our approach assumes that the mean and variance of the ensemble are exchangeable. These four parameters are estimated using a minimization of the average continuous ranked probability score (CRPS),
\[ CRPS = \frac{1}{n} \sum_{t=1}^{n} \sigma_t \left\{ \left( \frac{y_t - \mu_t}{\sigma_t} \right) \left[ 2\Phi \left( \frac{y_t - \mu_t}{\sigma_t} \right) - 1 \right] + 2\phi \left( \frac{y_t - \mu_t}{\sigma_t} \right) - \frac{1}{\sqrt{\pi}} \right\}, \] 

for a defined training period. Here \( \Phi(\ ) \) and \( \phi(\ ) \) indicate the evaluation of the CDF and PDF of the standard Gaussian distribution, respectively. In the CRPS equation, the \( y_t \) values are obtained from the training period, which is defined from the reforecast period of the NMME models. In this study, we fit the NGR-EMOS using four different training period lengths (15, 20, 25, and 30 years), and five ranges of ensemble numbers (10, 20, 30, 40, and 50).

3. Results and Discussion

3.1 SI-x climatology

Using mean day of year (DOY) values, the SI-x leaf index computed from the NMME is comparable to the observational pattern (Fig. 1). As in previous studies (e.g., Cayan et al., 2001; Ault et al., 2105), a prominent north-south gradient is present in both the observed data and the NMME ensemble mean. NMME means and standard deviations are computed over the full ensemble minus the multi-model mean without differences among individual model-specific distribution. Greater spatial heterogeneity is observed along the western Intermountain Region, which is not fully reproduced by the NMME mean (bottom panel of Fig 1). The standard deviation of the SI-x, in Fig. 2, shows less agreement between the observed pattern and model simulations. Although the simulations capture maximum variance in the Pacific Northwest and in the Southeastern US, it overestimates this variability within a range of four days in the Intermountain Region, west of the Rockies (bottom panel of Fig. 2). Thus, model biases are more apparent in the standard deviation than in the mean.
3.2 The SI-x Skill Score

We first evaluate the skill of NMME models in predicting SI-x without any statistical correction. The CRPS shows high values in the Intermountain Region (for January and February) and in the northeastern region of the domain (for March and April; Fig. 3). High CRPS values are associated with poor model performance. Nevertheless, these values are low (indicating good skill) in low-elevation terrain and when advancing through the season: e.g., they are better in March than February. However, the untreated input data does not outperform the CRPS climatology reference (CRPS[clim]; Fig. S1), which maximum value is in the order of 3.5 CRPS units. This CRPS climatology reference shows a coast-to-coast band around 35°N, which is consistent with the difference of variance between observation and NMME models (Fig. 2). This result shows that untreated input data is of minor use.

The previous results did not include the EMOS approach, so we applied it to the SI-x output to evaluate the dispersion error due to initialization and differences in model structure. We illustrate an example of the multi model-ensemble error dispersion using one-grid point for different model initialization times (Fig. S2). The temporal evolution from an early to late initialization (January to March) reveals the reduction of the error dispersion among the model-ensemble realizations. Therefore, the EMOS should use this information into its systematic approach to correct the final output. Using the NGR-EMOS approach, we found that this is indeed the case. Thus, the NGR-EMOS analysis of two datasets (untreated and JBC; Fig. 4) reveals a significant improvement respect to the untreated data (Fig. 3). Both NGR-EMOS results have similar spatial variability and validate the effectiveness of the approach. Although there is still a challenge to correct data in the Intermountain Region, the forecast skill is significant improved.
**Fig. 5** shows the continuous ranked probability score (CRPS) Skill Score (SScrps) for the entire set of forecasts, starting at January 1, February 1, March 1, and April 1 for the period from 1981 to 2012. The CRPS(clim) used to compute SScrps is shown in Fig. S1, and also an alternative SScrps field (when using the untreated CRPS as reference) is shown in **Fig. S3** to illustrate the valued added of the NGR-EMOS against the untreated dataset. However, an improved comparison from the climatology is a fair metric and hence it is used here. Thus, in January, several regions with improvement of 10% is observed in key places: the Southeast, the northwest Pacific, the Northeast, and the Southwest. Comparing with another metric (e.g. SSclim), similar SS for January shows the level of improvement of EMOS (**Fig. S4**). Nevertheless, the same can be inferred from the CRPS analysis (Figs. 3 and 4). In addition, results improve as the seasons progress, as should be expected because the initialization dates approach the onset day. In February, regions along the Southeast significantly improve. This is noted by the region with 30% positive percentage change (**Fig. 5**), which describes the improvement by NGR-EMOS. The major feature in February is the high percentage of error, on the order of negative 20%, in the Intermountain Region, as can be inferred from the analysis of the standard deviation anomaly (**Fig. 2**; bottom). In March, the region of improvement expands and migrates north, consistent with what was shown for January. Similar results are observed for the initializations starting in April. A region with positive SScrps change in all the months is located below 40°N—Missouri, Illinois, Ohio, Kentucky, West Virginia, and Virginia—where the major improvement is observed during January, February, and March. This was likely to happen because this region coincides with the maximum variability of the SI-x standard deviation (**Fig. 2**), near 85°W, 35°N. The SScrps spatial pattern is consistent with alternative metrics using one model and one-model ensemble (**Fig. S5**), its ensemble mean, and the NMME multi model ensemble (**Fig. S4**).
The agreement among the alternative metric means that the NMME ensemble mean of untreated input data does not significantly increase the forecast skill of the SI-x. This suggests that EMOS is able to add value by enhancing good SI-x individual forecast members in the NMME, which otherwise cancel out with the less sophisticated NMME ensemble mean.

