Toward the Application of Decadal Climate Predictions

ERIN TOWLER, DEBASISH PAIMAZUMDER, AND JAMES DONE

Capacity Center for Climate and Weather Extremes, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 27 April 2017, in final form 8 January 2018)

ABSTRACT

Decadal prediction is a relatively new branch of climate science that bridges the gap between seasonal climate forecasts and multidecadal-to-century projections of climate change. This paper develops a three-step framework toward the potential application of decadal temperature predictions using the Community Climate System Model, version 4 (CCSM4). In step 1, the predictions are evaluated and it is found that the temperature hindcasts show skill over some regions of the United States and Canada. In step 2, the predictions are manipulated using two methods: a deterministic-anomaly approach (like climate change projections) and a probabilistic tercile-based approach (like seasonal forecasts). In step 3, the predictions are translated by adding a delta (for the anomaly manipulation) and conducting a weighted resample (for the probabilistic manipulation), as well as using a new hybrid method. Using the 2010 initialized hindcast, the framework is demonstrated for predicting 2011–15 over two case-study watersheds [Ottawa (Canada) and Colorado]. For the Colorado watershed, there was a noticeable shift toward higher temperatures, and the delta, weighted resample, and hybrid translations all were better at capturing the observed temperatures than was an approach that used climatological values. For the Ottawa watershed, the observed temperatures over the period of prediction were only subtly different than the climatological values; therefore, the difference between the translation methods was less noticeable. The advantages and disadvantages of the manipulation and translation approaches are discussed, as well as how their use will depend on the user context. The authors emphasize that skill evaluations should be tailored to particular applications and identify additional steps that are needed before the decadal temperature predictions can be readily incorporated into applications.

1. Introduction

Decadal climate prediction is an evolving branch of climate science that fills the gap between seasonal climate forecasts and multidecadal-to-century projections of climate change. For seasonal climate forecasts, climate models are initialized to observation-based current conditions and are run out from months to a year. For multidecadal-to-century projections of climate change, climate models are run from a randomly selected preindustrial state and extended into the future using an external forcing, such as a greenhouse gas emission scenario (e.g., IPCC representative concentration pathways RCP4.5 or RCP8.5). Decadal predictions typically extend out between 5 and 30 years. Like seasonal forecasts, decadal predictions are initialized to observation-based current conditions, whereas, like climate projections, they are extended into the future using a scenario of future external forcing. The hybrid nature of decadal predictions can be an asset, because they can benefit from experience and lessons learned from predictions made at other time scales. This hybrid nature also makes them distinct from seasonal forecasts and uninitialized climate projections and thus worthy of new study.

Decadal predictions have many features in common with seasonal climate forecasts (Goddard et al. 2012a,b). In the United States, seasonal climate forecasts have been routinely issued by the NOAA Climate Prediction Center (CPC) and the International Research Institute for Climate and Society (IRI) at Columbia University since the 1990s. Seasonal climate forecasts are expressed probabilistically (Mason and Goddard 2001) using a...
tercile-based approach in which the likelihoods of being in the above-normal, normal, or below-normal tercile are forecast. Such forecasts commonly have a climatological base period of 1981–2010. Seasonal forecasts can provide a test bed for decadal predictions (Goddard et al. 2012a); for instance, adoption of seasonal forecasts can help to build trust, because seasonal-forecast performance can be evaluated over the recent past and over the near future (Goddard et al. 2012a). Use of seasonal forecasts can also increase the uptake of decadal predictions, because their use will add capacity for integrating climate information on longer time scales (Goddard et al. 2012a).

Decadal predictions also have many features in common with multidecadal and century projections of climate change. Climate-model projections have been widely disseminated in reports such as those from the Intergovernmental Panel on Climate Change (IPCC; Collins et al. 2013) and the National Climate Assessment (NCA; Walsh et al. 2014). In reports such as those from the IPCC and NCA, multidecadal and century projections of climate change are typically shown as relative deterministic changes, that is, deltas for temperature, from a 20- or 30-yr baseline period. In a similar way, experimental multimodel decadal predictions are now issued annually (Smith et al. 2013; Met Office 2016), and these predictions also present the 1- and 5-yr outlooks as deltas from a baseline period (1971–2000).

