AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2018JD028597

Key Points:

- The lead time for predicting the timing spring onset falls into two categories: 10–40 and 40–60 days
- The postprocessing work in this study improves the predictive skill for the timing of spring onset relative to the untreated input data set
- These findings suggest that the start of spring might be predictable on intraseasonal time horizons

Supporting Information:

- Supporting Information S1

Correspondence to:
C. M. Carrillo,
carrillo@cornell.edu

Citation:
Carrillo, C. M., Ault, T. R., & Wilks, D. S. (2018). Spring onset predictability in the North American multimodel ensemble. *Journal of Geophysical Research: Atmospheres*, *123*, 5913–5926. https://doi.org/10.1029/2018JD028597

Received 1 MAR 2018
Accepted 14 MAY 2018
Accepted article online 22 MAY 2018
Published online 12 JUN 2018

Spring Onset Predictability in the North American Multimodel Ensemble

Carlos M. Carrillo[1], Toby R. Ault[1], and Daniel S. Wilks[1]

[1]Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

Abstract The predictability of spring onset is assessed using an index of its interannual variability (the "extended spring index" or SI-x) and output from the North American Multimodel Ensemble reforecast experiment. The input data to compute SI-x were treated with a daily joint bias correction approach, and the SI-x outputs computed from the North American Multimodel Ensemble were postprocessed using an ensemble model output statistic approach—nonhomogeneous Gaussian regression. This ensemble model output statistic approach was used to quantify the effects of training period length and ensemble size on forecast skill. The lead time for predicting the timing of spring onset is found to be from 10 to 60 days, with the higher end of this range located along a narrow band between 35°N to 45°N in the eastern United States. Using continuous rank probability scores and skill score (SS) thresholds, this study demonstrates that ranges of positive predictability of SI-x fall into two categories: 10–40 and 40–60 days. Using higher skill thresholds (SS equal to 0.1 and 0.2), predictability is confined to a lower range with values around 10–30 days. The postprocessing work using joint bias correction improves the predictive skill for SI-x relative to the untreated input data set. Using nonhomogeneous Gaussian regression, a positive change in the SS is noted in regions where the skill with joint bias correction shows evidence of improvement. These findings suggest that the start of spring might be predictable on intraseasonal time horizons, which in turn could be useful for farmers, growers, and stakeholders making decisions on these time scales.

1. Introduction

Variations in the timing of spring onset affect ecosystems, forest fires, drought, pollen, and agriculture (Ault et al., 2013; Westerling et al., 2006). 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; Mo & Lettenmaier, 2014; Saha 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. Characterizing such transitions requires systematic indices that are consistent through space and time, such as the "extended" spring index (SI-x) of 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 (Cayan et al., 2001; Schwartz et al., 2006; Schwartz & Marotz, 1986) 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 characterized by the SI-x (Ault et al., 2015; Schwartz et al., 2013). We focus on SI-x because it integrates temporal and spatial atmospheric patterns of variability across synoptic to intraseasonal scales. As such, the SI-x serves 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 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 ensemble postprocessing technique—nonhomogeneous Gaussian regression (NGR)—is used to answer how the multimodel ensemble outperforms ensembles from individual models, and also whether longer reforecast training periods improve postprocessing capacity by enhancing prediction skill.

Table 1
The NMME Models and Organizations

Organization	Model
Geophysical Fluid Dynamics Laboratory (NOAA-GFDL)	FLORB01
Global Modeling and Assimilation Office (NASA-GMAO)	GEOS5
Environmental Canada	CMC1-CanCM3
	CMC2-CanCM4
National Center for Atmospheric Research (NCAR)	CESM1

Note. NMME = North American Multimodel Experiment; NASA = National Aeronautics and Space Administration; NOAA = National Oceanic and Atmospheric Administration.

2. Data and Methodology

2.1. Observational Data and SI-x

The SI-x used in this study was originally developed in Schwartz and Marotz (1986) and Schwartz et al. (2006) and 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 daily minimum and maximum temperatures, 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 data set (Rohde et al., 2013), which includes daily maximum (T_{max}) and minimum (T_{min}) temperatures at 1° lat/lon spatial resolution, obtained from http://www.BerkleyEarth.lbl.org/data/. We use observational data over the period 1981 to 2012 for comparison with the NMME reforecast data. Two variables were evaluated in this study: the “leaf” and “bloom” indices. The leaf and bloom indices pertain to the first leaf and bloom of plants related to seasonal transition of phenology due to variations of weather and climate. 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 Forecast Data

Forecasts of daily maximum and minimum temperature are obtained from the North American Multimodel Ensemble (NMME) Phase 2 data set (Kirtman et al., 2014), which includes multiple models and multiple ensemble members from individual models over the period 1981 to present (http://www.cpc.noaa.gov/products/NMME/data.html). All model fields were bilinearly 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 1). Ten members, fixed ensemble members numbered from 1 to 10, per each model ensemble were used to be consistent with the weighting among models, and each ensemble member yields a single SI-x field that is considered for the construction of the statistical approaches.

2.3. Skill Score Metrics

We quantify the skill of postprocessed 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 & Winkler, 1976) skill score (SScrps). Both of these skill scores are variations of the generalized skill score (SS)

$$\mathrm{SS}=\frac{A-A_{\mathrm{ref}}}{A_{\mathrm{perf}}-A_{\mathrm{ref}}}*100\%, \tag{1}$$

which measures the accuracy improvement (in units of percentages) of a given forecast (A) over the departure of reference metric (A_{ref}) from the perfect forecast (A_{perf}; Wilks, 2011).

The SSclim,

$$\mathrm{SS}_{\mathrm{clim}}=\frac{\mathrm{MSE}-\mathrm{MSE}_{\mathrm{clim}}}{0-\mathrm{MSE}_{\mathrm{clim}}}*100\%, \tag{2}$$

is based on the mean square error (MSE),

$$\mathrm{MSE}=\frac{1}{n}\sum\nolimits_{k=1}^{n}\left(y_k-o_k\right)^2, \tag{3}$$

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

$$\mathrm{MSE}_{\mathrm{clim}}=\frac{1}{n}\sum\nolimits_{k=1}^{n}\left(o_k-\overline{o}\right)^2, \tag{4}$$

where $\overline{o}$ is the observed climatological average, and the perfect forecast, A_{perf}, is zero as it has zero MSE.