3.3 The SI-x predictability range

The SScrps evaluates the forecast skill of SI-x as a percentage of the reference climatology (Fig. S1). However, to determine how many days in advance SI-x can estimate spring onset, the time-dependence of SScrps needs to be characterized. Fig. 6 shows how this additional metric is constructed for a region in the Great Plains (100°-90°W, 35°-45°N). First, the SScrps values for every initialization (January 1 through April 1) are calculated using the valid reference (Fig. S1). Second, the SScrps dependence on time is constructed based on the different model initialization dates, allowing us to compute the SI-x predictability range for a given SScrps level of improvement. Logically, predictive skill increases as the initialization date approaches the target date. In the worst forecast skill scenario, we could expect to have at least the same chance to make a forecast as good as the climatology, which means when SScrps=0.0. A SScrps value of 0.2 therefore represents a 20% improvement over the reference climatology, and similarly SScrps=0.10 corresponds to a 10% improvement. Therefore, the SScrps is used to define the level of improvement. Using SScrps=0.20, we estimate the DOY (Julian day) in the fitted SScrps time variation (dashed line in Fig. 6). In this example, a SI-x predictability range of 20 days is obtained according to the fitted SScrps versus time relationship.

The observed 20-day forecast skill is in the range of a model such as the Climate Forecast System version 2 (CFSv2) reported by Saha et al. (2014). This 20-day predictability range is the
one that Climate Services would potentially use as the most accurate information that includes a high level of probability (q=0.70), with a level of improvement that departs from the mean climatology by 20% (SScrps=0.20). As the SScrps is a specific characteristic of each model’s performance, different models have different SScrps values and predictability ranges for the same region (see Table II), and this information can be used to weight a final product or to eliminate some models in an optimal operational forecast. We repeated the analysis for the entire domain of the models and for each individual grid.

The SI-x predictability for the continental US is in the range of 10-60 days for the NGR-EMOS NMME (Fig. 7a). The SScrps threshold used here is 0.0, which is a threshold that is comparable with climatology. We extended the analysis to SScrps = 0.1 (Fig. 7b) and 0.2 (Fig. 7c), reducing both the temporal range and geographic extension of high forecast skill. This SI-x predictability, in the range of 10-60 days for SScrps=0.0, can be confirmed by the behavior of individual models (Fig. S6). The scale bar groups the forecast skill range into low (10-40 days) and high (40-60 days) to highlight the results in the intraseasonal and seasonal scales, with the goal of identifying them, but without assessing the source of what produces better results in these ranges. The relative low range of 10-40 days is characteristic of the northern Great Plains and part of the Intermountain Region, which was suggested from the analysis of the mean and variance shown previously. The high range of 40-60 days is shown north of 45°N and marginally in the Intermountain Region. These bands reflect the region of minimum variability in observed standard deviation (Fig. 2). The SI-x predictability range shows a north-south gradient as a typical characteristic seen in the SI-x climatology that reflects the seasonal march from winter to summer.
The multi-model ensemble NMME NGR-EMOS (Fig. 7) agrees well with the individual-model ensemble (Fig. S6), which portrays differences in the SI-x forecast skill when applying the TT-JBC approach. As expected, the spatial pattern of predictability differs among models. Although the GEOS-5 model shows the lowest range of predictability in the Intermountain Region, it shows better improvement after applying the TT-JBC, similarly for CanCM3 and CanCM4 along 45°N. The results with the TT-JBC are consistent with the biased temperature (Fig. S6; left panel), and in addition they show that the TT-JBC adds value in regions that already have considerable forecast skill. The improvement occurs mainly in both the Canadian (CanCM3 and CanCM4) and the NOAA (GEOS-5) models. As synoptic events (Schwartz and Marotz, 1986) are modeled with daily maximum and minimum temperature, the JBC applied on both temperatures might influence on the corrected final calculation of SI-x. Therefore, the bias correction applied over the individual models improves the forecast skill, however, it does not outperform NGR-EMOS (Fig. 4).