There has been a demand for decadal predictions because there is a societal need for time-evolving predictions on this time scale (Vera et al. 2010; Barsugli et al. 2009). To identify their climate sensitivities, planners often use impact models to translate climate information into parameters that are relevant to their system (Raucher et al. 2015). For instance, many water utilities are already considering seasonal and long-term climate information in their planning (Raucher et al. 2015) and have identified improved climate information on decadal time scales as a desired need for their planning horizons (Barsugli et al. 2009). This is true even for variables such as temperature, for which the increasing signal is robust but for which water utilities seek to narrow the range of uncertainty in the near term (Barsugli et al. 2009).

A recent survey found that potential users prefer familiar formats for presenting the decadal predictions (Taylor et al. 2015). Further, potential users perceive decadal information to be useful but need assistance in processing and interpreting the predictions, especially the uncertainty (Taylor et al. 2015). Decadal climate predictions are still experimental, and several scientific and technical challenges need to be resolved before they are ready to be used operationally. Climate centers around the globe are working to improve the predictions, and real-time multimodel decadal predictions are being issued annually as a community resource (Smith et al. 2013; Met Office 2016). This creates a research opportunity to understand their potential role for users and impact modelers. To this end, the purpose of this paper is to explore how the decadal temperature predictions could be applied by potential users. We achieve this by developing a three-step framework that is demonstrated using the decadal temperature predictions from the National Center for Atmospheric Research (NCAR) Community Climate System Model, version 4 (CCSM4).

In step 1, we evaluate the CCSM4 temperature predictions. Although the purpose of this paper is not to evaluate skill per se, which has been done more comprehensively elsewhere (e.g., Meehl et al. 2014; Kirtman et al. 2013, Kim et al. 2012), we perform diagnostics on the CCSM4 decadal temperature predictions to give users some familiarity with the skill scores, with an emphasis on how they can be tailored for a particular application. In step 2, we manipulate the temperature predictions using 1) a deterministic-anomaly approach (as in projections of climate change) and 2) a probabilistic tercile-based approach (as with seasonal forecasts). In step 3, for both of the manipulation approaches from step 2, we demonstrate how the temperature predictions could be translated for application by adding a delta (for the deterministic-anomaly manipulation), conducting a weighted resample (for the probabilistic tercile-based manipulation), or using a new hybrid approach. To do this, we examine two case-study watersheds.

The paper is organized as follows. In section 2, we provide background on decadal climate prediction science, with a focus on what is known about the skill. In section 3, we introduce the decadal predictions and observational data used in the study. In section 4, we provide details on the three-step framework including evaluating the predictions (step 1), manipulating the predictions (step 2), and translating the predictions (step 3). Section 5 presents the results. Section 6 features the discussion and our conclusions, and in it we consider the pros and cons of the different approaches, as well as additional steps needed before the application of decadal predictions can be realized in practice.

2. State of the science

The most recent Coupled Model Intercomparison Project (phase 5; CMIP5) included decadal hindcast and prediction experiments (Taylor et al. 2012). The hindcasts are a historical reforecast and, thus, allow for evaluation of the skill. As such, as referred to up to this
point and from this point forward, the term “decadal predictions” will be used to refer generally to the decadal climate data. Note that in our analysis, or in any skill quantification, the term “decadal hindcasts” will be used to clarify that it only includes historical reforecasts that can be evaluated for skill.

Decadal climate predictions have two potential sources for skill: 1) the initialization and 2) the external forcing. In terms of the former, Meehl et al. (2009) identify several phenomena that could provide skill on decadal time scales, both in the Pacific Ocean [e.g., the Pacific decadal oscillation (PDO), the North Pacific index, and the interdecadal Pacific oscillation (IPO)] and the Atlantic Ocean [e.g., the Atlantic meridional overturning circulations, the North Atlantic Oscillation, and the Atlantic multidecadal oscillation (AMO)]. This has parallels with seasonal climate forecasts, which derive much of their skill from the El Niño–Southern Oscillation (ENSO) phenomenon. Just as stakeholders whose regions are impacted by ENSO have more potential to benefit from seasonal forecasts, the potential skill from decadal phenomena will also be regionally dependent. For instance, in the United States, Dong and Dai (2015) found the IPO to be associated with precipitation in the southwestern United States and McCabe et al. (2004) found both the AMO and PDO to be associated with U.S. drought. In terms of the latter source of skill, that is, skill derived from external forcing, decadal predictions have built-in skill from climate change commitment (i.e., greenhouse gases that have already been emitted) and the forcing from increasing greenhouse gases (i.e., the greenhouse gases estimated to be emitted over the prediction period) (Meehl et al. 2009; Lee et al. 2006). As such, a fundamental difference between the uninitialized predictions (such as projections of climate change) and initialized predictions (such as decadal) is that uninitialized predictions only aim to predict the forced response to greenhouse gases, whereas the initialized predictions aim to predict both the forced response and the natural variability.