The SScrps is defined as

$$\mathrm{SS}_{\mathrm{crps}} = \frac{\mathrm{CRPS} - \mathrm{CRPS}_{\mathrm{ref}}}{0 - \mathrm{CRPS}_{\mathrm{ref}}} * 100\%, \tag{5}$$

which is based on the CRPS:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 \mathrm{d}y, \tag{6}$$

where $F(y)$ is the continuous cumulative distribution function (CDF) of the predictand y. The term F_o is the cumulative probability step function defined by

$$F_o(y) = \begin{cases} 0, y < \text{observed value} \\ 1, y \geq \text{observed value} \end{cases}. \tag{7}$$

As SI-x follows an approximately Gaussian distribution with mean μ and variance σ^2 (e.g., Ault et al., 2015), CRPS for a given observation o can be calculated using

$$\mathrm{CRPS}(\mu, \sigma^2, o) = \sigma * \left\{ \left(\frac{o - \mu}{\sigma}\right) \left[2\Phi\left(\frac{o - \mu}{\sigma}\right) - 1 \right] + 2\varnothing\left(\frac{o - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\}, \tag{8}$$

where Φ() and ϕ() are the CDF and PDF, respectively, of the standard Gaussian distribution. Equation (8) is used when CRPS is used to evaluate NGR-based forecasts, where the forecast is defined as Gaussian distribution. Alternatively, we employed an ensemble version of the CRPS, which operates on the full discrete ensemble. The ensemble CRPS (eCRPS) is based on the alternative formulation for equation (8) (Gneiting & Raftery, 2007):

$$\mathrm{CRPS}(F, o) = E_F|X - o| - \frac{1}{2} E_F \left\lceil X - X^{'} \right\rceil, \tag{9}$$

where E_F denotes statistical expectation with respect to the predictive distribution F (x), and X and X' are independent realizations from F (x). Substitution of sample averages from the forecast ensemble for the expectations in equation (8) (Ferro et al., 2008; Van Schaeybroeck & Vannitsem, 2015) yields

$$e\mathrm{CRPS} = \frac{1}{n} \sum_{t=1}^{n} \left[\frac{1}{m} \sum_{k=1}^{m} \left(x_{t,k} - y_t\right) - \frac{\delta_t}{2} \right], \tag{10}$$

where

$$\delta_t = \frac{1}{m(m-1)} \sum_{j=1}^{m} \left[\sum_{k=1}^{m} |x_{t,j} - x_{t,k}| \right]. \tag{11}$$

For our application $x_{t,\ j}$ and $x_{t,\ k}$ are raw ensemble members, m is the total number of members, 50, and y_t is observation.

2.4. Bias Correction

A joint bias correction (JBC) technique (e.g., Thrasher et al., 2012) is applied to remove systematic model errors in both T_{max} and T_{min} temperature while preserving their covariance. This correction is required because SI-x is sensitive to the covariance of T_{max} and T_{min}, 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 joint distribution of daily maximum and minimum temperatures to be bivariate Gaussian (Wilks, 2011), which is motivated by the high correlation between daily T_{max} and T_{min}. After fitting the parameters of the joint distribution to gridded observations and NMME temperatures, we follow a quantile remapping approach similar to the one described in Li et al. (2014, their Figure 1). First, we estimate the quantile of a T_{min} value in the forecast CDF and then match this value to the same quantile in the (marginal) observational CDF; a bias corrected value for T_{min} is therefore obtained by identifying the appropriate observed T_{min} value for that quantile. Next, to bias-correct T_{max}, we condition its CDF on T_{min} and then

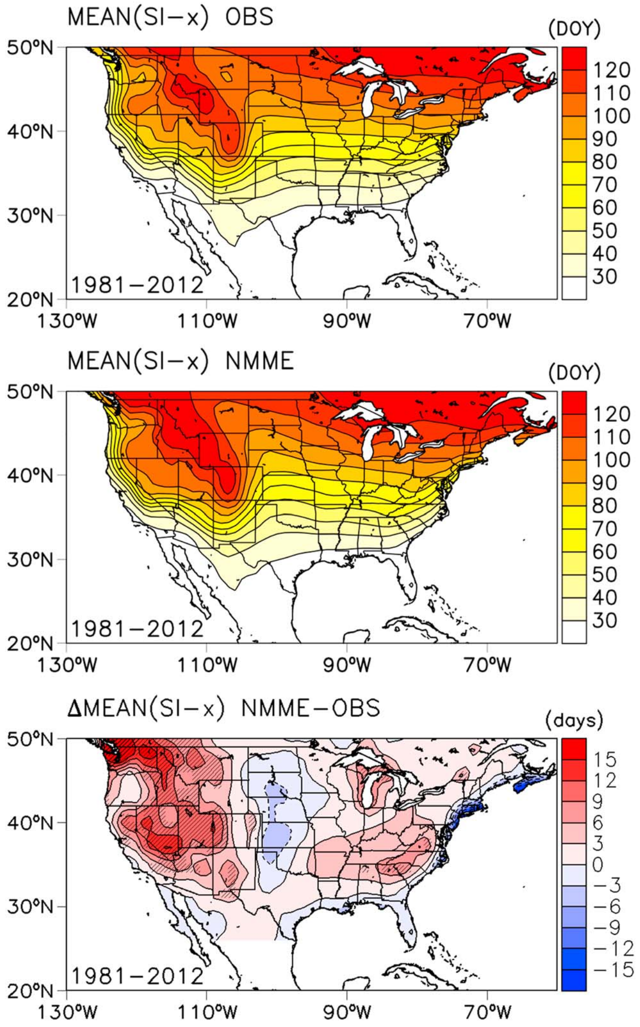

Figure 1. Mean day of the year (DOY) of the SI-x leaf index during the period 1981–2011 for the Berkeley Earth surface temperature (OBS; top), for the multimodel ensemble (middle) of the North America Multimodel Ensemble (NMME), and the NMME minus OBS difference (bottom). Oblique lines define regions with global statistically significant values (α_{GLOBAL} = 0.05) of this difference based on the false discovery rate (FDR) of multiple hypothesis tests (Wilks, 2016).