Using a multi-model ensemble NGR-EMOS (Fig. 7), the results for the five models can be summarized in two major points. Firstly, there is signal in the range of intraseasonal variability (10-60 days) in the NMME models when compared to climatology (SScrps=0.0), meaning the multi-model ensemble outperforms two months before the beginning of the spring, with a 50% chance of failure after the NGR-EMOS is applied. These changes are localized in two regions: the “corn belt” along 40°N (Nebraska, Iowa, Minnesota, and Illinois), and the Intermountain Region. Secondly, when using higher thresholds (SScrps = \{0.1,0.2\}), this range is reduced by 10 days (with some exceptions in small localized regions), with a lower reduction in the Great Plains. Thus, a large range is still found in the vicinity of the Corn Belt region that looks promising for potential agriculture related applications.
In addition, for different training periods and number of ensemble members (Fig. S7), the continuous ranked probability score (CRPS) shows two important aspects to consider when applying the EMOS in SI-x related products: (1) a long training period significantly increases the predictability score (e.g. from 15-year to 30-year; top panel Fig. S4); and (2) a large number of ensemble members also improves the RPS skill score (e.g. from 10 to 20 ensembles; bottom panel Fig. S7). Although the forecast skill was significantly improved when the skill was low, it is not improved much when the skill was already high. For example, the initialization in January (1-month) shows a smooth transition from 1.7 with 10 ensembles to 1.5 when using 20 ensemble members (lower CRPS numbers represent better score). When the skill is good (e.g. initialization in March at 3-month), increasing of the number of ensembles does not add much value to the forecast skill.

A spatial description of the Skill Score, after using the NGR-EMOS, reveals a significant improvement in the Corn Belt region (Fig. 5). It portrays the positive effect of NGR-EMOS for the four initializations (Jan, Feb, Mar, and Apr) using the SScrps skill score. When we compare the difference between the model-ensemble NMME mean and EMOS, the percentage of improvement is of the order of 50% (from 10% to 80% SS) and the extension of this improvement expand significantly respect to the untreated results. For example, in February and March, the Corn Belt region seems an important improvement, which is verified with the similar results obtained by other two EMOS: the logistic regression (LR) and Gaussian ensemble dressing (GED; results not shown). Therefore, EMOS adds significant value to the SI-x forecast products at all initialization stages.
4. Conclusions

This study assesses the seasonal predictability of spring onset using an index previously calibrated with plant phenology and variability of temperature (SI-x; Ault et al., 2015). A set of NMME models was treated with a daily joint bias correction approach and an ensemble model output statistic approach. Our findings show that untreated input data is of minor use, as it does not significantly increase the forecast skill of the SI-x. Also, the selected training period length and ensemble size affect the SI-x forecast skill. Long training periods and a large number of ensemble members significantly improve the SI-x predictability skill score. Because SI-x integrates temporal variations in the atmosphere at a continental scale, it helps us identify regions where maximum skill occurs over North America. This study provides insight into how reliable climate-based information helps to evaluate lead time on which spring onset can be forecasted skillfully.