To quantify the skill of the decadal temperature hindcasts, several metrics have been used (e.g., Goddard et al. 2012b; Saha et al. 2006). Anomaly correlation skill score (ACC) shows widespread skill of surface air temperature using the CMIP5 multimodel decadal hindcasts (Meehl et al. 2014), although the CMIP3/CMIP5 multimodel uninitialized projections had already shown widespread predictive skill of temperature (Meehl et al. 2009, 2014). The mean-square skill score (MSSS) can be used to quantify the skill added from initialization, which it does for temperature in a few regions, notably the North Atlantic Ocean (Meehl et al. 2014). The MSSS can also be used to compare decadal hindcast skill with use of climatological values (hereinafter referred to as “climatology”), that is, to identify areas where it would have been better to use the hindcast than climatology. The continuous rank probability skill score (CRPSS) is used to understand whether the prediction ensemble captures the observed uncertainty on average. Lack of skill has been shown using the CRPSS (Goddard et al. 2012b), indicating that more decadal prediction ensembles, such as making a decadal prediction with 40 members (rather than 10), similar to the large ensemble in Kay et al. (2015), are needed to adequately capture the uncertainty. Precipitation shows less skill than temperature (Meehl et al. 2014) and is not considered in this study.

In short, decadal predictions are still considered to be exploratory (Goddard et al. 2012a; Taylor et al. 2012). Kirtman et al. (2013) identifies the three remaining technical challenges to decadal prediction as 1) limited availability of observations, 2) limitations of dynamical modeling, and 3) a need to assess the methods used to initialize the models (i.e., the data-assimilation techniques). In terms of the third point, there are two main initialization methods: “full field” and “anomaly” initialization (CMIP–WGCM–WGSIP Decadal Climate Prediction Panel 2011). Both still have technical challenges (Goddard et al. 2012b) and are an area of active research. Researchers are currently favoring the full-field initialization method (Meehl et al. 2014), which keeps the model’s initial values close to the actual observed values, but it requires an adjustment for bias in the predictions because the model will inevitably drift toward its preferred climatology. As such, the decadal predictions need to be adjusted for bias, which in decadal predictions is called drift correction (Meehl et al. 2014). Further, the drift-correction method required is different and more complicated than the bias-correction approach that is typically used for projections of climate change (Meehl et al. 2014).

3. Data

a. Decadal temperature hindcasts

We use decadal temperature hindcast data from CCSM4 that were initialized on 1 January every year from 1980 to 2010 and run for 10 years of simulation. Model runs were initialized using the full-field initialization method; details of the initialization of the CCSM4 ocean component can be found in Karspeck et al. (2013). These CCSM4 decadal prediction hindcast runs have been used in several other research efforts (Meehl and Teng 2014a,b; Yeager et al. 2015; Teng et al. 2017) and were internally accessed at NCAR. CCSM4 is a state-of-the-art coupled model, consisting of an

Unauthenticated | Downloaded 09/30/22 11:47 AM UTC
atmospheric model and components of ocean, land, and sea ice (Gent et al. 2011). Several studies suggest that, because of the substantial influence of internal climate variability on climate trajectories, single realizations from climate models are often insufficient for model comparison with the observations and uncertainty communication (Deser et al. 2014; Kay et al. 2015). Thus, a 10-member ensemble is used in this study; members were generated by perturbing the initial conditions.