associate conditioned quantiles of simulated T_{max} values with observational ones. This procedure yields bias-corrected values of T_{min} and T_{max} for every grid point for every day of each year and preserves the covariance structure of T_{min} and T_{max} 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, multimodel ensemble forecasts are amenable to estimating forecast-uncertainty distributions, which can be used to calibrate these ensembles probabilistically. Here we use the nonhomogeneous Gaussian regression-ensemble model output statistic (NGR-EMOS) method (Gneiting et al., 2005) to post process the NMME direct model output forecast in order to improve SI-x forecast skill. Under this approach, the forecast-uncertainty distribution is assumed to be defined by a Gaussian distribution as indicated in equation (12), which describes the cumulative probability that a future observation V will be less than a forecast quantile q:

$$\Pr(V\leq q)=\Phi\left[\frac{q-\left(a+b_1*\overline{x}_{\mathrm{ens1}}+b_2*\overline{x}_{\mathrm{ens2}}+b_3*\overline{x}_{\mathrm{ens3}}+b_4*\overline{x}_{\mathrm{ens4}}+b_5*\overline{x}_{\mathrm{ens5}}\right)}{\left(c+d*s_{\mathrm{ens}}^2\right)^{\frac{1}{2}}}\right], \tag{12}$$

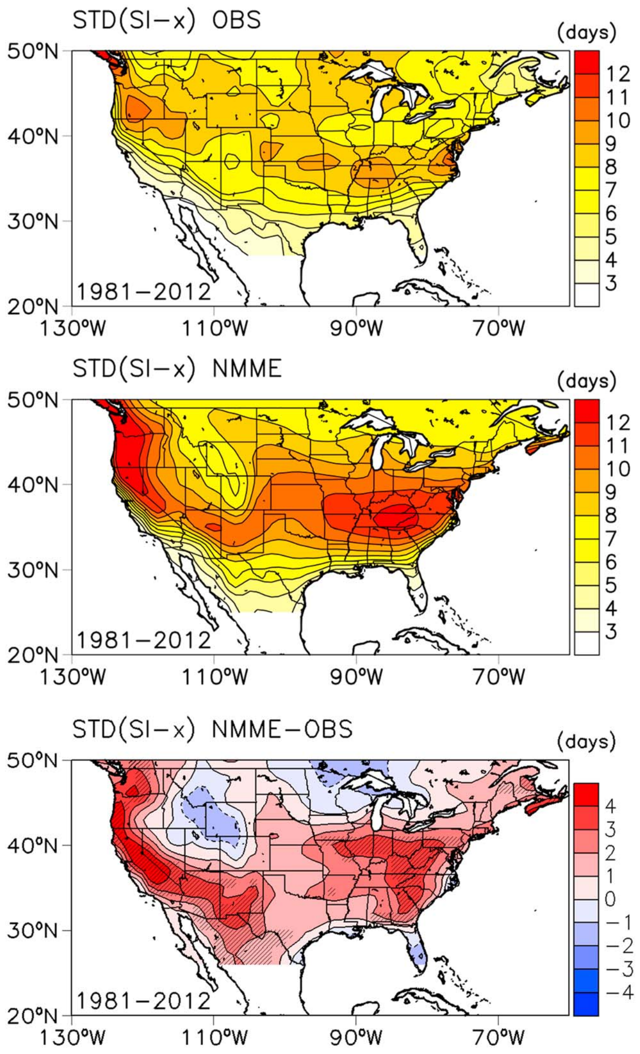

Figure 2. Standard deviation (STD) of the SI-x leaf index during the period 1981–2012 for the Berkeley Earth surface temperature data set (OBS; top), for the multimodel ensemble (middle) of the North America Multimodel Ensemble (NMME), and the NMME minus OBS difference (bottom). Oblique lines define regions with global statistically significant values (α_{GLOBAL} = 0.05) of this difference based on the false discovery rate (FDR) of multiple hypothesis tests (Wilks, 2016).

where Φ[ ] indicates the evaluation of the cumulative distribution function, $\overline{x}_{\text{ens}M}$ is the ensemble average of each model from Table 1, and s^2_{ens} is the ensemble variance. The parameters a, b_1, b_2, b_3, b_4, b_5, c, and d define the adjusted mean ($\mu = a + b_1 * \overline{x}_{\text{ens1}} + b_2 * \overline{x}_{\text{ens2}} + b_3 * \overline{x}_{\text{ens3}} + b_4 * \overline{x}_{\text{ens4}} + b_5 * \overline{x}_{\text{ens5}}$) and variance ($\sigma^2 = c + d * s^2_{\text{ens}}$), where $\overline{x}_{\text{ens}M} = \frac{1}{10}\sum_{i=1}^{10} x_{M,i}$ and $s^2_{\text{ens}} = \frac{1}{49}\sum_{i=1}^{50} (x_i - \overline{x})^2$ with $\overline{x} = \frac{1}{50}\sum_{i=1}^{50} x_i$. These parameters are used to generate the calibrated SI-x data. The mean is calculated with six parameters because our approach does not assume that the ensembles derived from the individual models are exchangeable. These eight parameters are estimated using a minimization of the average of CRPS (equation (8)) over the n training-period samples.

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), which training data are out of sample (observations to be forecasted are not included in the training data). Drawing for the fitting of these parameters is randomly selected from the maximum of year length and ensemble number without repetition.