The results presented here show that the best predictability for the spring onset is in the range from 10-60 days located along a narrow band between 35°-45°N. Using a forecast threshold (q) of 0.5, the range of predictability falls into two categories 10-40 and 40-60 days. Using higher thresholds (q=0.6 and 0.7) predictability shows lower range with values around 10-30 days (Fig. 5). The 40-60 day time horizon is notable, as it extends well beyond the 10 days barrier inherent to most meteorological forecasts. It is, however, broadly consistent with Koster et al. (2011), which found some skill in air temperature predictions on similar timescales, though the motivation and metrics of that study were different from ours. The region with better skill is in the core of the continent along 40°N, where the major variability of the SI-x is observed. This region is relevant because of its vicinity to the Corn Belt states that has great impacts to the local and global economy. Also, it is where the early and late spring variability was found (Schubert et
Becker et al. (2014) also show that NMME has good results in central US, which further supports our interpretation. Although the regions with better skill found in this study are narrow and localized, this opens for targeting other approaches in these sensitive regions to improve model skill with advanced post-processing techniques (e.g. EMOS).

Future work could include assessment of the atmospheric processes linked to early versus late spring onset. The dominant driver is potentially the Pacific Jetstream transition from winter into spring because its impact in western North America. Indices have been constructed that characterize the timing, position, and structure of the Pacific Jetstream (Newman and Sardeshmukh, 1998). The Pacific Jetstream migrates north, splits, and weakens. Therefore, the timing of this breakdown can be characterized in the intraseasonal range, which typically occurs between mid-March and mid-April. The range of predictability found in this study potentially support the existence of driving mechanism at this scale that might be orchestrating these ranges of predictability skill. Extreme variations of the spring onset affect ecosystems, fire forests, and especially agriculture (Ault et al., 2013). In the US, loss of maize during the 2012 extreme early spring onset was counted by 38% in Indiana (Nielson, 2012), and it is unknown how the system will respond when climate change unfolds. Therefore, right timing of seasonal transitions can potential protect economic assets much as synoptic forecast does today but with a longer time range.

Finally, our findings suggest that there is potential forecast skill in NMME products, but a sophisticated post-processing is necessary to achieve that potential. We portrayed how the predictability skill of NMME models to forecast spring onset in North America is improved with two post-processing techniques—the joint bias correction (JBC) and ensemble model output statistic (EMOS). The JBC outperforms the biased temperature SI-x product, and the
improvement mainly occurs in both the Canadian and NOAA models, however, it does not
outperform the multi-model ensemble EMOS. Using EMOS, a significant positive change in the
skill score is noted in regions where the skill with non-treated data is low. The consensus of both
techniques shows that regions of better predictability can be expanded (e.g. the Corn Belt
region). Therefore, adding these corrections would be important for any future operation use.

5. References

Ault R.T., Macalady AK., Pederson GT., Betancourt JL., Schwart MD., 2011. Northern
hemisphere modes of variability and the timing of spring in western North America. Journal of
Climate 224: 4003-4014.

Ault R.T., Henebry GM, de Beurs KM, Schwartz MD, Betancourt JL. Moore D. 2013. The False

Ault R.T., Zurita-Milla R., Schwartz M.D., A Matlab toolbox for calculating spring indices from

Ault TR, Labe Z, Zurita-Milla R. 2016. Identifying anomalously early spring onsets in the
CESM large ensemble project. Climate Dynamics :1-18.


Caprio J.M., 1993. Western regional phonological summary of information on honeysuckle and
lilac first bloom phase covering the period 1956-1991. Montana Agricultural Experiment Station
State Climate Center Circular. No. 3, 92 pp.

the onset of spring in the western United States. Bulletin of the American Meteorological Society
82:399-415.

Epstein ES. 1969. A scoring system for Probability Forecast of Ranked Categories. Journal of
Applied Meteorology 8: 985-987.

using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review
133: 1098-1118.


Table Captions

**Table I**: The North America Multi-Model Experiment (NMME) models and organizations.

**Table II**: The SI-x predictability range (in days) for a region in the Great Plains (100°-90°W,40°-45°N). These values are calculated from the untreated datasets (noJBC) for each individual model (CanCM3, CanCM4, FLORB01, and GEOS-5), individual ensemble members (E1, …, E10), and the multi model-ensemble mean (NMME). The last column is the same data but after treated with the Joint Bias Correction (JBC) approach.
Figure Captions

**Figure 1:** Mean day of the year (DOY) of the Spring Index (SI-x) for the leaf parameter during the period 1981-2011 for the Berkeley Earth surface temperature (OBS; top), for the multi-model ensemble (middle) of the North America Multi-Model Ensemble (NMME), and the OBS minus NMME difference (bottom).