Here, we examine 2-m temperature over North America. As mentioned previously, the decadal predictions need to be drift corrected, which was done as in other studies (e.g., Meehl and Teng 2012; Meehl and Teng 2014a, b) in accordance with the “CLIVAR” protocol (CMIP–WGCM–WGSIP Decadal Climate Prediction Panel 2011). Following from Meehl and Teng (2014a, b), the drift correction for full-field initialization is calculated as

$$\hat{Y}_j = Y_j - \frac{1}{N-1} \sum_{k=1, k \neq j}^N (Y_{kt} - O_{kt}),$$

where $\hat{Y}_j$ and $Y_j$ are the drift-corrected and raw values for hindcast $j$ at lead year $t$. In addition, $O_{kt}$ corresponds to the observed values at $Y_j$ and $N$ is the number of hindcasts (initialization years) that can be used. The drift correction is calculated in a cross-validated manner, so that the hindcast being corrected does not contribute to the average model drift being removed.

b. Observational temperature data

For the observations, we use the Daymet dataset, version 3 (V3; Thornton et al. 1997), which provides 1 km × 1 km gridded estimates of daily weather parameters for the United States, including minimum temperature Tmin and maximum temperature Tmax. As defined in Thornton et al. (1997), we calculate daily average temperature as $0.4 \times T_{\text{min}} + 0.6 \times T_{\text{max}}$.

4. Methods

We develop a three-step framework toward the potential application of decadal temperature predictions. As shown in the flowchart of Fig. 1, the three steps are 1) evaluate predictions, 2) manipulate predictions, and 3) translate predictions; these three steps are detailed in the following sections.

a. Step 1: Evaluate predictions

To give users a sense of the skill of the CCSM4 temperature hindcasts, we examine two measures: the ACC and MSSS. We emphasize that, for a particular application, the skill assessment will be more salient if it is tailored to particular user needs. In this paper, we examine two case-study watersheds. Thus, we tailor the skill score by evaluating water years, which are defined to be from 1 October to 1 September, rather than calendar years. This is because water years are more salient to watershed hydrology in the midlatitudes. In addition, even though the decadal predictions are run out for 10 years, the particular prediction period of interest will depend on the user; here we look out 5 years, because that period is when the initialization is the most influential (Boer et al. 2016).

To calculate the ACC and MSSS, we examine decadal hindcast runs that were initialized every year from 1980 to 2009, resulting in a sample size of 30 runs. For each initialization year, we examine the 5-yr average over months 10–69 (we start at month 10 because the water year starts on 1 October), resulting in 30 five-year averages. This is repeated for all 10 ensemble members.
First, we examine the ACC; the ACC is primarily used to evaluate skill because it is inherently insensitive to mean bias corrections. From Saha et al. (2006), the ACC is calculated as

$$\text{ACC} = \frac{\frac{1}{n} \sum_{i=1}^{n} H_i O_i}{\left[\frac{1}{n} \sum_{i=1}^{n} H_i H_i \left(\frac{1}{n} \sum_{i=1}^{n} O_i O_i\right)\right]^{1/2}},$$

where $H_i$ is the ensemble-average hindcast anomaly, $O_i$ is the observed anomaly, and $n$ is 30 initialization years. Anomalies are calculated using a 30-yr baseline period from October 1980 to September 2010, henceforth referred to as “1981–2010.” Although spatial smoothing can help to show regional skill, many impact users ingest model information at the grid scale (Goddard et al. 2012b), and therefore that is what is done here. The observed data are interpolated to the model resolution, and calculations are done at the grid scale of the model, which is 100 km × 100 km.

The MSSS combines the square of the correlation coefficient and the square of the conditional prediction bias (Goddard et al. 2012b). Here, the MSSS is used to compare the decadal hindcast skill with climatology; positive MSSS indicates that the hindcasts have lower mean-square error (MSE) than a forecast of climatological bias (Goddard et al. 2012b), and therefore that is what is done here. The MSSS is calculated as

$$\text{MSSS} = \frac{\sum_{i=1}^{n} (H_i - O_i)^2}{\text{MSE}_{\text{ref}}},$$

where $H_i$ is the ensemble-average hindcast anomaly, $O_i$ is the observed anomaly, and $n$ is 30 initialization years. Anomalies are calculated using a 30-yr baseline period from October 1980 to September 2010, henceforth referred to as “1981–2010.” Although spatial smoothing can help to show regional skill, many impact users ingest model information at the grid scale (Goddard et al. 2012b), and therefore that is what is done here. The observed data are interpolated to the model resolution, and calculations are done at the grid scale of the model, which is 100 km × 100 km.