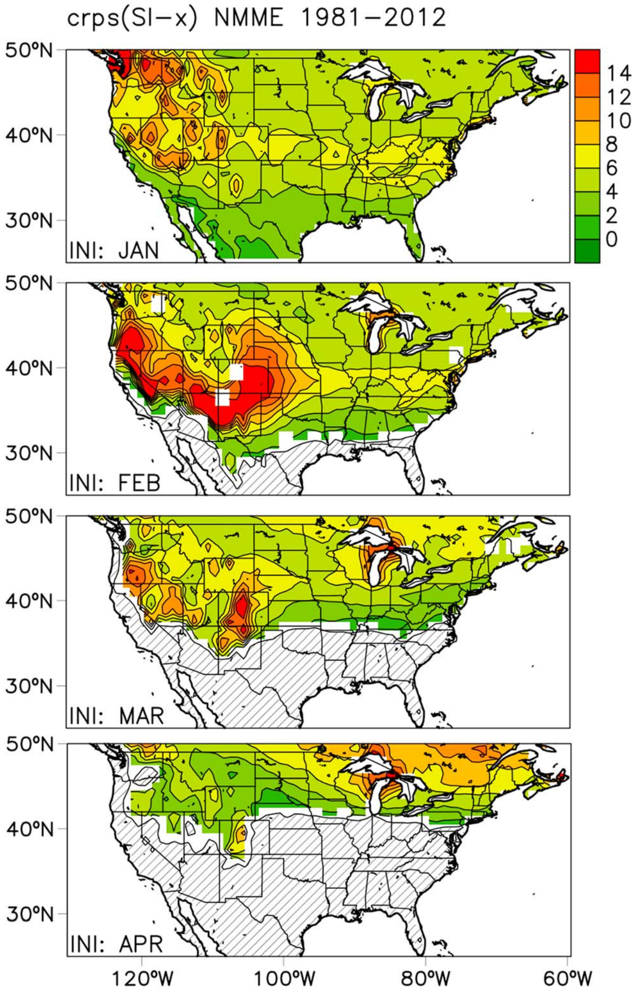

Figure 3. Spatial pattern of the ensemble continuous ranked probability score (eCRPS) of the SI-x leaf index. The eCRPS is computed using models from the North America Multimodel Experiment (NMME) without postprocessing treatment for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). A total of 50 ensemble members are used for each realization, and the period of analysis is from 1981 to 2012. High values of eCRPS indicate poor model performance. Oblique lines show regions where the spring index has already been reached leaf stage for each panel.

3. Results and Discussion

3.1. SI-x Climatology

Using mean DOY values, the SI-x leaf index computed from the NMME is comparable to the observational pattern (Figure 1). As in previous studies (e.g., Ault et al., 2015; Cayan et al., 2001), a prominent north-south gradient is present in both the observed data and the NMME ensemble mean. NMME mean and standard deviation are computed over the multimodel ensemble without distinguishing individual model-specific distributions. Greater spatial heterogeneity is observed along the western Intermountain Region, which is not fully reproduced by the NMME mean (bottom panel of Figure 1). The standard deviations of the SI-x in Figure 2 show less agreement between the observed pattern and model simulations. Although the simulations capture maximum interannual variance in the Pacific Northwest and in the southeastern United States, NMME overestimates this variability within a range of 4 days in the Intermountain Region, west of the Rockies (bottom panel of Figure 2). Thus, model biases are more apparent in the standard deviation than in the mean.

3.2. The SI-x SS

We first evaluate the skill of NMME models in predicting SI-x without any statistical correction. The ensemble CRPS, computed using equation (10), shows high values in the Intermountain Region (for January and February) and in the northeastern region of the domain (for March and April; Figure 3). High CRPS values

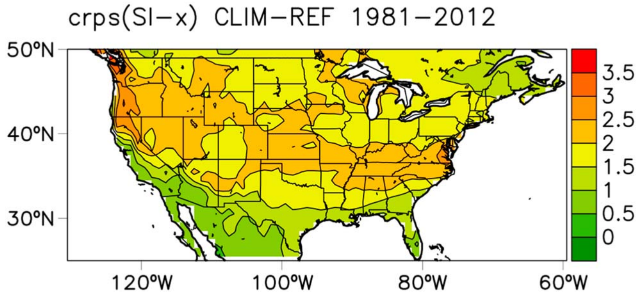

Figure 4. Spatial pattern of the continuous ranked probability score (CRPS) of the spring index (SI-x) for the observed leaf parameter, which is considered as the reference climatology (CLIM-REF) used in the computation of the skill score. The SI-x is computed for the period 1981–2012, using observed Berkeley Earth surface temperature.

are associated with poor model performance. CRPS values are low (indicating good skill) in low-elevation terrain. However, the raw ensemble forecasts do not outperform the CRPS climatology reference (CRPS [clim]; Figure 4), which maximum value is on 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 observations and NMME models (Figure 2). This result shows that the raw ensemble forecasts are of limited utility, as they exhibit negative skill with respect to the climatology (Figure S1).

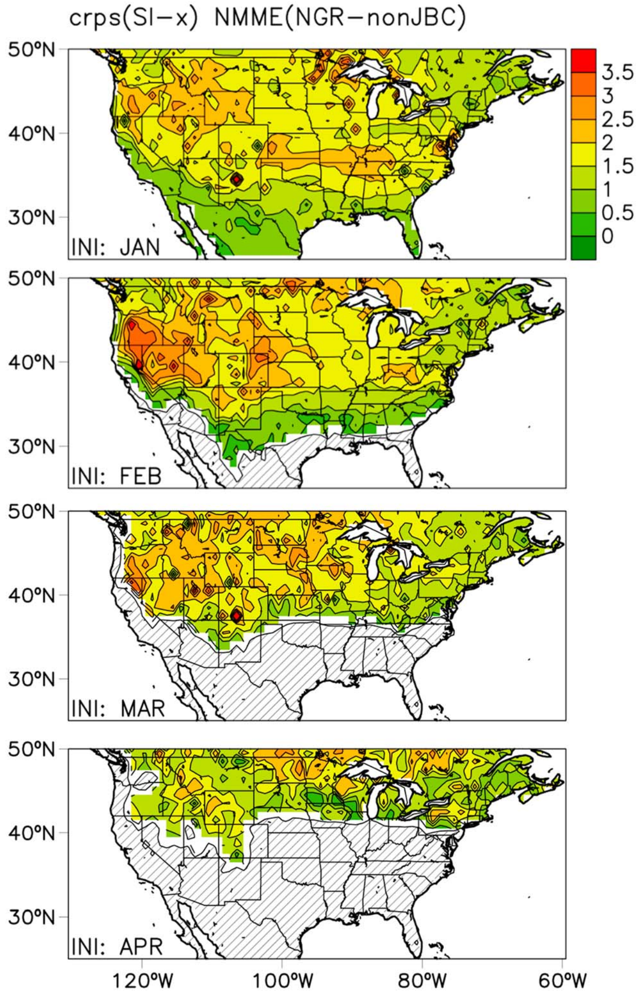

Figure 5. As in Figure 3 but continuous ranked probability score is computed after the nonhomogeneous Gaussian regression (NGR) is applied to the SI-x.