**Figure 2:** Standard deviation (STD) of the Spring Index (SI-x) for the leaf parameter during the period 1981-2012 for the Berkeley Earth surface temperature dataset (OBS; top), for the multi-model ensemble (middle) of the North America Multi-Model Ensemble (NMME), and the OBS minus NMME difference (bottom).

**Figure 3:** Spatial pattern of the Continuous Ranked Probability Score (crps) of the Spring Index (SI-x) for the leaf parameter. The crps is computed using models from the North America Multi-Model Experiment (NMME) without post-processing treatment for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). A total of 50 model-ensemble are using for each realization and the period of analysis is from 1981 to 2012. High values of crps indicate poor model performance. Oblique lines show regions where the spring index is already reach for each panel.

**Figure 4:** Left panel: as in Fig. 3 but crps is computed after the non homogeneous Gaussian regression (NGR) is applied to SI-x (untreated dataset). Right panel: as in Fig. 3 but crps is computed after the NGR is applied to SI-x , which was computed with maximum and minimum temperature bias corrected using the joint bias correction (JBC) approach.

**Figure 5:** Spatial pattern of the skill score (SS) for the Spring Index (SI-x). The SS is computed using the Continuous Ranked Probability Score (crps) as measurement metric for the SI-x computed using models from the North America Multi-Model Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement in percentage respect to the reference climatology (Fig. S1) and oblique lines show regions where the Spring Index is already reach for each panel.

**Figure 6:** The continuous ranked probability skill score (SScrps) versus time plot for initializations in January 1, February 1, March 1, April 1, and May 1. The SScrps is calculated for an average region in the Great Plains (100-90°W, 35-45°N). The solid line is the fitted curve with a second order polynomial, and the dashed lines highlight the value for the 0.2 SScrps.

**Figure 7:** The SI-x predictability range measured in days before the occurrence of the spring onset. The SI-x predictability range is computed as in Fig. 6 and for three Skill Score (SScrps) thresholds: 0.0 (a), 0.1 (b), and 0.2 (c). Oblique lines indicating regions where the calculation of the metric is not possible.

Supplementary Material
**Figure S1:** Spatial pattern of the Continuous Ranked Probability Score (crps) of the Spring Index (SI-x) for the observed leaf parameter, which is considered as the reference climatology (CLIM-REF) used in the computation of the skill score. The SI-x is computed for the period 1981-2012, using observed Berkeley Earth surface temperature.

**Figure S2:** Spring onset index (SI-x) frequency distribution of the multi model-ensemble realizations of the North America Multi-Model Ensemble (NMME) experiment for a grid at 40°N, 82.5°W. Each frequency distribution is for different initialization time: January (top), February (middle), and March (bottom) during the 2005 forecast. The red vertical line is the observed SI-x for that year, and the blue vertical line is added as reference for the forecast initialization.

**Figure S3:** Spatial pattern of the skill score (SS) for the Spring Index (SI-x). The SS is computed using the Continuous Ranked Probability Score (crps) as measurement metric. The SI-x is computed using models from the North America Multi-Model Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement in percentage respect to the untreated crps (Fig. 3) and oblique lines show regions where the Spring Index is already reach for each panel.

**Figure S4:** Spatial pattern of the Reduction of Variance Skill Score (SSclim) of the Spring Index (SI-x) for the leaf parameter using the multi model-ensemble mean of the North America Multi-Model Ensemble (NMME) for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement respect to the climatology and oblique lines show regions where the Spring Index is already reach for each panel.

**Figure S5:** Left panel: spatial pattern of the Reduction of Variance Skill Score (SSclim) of the Spring Index for the leaf parameter using one ensemble member (E1) of the CanCM4 model for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement respect to the climatology and oblique lines show regions where the Spring Index is already reach for each panel. Right panel: same as left panel but for the CanCM4 ensemble mean.

**Figure S6:** Left panel: the SI-x predictability range measured in days before the occurrence of the spring onset. The SI-x predictability range is computed based on a Skill Score (SScrps) threshold equal to 0.0. These spatial patterns are calculated for each model ensemble individually (CanCM3, CanCM4, CESM1, FLORB01, and GEOS-5) as indicated in the lower left corner. These spatial pattern maps were obtained with untreated temperatures. Oblique lines indicate regions where the calculation of the metric is not possible. Right panel: same as left panel but for SI-x with bias corrected temperatures using a joint bias correction approach.