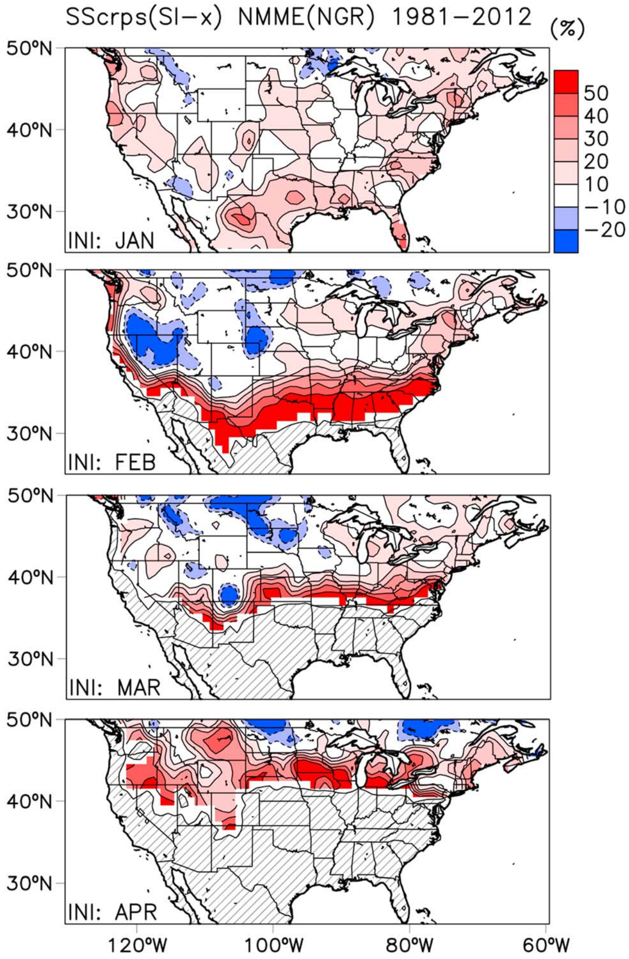

Figure 6. Spatial pattern of the skill score (SS) for the SI-x leaf index. The SS is computed using the continuous ranked probability score (CRPS) as measurement metric for the SI-x computed using models from the North America Multimodel Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate that improvement in percentage with respect to the reference climatology (Figure 4) and oblique lines show regions where the spring index is already reached for each panel.

The previous results did not include the NGR-EMOS approach, so we applied it to the SI-x output to evaluate the effects of correcting mean and dispersion errors. We illustrate an example of the multimodel ensemble error dispersion using one grid point for different model initialization times (Figure 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 NGR-EMOS should use this information in 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 (Figure 5) reveals an improvement with respect to the raw ensemble data (Figure 3) in all the different initializations. Although there is still a challenge to correct data in the Intermountain Region, the lower CRPS values show that the NGR-EMOS postprocessed forecasts are improved relative to the raw direct model output.

Figure 6 shows the SScrps for the entire set of forecasts, starting on 1 January, 1 February, 1 March, and 1 April for the period from 1981 to 2012. The CRPS [clim] used to compute SScrps is shown in Figure 4, and also an alternative SScrps field (when using the untreated NMME CRPS as reference) is shown in Figure 7 to illustrate the value added of the NGR-EMOS against the untreated data set. However, a comparison relative to the climatology is a fair metric and hence it is used here. Thus, in January (Figure 6), several regions with improvement of at least 10% are observed in the southeast, the northwest Pacific, the northeast, and the southwest. 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 southern states improve. This is noted by the

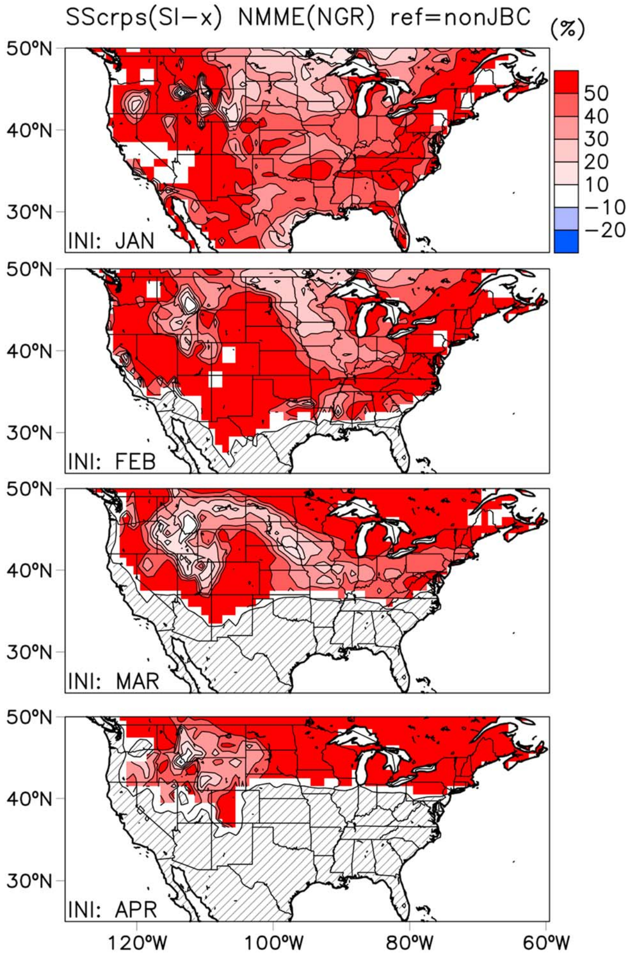

Figure 7. Spatial pattern of the skill score (SS) for the SI-x leaf index. The SS is computed using the continuous ranked probability score (CRPS) as measurement metric. The SI-x is computed using models from the North America Multimodel Experiment (NMME) and for four initializations corresponding to January (JAN), February (FEB), March (MAR), and April (APR). Positive values indicate that improvement in percentage with respect to the untreated CRPS (Figure 3) and oblique lines show regions where the spring index is already reached for each panel.