**Figure S7:** Top panel: the continuous ranked probability score (CRPS) for the 50 ensemble-model realizations for different training periods (15, 20, 25, and 30 years) and lead time (x-axis: 1, 2, 3, and 4 months). Low CRPS values represent better score. Bottom panel: similar as top panel but fixing the training period to 30 years and changing both the ensemble size (x-axis: 10, 20, 30, 40, and 50) and lead time (1, 2, 3, and 4 months).
<table>
<thead>
<tr>
<th>Organization</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Centers for Environmental Prediction (NOAA-NCEP)</td>
<td>CFSv2</td>
</tr>
<tr>
<td>Geophysical Fluid Dynamics Laboratory (NOAA-GFDL)</td>
<td>FLOR</td>
</tr>
<tr>
<td>Global Modeling and Assimilation Office (NASA-GMAO)</td>
<td>GEOS5</td>
</tr>
<tr>
<td>Environmental Canada</td>
<td>CMC1-CanCM3</td>
</tr>
<tr>
<td></td>
<td>CMC2-CanCM4</td>
</tr>
<tr>
<td>National Center for Atmospheric Research (NCAR)</td>
<td>CESM1</td>
</tr>
<tr>
<td></td>
<td>CCSM4</td>
</tr>
</tbody>
</table>

Table I: The North America Multi-Model Experiment (NMME) models and organizations
<table>
<thead>
<tr>
<th>Model</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
<th>E8</th>
<th>E9</th>
<th>E10</th>
<th>noJBC</th>
<th>JBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CanCM3</td>
<td>18.8</td>
<td>30.7</td>
<td>23.3</td>
<td>27.6</td>
<td>21.8</td>
<td>26.4</td>
<td>26.4</td>
<td>23.8</td>
<td>22.0</td>
<td>19.3</td>
<td>24.0</td>
<td>24.8</td>
</tr>
<tr>
<td>CanCM4</td>
<td>23.2</td>
<td>27.4</td>
<td>22.6</td>
<td>27.2</td>
<td>24.1</td>
<td>23.0</td>
<td>24.6</td>
<td>25.2</td>
<td>25.8</td>
<td>25.0</td>
<td>24.8</td>
<td>27.3</td>
</tr>
<tr>
<td>CESM1</td>
<td>13.8</td>
<td>14.0</td>
<td>14.8</td>
<td>17.1</td>
<td>19.6</td>
<td>13.1</td>
<td>17.2</td>
<td>12.2</td>
<td>12.9</td>
<td>19.3</td>
<td>15.4</td>
<td>23.6</td>
</tr>
<tr>
<td>FLORB01</td>
<td>24.7</td>
<td>25.3</td>
<td>30.9</td>
<td>30.0</td>
<td>14.3</td>
<td>16.1</td>
<td>11.3</td>
<td>22.8</td>
<td>13.0</td>
<td>10.9</td>
<td>19.9</td>
<td>19.5</td>
</tr>
<tr>
<td>GEOS-5</td>
<td>18.2</td>
<td>18.9</td>
<td>12.8</td>
<td>19.1</td>
<td>25.2</td>
<td>21.4</td>
<td>22.6</td>
<td>22.8</td>
<td>17.7</td>
<td>22.2</td>
<td>20.1</td>
<td>23.1</td>
</tr>
<tr>
<td>NMME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20.9</td>
<td>23.7</td>
</tr>
</tbody>
</table>