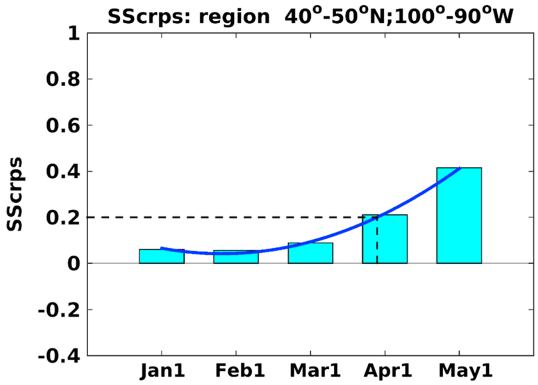

Figure 8. The continuous ranked probability skill score (SScrps) versus time plot for initializations on 1 January, 1 February, 1 March, 1 April, and 1 May. The SScrps is calculated for an average region in the Great Plains (100°W–90°W, 40°N–50°N). The solid line is the fitted curve with a second-order polynomial, SScrps = $0.0363\,x^2 - 0.1310\,x + 0.1599$, and the dashed lines highlight the value for the SScrps = 0.2 and $x = 3.89$ in months unit (day of year [DOY] = 86). The average observed SI-x for this region is 104 DOY.

Table 2
SI-x Predictability Ranges

Model	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	noJBC	JBC
CanCM3	18.8	30.7	23.3	27.6	21.8	26.4	26.4	23.8	22.0	19.3	24.0	24.8
CanCM4	23.2	27.4	22.6	27.2	24.1	23.0	24.6	25.2	25.8	25.0	24.8	27.3
CESM1	13.8	14.0	14.8	17.1	19.6	13.1	17.2	12.2	12.9	19.3	15.4	23.6
FLORB01	24.7	25.3	30.9	30.0	14.3	16.1	11.3	22.8	13.0	10.9	19.9	19.5
GEOS-5	18.2	18.9	12.8	19.1	25.2	21.4	22.6	22.8	17.7	22.2	20.1	23.1
MMEM											20.9	23.7

Note. SI-x predictability ranges (in days) average for a region over the U.S. Great Plains (100°W–90°W, 40°N–50°N) using mean square error (MSE) as metric to compute the skill score (SS) as in equation (2). These values are calculated from forecasted SI-x with untreated temperature data sets (noJBC for non-joint bias correction) for each individual model (CanCM3, CanCM4, FLORB01, and GEOS-5) and individual ensemble members (E1, …, E10). The next column shows the ensemble average for each model (noJBC). The last column is the same as in the *noJBC* column but after temperature data sets to compute SI-x were treated with the joint bias correction (JBC) approach. The multimodel ensemble mean (MMEM) is added at the bottom for both approaches.

region with 30–50% of positive change (Figure 6), which describes the improvement by NGR-EMOS. The major feature in February is the percentage of negative skill, on the order of 20%, in the Intermountain Region, as can be inferred from the analysis of the standard deviation anomaly (Figure 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 (Figure 2), near 85°W, 35°N. This suggests that NGR-EMOS is able to add value by enhancing good SI-x individual forecast members in the NMME.

3.3. The SI-x Predictability Range

The SScrps evaluates the forecast skill of SI-x as a percentage of the reference climatology (Figure 6). However, to determine how many days in advance SI-x computed from NMME forecasts can estimate spring onset, the *time dependence* of SSclim needs to be characterized. Figure 8 shows how this additional metric is constructed for a region in the Great Plains (100°W–90°W, 40°N–50°N). First, the SScrps values for every initialization (1 January through 1 May) are calculated using the climatological reference (Figure 4). 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. In Figure 8, the solid line is the fitted curve with a second-order polynomial, $\mathrm{SScrps} = 0.0363\,x^2 – 0.1310\,x + 0.1599$, with x in months units. Thus, using SScrps = 0.20, we estimate $x = 3.89$ months or 86 days (DOY) in the fitted SScrps time variation (dashed line). In this example, a SI-x predictability range of about 20 days is obtained according to the fitted SScrps versus time relationship, as climatology for the region is 104 DOY. Values for the SI-x predictability range of the same order are obtained with the alternative SSclim metric (Table 2).

The observed 20-day forecast skill is in the range of a model such as the Climate Forecast System version 2 reported by Saha et al. (2014), and it might justify the use of SS equal to 20% as a meaningful benchmark. This 20-day predictability range is the one that Climate Services would potentially use as information that includes a level of improvement of 20% (SScrps = 0.20) with respect to climatology as forecast. 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, and this information can be used to weight a final product or to eliminate some models in an optimal operational forecast. For example, this weighting is objectively achieved by the b parameters in equation (12), which indicates the contribution of the five models to form the best postprocessed SI-x forecast, with high values defining the best models and near-zero values suggesting less useful models (Table 3).

Table 3
Regression Values of the Parameters b_i

Initialization	b_1	b_2	b_3	b_4	b_5
January	0.39	0.7	0.06	−0.01	−0.11
February	0.19	0.11	0.58	−0.03	0.14
March	0.27	0.26	0.36	0.28	−0.18
April	−0.01	0.65	0.15	−0.09	0.26
	CanCM3	CanCM4	CESM1	FLORB01	GEOS-5

Note. Regression values of the parameters b_i (i = 1, 2, …, 5) in the probabilistic nonhomogeneous Gaussian regression equation (12). The parameters are described for each initialization (January, …, April) and for the five NMME models: CanCM3 (b_1), CanCM4 (b_2), CESM2 (b_3), FLORB1 (b_4), and GEOS-5 ($b5$). The parameter values are computed over an average region in the U.S. Great Plains (100°W–90°W, 40°N–50°N).

The SI-x predictability for the continental United States is in the range of 10–60 days for the NGR-EMOS NMME (Figure 9a). 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 (Figure 9b) and 0.2 (Figure 9c), 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 (Figure S3). 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 relatively 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 (Figure 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.

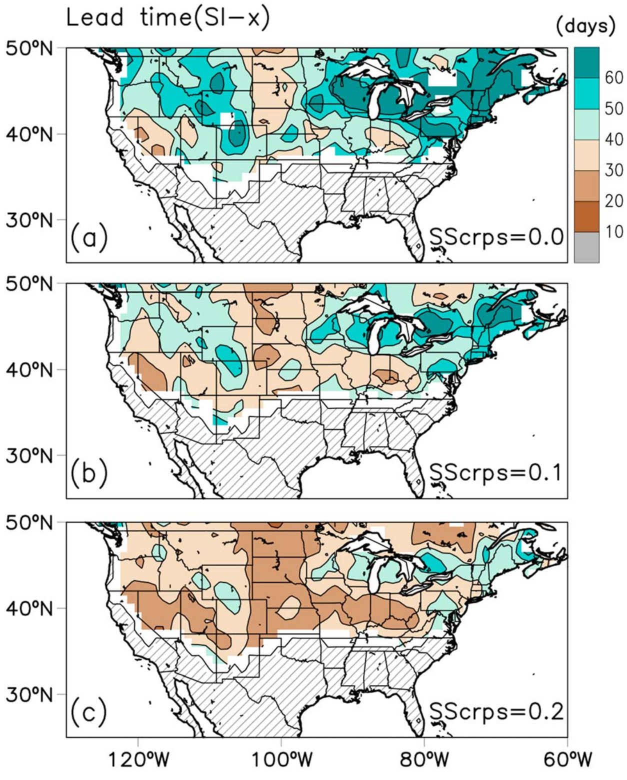

Figure 9. The SI-x leaf index predictability range measured in days before the occurrence of the spring onset. The SI-x leaf index predictability range is computed as in Figure 6 and for three skill score (SScrps) thresholds: 0.0 (a), 0.1 (b), and 0.2 (c). Oblique lines indicating regions where the calculation of the metric is not possible.

The multimodel ensemble NMME NGR-EMOS (Figure 9) agrees well with the individual model ensembles (Figure S3), 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 Goddard Earth Observing System Model, Version 5 model shows the lowest range of predictability in the Intermountain Region, it shows better improvement after applying the TT-JBC, which is also true for CanCM3 and CanCM4 near 45°N. The results with the TT-JBC are consistent with the biased temperature (Figure S3; 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 National Oceanic and Atmospheric Administration (Goddard Earth Observing System Model, Version 5) models. As abrupt warming events in the SI-x calculation are modeled with daily maximum and minimum temperature (Schwartz & Marotz, 1986), 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 (Figure 6).

Using a multimodel ensemble NGR-EMOS (Figure 9), the results for the five models can be summarized in two major points. First, 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 multimodel ensemble outperforms 2 months before the beginning of the spring. These changes are localized in two regions: the "corn

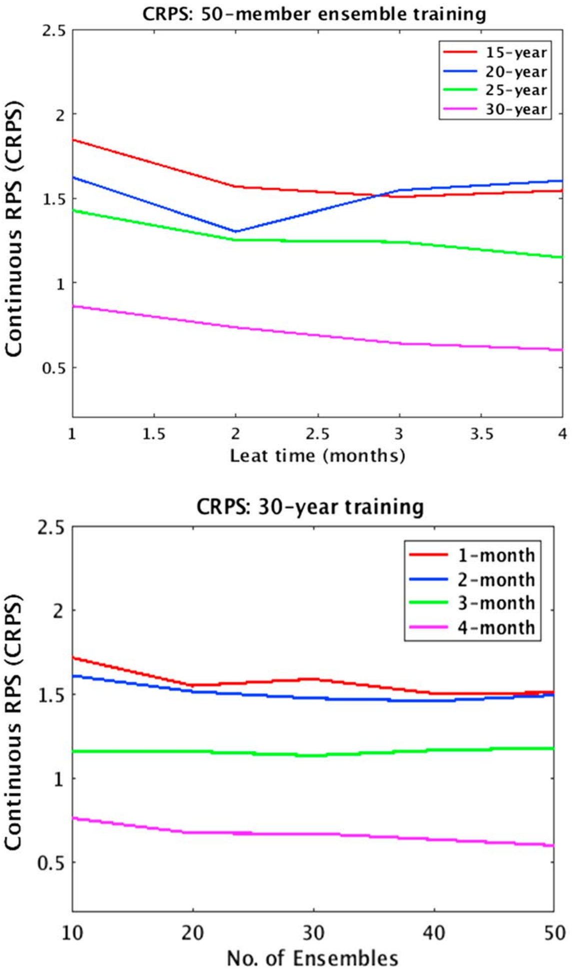

Figure 10. (top) The continuous ranked probability score (CRPS) for the 50 ensemble model realizations for different training periods (15, 20, 25, and 30 years) and lead time (*x* axis: 1, 2, 3, and 4 months). Low CRPS values represent better score. (bottom) Similar as Figure 10 (top) but fixing the 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).

belt" along 40°N (Nebraska, Iowa, Minnesota, and Illinois) and the Intermountain Region. Second, 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 (Figure 10), the CRPS shows two important aspects to consider when applying the NGR-EMOS in SI-x related products: (1) a long training period significantly increases the predictability score (e.g., from 15 to 30 years; top panel Figure 10) and (2) a large number of ensemble members marginally improves the RPS SS (e.g., from 10 to 20 members; bottom panel Figure 10). 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 members to 1.5 when using 20 ensemble members. 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 SS, after using the NGR-EMOS, reveals a significant improvement in the Corn Belt region (Figure 6). It portrays the positive effect of NGR-EMOS for the four initializations (January–April) using the SScrps SS. When we compare the difference between the model ensemble NMME mean and NGR-EMOS, the percentage of improvement is of the order of 50 percentage points (from 10% to 80% SS) and the extension of this improvement expands significantly relative to the untreated results. For example, in February and March, the Corn Belt region sees an important improvement, which is verified with the similar results obtained by two other EMOS methods: logistic regression and Gaussian ensemble dressing (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, and references therein). A set of NMME models was treated with a daily JBC approach and an ensemble model output statistics approach. Our findings show that untreated input data are of limited use, as it exhibits negative skill relative to climatology. 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 improve the SI-x predictability SS. 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 to 60 days located along a narrow band between 35°N and 45°N. Using a forecast threshold of SScrps = 0.0, the range of predictability falls into two categories: 10–40 and 40–60 days. Using higher thresholds (SScrps = 0.1 and 0.2), predictability shows a lower range with values around 10–30 days (Figure 9). The 40–60 day time horizon is notable, as it extends well beyond the 10-day 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 time scales, 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 early and late spring variability is significant (Shubert et al., 2016). Becker et al. (2014) also show that NMME has good results in the central United States, which further supports our interpretation, although the regions with better skill found in this study are narrow and localized.