Table II: The SI-x predictability range (in days) for a region in the Great Plains (100°-90°W, 40°-45°N). These values are calculated from the untreated datasets (noJBC) for each individual model (CanCM3, CanCM4, FLORB01, and GEOS-5), individual ensemble members (E1, …, E10), and the multi model-ensemble mean (NMME). The last column is the same data but after treated with the Joint Bias Correction (JBC) approach.
Figure 1: Mean day of the year (DOY) of the Spring Index (SI-x) for the leaf parameter during the period 1981-2011 for the Berkeley Earth surface temperature (OBS; top), for the multi model-ensemble (middle) of the North America Multi-Model Ensemble (NMME), and the OBS minus NMME difference (bottom).
Figure 2: Standard deviation (STD) of the Spring Index (SI-x) for the leaf parameter during the period 1981-2012 for the Berkeley Earth surface temperature dataset (OBS; top), for the multi-model ensemble (middle) of the North America Multi-Model Ensemble (NMME), and the OBS minus NMME difference (bottom).
Figure 3: Spatial pattern of the Continuous Ranked Probability Score (crps) of the Spring Index (SI-x) for the leaf parameter. The crps is computed using models from the North America Multi-Model Experiment (NMME) without post-processing treatment for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). A total of 50 model-ensemble are using for each realization and the period of analysis is from 1981 to 2012. High values of crps indicate poor model performance. Oblique lines show regions where the spring index is already reach for each panel.
**Figure 4:** Left panel: as in Fig. 3 but crps is computed after the non homogeneous Gaussian regression (NGR) is applied to SI-x (untreated dataset). Right panel: as in Fig. 3 but crps is computed after the NGR is applied to SI-x, which was computed with maximum and minimum temperature bias corrected using the joint bias correction (JBC) approach.
Figure 5: Spatial pattern of the skill score (SS) for the Spring Index (SI-x). The SS is computed using the Continuous Ranked Probability Score (crps) as measurement metric for the SI-x computed using models from the North America Multi-Model Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement in percentage respect to the reference climatology (Fig. S1) and oblique lines show regions where the Spring Index is already reach for each panel.
**Figure 6:** The continuous ranked probability skill score (SScrps) versus time plot for initializations in January 1, February 1, March 1, April 1, and May 1. The SScrps is calculated for an average region in the Great Plains (100°-90°W, 35°-45°N). The solid line is the fitted curve with a second order polynomial, and the dashed lines highlight the value for the 0.2 SScrps.
Figure 7: The SI-x predictability range measured in days before the occurrence of the spring onset. The SI-x predictability range is computed as in Fig. 6 and for three Skill Score (SScrps) thresholds: 0.0 (a), 0.1 (b), and 0.2 (c). Oblique lines indicating regions where the calculation of the metric is not possible.
Figure S1: Spatial pattern of the Continuous Ranked Probability Score (crps) of the Spring Index (SI-x) for the observed leaf parameter, which is considered as the reference climatology (CLIM-REF) used in the computation of the skill score. The SI-x is computed for the period 1981-2012, using observed Berkeley Earth surface temperature.
Figure S2: Spring onset index (SI-x) frequency distribution of the multi model-ensemble realizations of the North America Multi-Model Ensemble (NMME) experiment for a grid at 40°N, 82.5°W. Each frequency distribution is for different initialization time: January (top), February (middle), and March (bottom) during the 2005 forecast. The red vertical line is the observed SI-x for that year, and the blue vertical line is added as reference for the forecast initialization.
Figure S3: Spatial pattern of the skill score (SS) for the Spring Index (SI-x). The SS is computed using the Continuous Ranked Probability Score (crps) as measurement metric. The SI-x is computed using models from the North America Multi-Model Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement in percentage respect to the untreated crps (Fig. 3) and oblique lines show regions where the Spring Index is already reach for each panel.
**Figure S4:** Spatial pattern of the Reduction of Variance Skill Score (SSclim) of the Spring Index (SI-x) for the leaf parameter using the multi model-ensemble mean of the North America Multi-Model Ensemble (NMME) for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement respect to the climatology and oblique lines show regions where the Spring Index is already reach for each panel.
Figure S5: Left panel: spatial pattern of the Reduction of Variance Skill Score (SSclim) of the Spring Index for the leaf parameter using one ensemble member (E1) of the CanCM4 model for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate improvement respect to the climatology and oblique lines show regions where the Spring Index is already reach for each panel. Right panel: same as left panel but for the CanCM4 ensemble mean.
Figure S6: Left panel: the SI-x predictability range measured in days before the occurrence of the spring onset. The SI-x predictability range is computed based on a Skill Score (SScrps) threshold equal to 0.0. These spatial patterns are calculated for each model ensemble individually (CanCM3, CanCM4, CESM1, FLORB01, and GEOS-5) as indicated in the lower left corner. These spatial pattern maps were obtained with untreated temperatures. Oblique lines indicate regions where the calculation of the metric is not possible. Right panel: same as left panel but for SI-x with bias corrected temperatures using a joint bias correction approach.
**Figure S7:** Top panel: the continuous ranked probability score (CRPS) for the 50 ensemble-model realizations for different training periods (15, 20, 25, and 30 years) and lead time (x-axis: 1, 2, 3, and 4 months). Low CRPS values represent better score. Bottom panel: similar as top panel but fixing the training period to 30 years and changing both the ensemble size (x-axis: 10, 20, 30, 40, and 50) and lead time (1, 2, 3, and 4 months).