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 of its impact in western North America. Indices have been constructed that characterize the position, structure, and strength of the Pacific Jetstream (Newman & Sardeshmukh, 1998) as it migrates north, splits, and weakens each spring. 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 supports the existence of driving mechanisms at this scale that might be orchestrating these ranges of predictability skill.

Finally, our findings suggest that there is potential spring onset forecast skill in NMME products, but a sophisticated postprocessing is necessary to achieve that potential. We delineate how the predictability skill of NMME models to forecast spring onset in North America is improved with two postprocessing techniques—the JBC and nonhomogeneous Gaussian regression EMOS. The JBC outperforms the biased temperature SI-x product, and the improvement mainly occurs in both the Canadian and National Oceanic and Atmospheric Administration models; however, it does not outperform the multimodel ensemble NGR-EMOS. Using NGR-EMOS, a significant positive change in the SS is noted in regions where the skill of the raw NMME ensemble 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 operational use.

Acknowledgments
This material is based on work supported partially by USDA grant 1010630 (project NYC-124439); additional funding was provided by NSF grant 1702697. The authors thank the North American Multimodel Ensemble (NMME) project for providing the data set. NMME project is supported by NOAA, NFS, NASA, and DOE. NMME data were obtained from http://www.cpc.noaa.gov/products/NMME/data.html.%20 Observed temperature was obtained from http://www.BerkleyEarth.lbl.org/data/. Code to compute the SI-x is available at https://github.com/cornell-eas/SI-X.

References

Rohde, R., Muller, A. M., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., et al. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinformatics and Geostatistics: An Overview*, *1*, 1–7. https://doi.org/10.4172/gigs.1000101

Ault, R. T., Henebry, G. M., de Beurs, K. M., Schwartz, M. D., Betancourt, J. L., & Moore, D. (2013). The false spring of 2012, earliest in North American record. *Eos*, *94*(20), 181–182. https://doi.org/10.1002/2013EO200001

Ault, R. T., Macalady, A. K., Pederson, G. T., Betancourt, J. L., & Schwart, M. D. (2011). Northern Hemisphere modes of variability and the timing of spring in western North America. *Journal of Climate*, *224*, 4003–4014.

Ault, R. T., Zurita-Milla, R., & Schwartz, M. D. (2015). A Matlab toolbox for calculating spring indices from daily meteorological data. *Computers and Geosciences*, *83*, 46–53. https://doi.org/10.1016/j.cageo.2015.06.015

Becker, E., van den Dool, H., & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate*, *27*(15), 5891–5906. https://doi.org/10.1175/JCLI-D-13-00597.1

Cayan, D. R., Kammerdiener, S. A., Dettinger, M. D., Caprio, J. M., & Peterson, D. H. (2001). Changes in the onset of spring in the western United States. *Bulletin of the American Meteorological Society*, *82*(3), 399–415. https://doi.org/10.1175/1520-0477(2001)082%3C0399:CITOOS%3E2.3.CO;2

Ferro, C. A. T., Richardson, D. S., & Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, *15*(1), 19–24. https://doi.org/10.1002/met.45

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L. III, Paolino, D. A., Zhang, Q., et al. (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

Koster, R. D., Mahanama, S. P. P., Yamada, T. J., Balsamo, G., Berg, A. A., Boisserie, M., et al. (2011). The second phase of the global land-atmosphere coupling experiment: Soil moisture contributions to subseasonal forecast skill. *Journal of Hydroclimatology*, *12*, 805–822.

Li, C., Sinha, E., Horton, D. E., Diffenbaugh, N. S., & Michalak, A. M. (2014). Joint bias correction of temperature and precipitation in climate model simulations. *Journal of Geophysical Research: Atmospheres*, *119*, 13,153–13,162. https://doi.org/10.1002/2014JD022514

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Mo, K. C., & Lettenmaier, D. P. (2014). Hydrologic prediction over the conterminous United States using the National Multi-Model Ensemble. *Journal of Hydrometeorology*, *15*(4), 1457–1472. https://doi.org/10.1175/JHM-D-13-0197.1

Monahan, W. B., Rosemartin, A., Gerst, K. L., Fisichelli, N. A., Ault, T., Schwartz, M. D., et al. (2016). Climate change is advancing spring onset across the U.S. national park system. *Ecosphere*, *7*(10), 1–17. https://doi.org/10.1002/ecs2.1465

Newman, M., & Sardeshmukh, P. D. (1998). The impact of the annual cycle on the North Pacific/North American response to remote low-frequency forcing. *Journal of the Atmospheric Sciences*, *55*, 1336–1353.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Schwartz, M. D., Ahas, R., & Aasa, A. (2006). Onset of spring starting earlier across the Northern Hemisphere. *Global Change Biology*, *12*(2), 343–351. https://doi.org/10.1111/j.1365-2486.2005.01097.x

Schwartz, M. D., Ault, T. R., & Betancourt, J. L. (2013). Spring onset variations and trends in the continental United States: Past and regional assessment using temperature-based indices. *International Journal of Climatology*, *33*(13), 2917–2922. https://doi.org/10.1002/joc.3625

Schwartz, M. D., & Marotz, G. A. (1986). An approach to examining regional atmosphere plant interactions with phenological data. *Journal of Biogeography*, *13*(6), 551–161. https://doi.org/10.2307/2844818

Shubert, S., Chang, Y., Wang, H., Koster, R., & Suarez, M. (2016). A modeling study of the causes and predictability of the spring 2011 extreme U.S. weather activity. *Journal of Climate*, *29*(21), 7869–7887. https://doi.org/10.1175/JCLI-D-15-0673.1

Thrasher, B. E., Maurer, E. P., McKellar, C., & Duffy, P. V. (2012). Technical note: Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrology and Earth System Sciences*, *16*(9), 3309–3314. https://doi.org/10.5194/hess-16-3309-2012

Van Schaeybroeck, B., & Vannitsem, S. (2015). Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 807–818. https://doi.org/10.1002/qj.2397

Westerling, A. L., Hidalgo, H. G., Cayan, D. R., & Swetnam, T. W. (2006). Warming and earlier spring increase western U.S. forest wildfire activity. *Science*, *313*(5789), 940–943. https://doi.org/10.1126/science.1128834

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed., p. 676). Academic Press.

Wilks, D. S. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1.