The skill evaluation from step 1 serves as critical context in terms of how skillful the predictions were in the past (Taylor et al. 2015), but in real time a practitioner would only have access to a prediction from a single initialization. Thus, in this step, we focus on a single initialization: we use the hindcast initialized in January of 2010 to examine October 2010–September 2015, which henceforth we will refer to simply as “2011–15,” although it technically refers to water years 2011–15. We underscore that the hindcast initialized in 2010 was not part of the skill evaluation conducted in step 1; rather, starting in step 2, we examine how a particular decadal prediction could be examined in practice.

b. Step 2: Manipulate predictions

In this step, we present two different approaches to manipulating the decadal predictions (Fig. 1). Users of climate information prefer familiar formats (Taylor et al. 2015), and therefore we manipulate the information using a deterministic-anomaly approach (as is done in projections of climate change) as well as using a probabilistic tercile-based approach (as in seasonal forecasts).

In the previous step, we use the hindcasts initialized from 1980 to 2009 to evaluate the skill (i.e., the ACC and MSSS; Fig. 2). The skill evaluation from step 1 serves as critical context in terms of how skillful the predictions were in the past (Taylor et al. 2015), but in real time a practitioner would only have access to a prediction from a single initialization. Thus, in this step, we focus on a single initialization: we use the hindcast initialized in January of 2010 to examine October 2010–September 2015, which henceforth we will refer to simply as “2011–15,” although it technically refers to water years 2011–15. We underscore that the hindcast initialized in 2010 was not part of the skill evaluation conducted in step 1; rather, starting in step 2, we examine how a particular decadal prediction could be examined in practice.

1) Deterministic-anomaly approach

For the deterministic-anomaly approach, we calculate the average over the previously drift-corrected hindcasts to get an anomaly with respect to the climatological average (1981–2010). The hindcasts span the 5-yr period (2011–15) and each year has 10 members; hence, the average is taken over the 50 (= 10 × 5) annual temperatures for each grid box. We note that this anomaly calculation is identical to taking the average of the 5-yr period of ensemble member 1 and then averaging it over all 10 ensemble members.

2) Probabilistic tercile-based approach

Terciles divide a variable into three equally probable categories on the basis of the value of the variable. These three categories are known as tercile categories and can be referred to as the below-normal, normal, and above-normal categories. The tercile categories are calculated from the 30 years of observations from 1981 to 2010. Thus, by definition, each tercile category contains 10 years of observations.

For the probabilistic approach, we examine the annual-average temperature for each ensemble member over 2011–15, counting how many of the members are in each of the observed tercile categories. The number of hindcast values in each tercile category is expressed as a probability. For example, if 32 of the 50 members are in the above-normal tercile category, then the probability
of the average annual temperature for water years 2011–15 being in the above-normal tercile category in any particular year during 2011–15 would be 64% (=32/50).

It is important to note that, as the prediction goes farther into the future from 2011 to 2015, the variance among the ensemble members increases; hence, the 50-member sample used here to calculate the probabilistic terciles is not uniform. This approach has the advantage of using the annual averages, which results in a distribution of predicted annual-average temperatures for the near term, which is relevant to practitioner needs (Barsugli et al. 2009), and relevance has been found to be a key criterion for use of scientific information (Cash et al. 2003). Another way to calculate the tercile probabilities would be to use 5-yr-mean predictions, which would result in a 10-member sample. This would have the advantage of being a uniform sample that is more readily comparable with the samples used to calculate the skill scores in step 1 but is a smaller sample size, and 5-yr averages are temporally coarse for use in management applications, such as hydrologic models. From this point forward, we present results from the probabilistic tercile-based approach using the 50-member annual sample, but results using the 10-member sample for the 5-yr means are shown in the online supplemental material.

c. Step 3: Translate predictions

In this step, we demonstrate how the output from step 2 can be translated for potential users. Here, we continue with the 2011–15 hindcast but focus on two case-study watersheds: the Colorado/South Platte watershed in Colorado and the Ottawa watershed in Ontario, Canada. We selected the Colorado watershed because it shows some skill (see upcoming results in section 5a) and it is of interest to a hydrologic-modeling effort that will be examined in future work. We selected the Ottawa watershed because it is a region that shows skill in terms of both ACC and MSSS (see section 5a).

We examine two translation methods, delta and weighted resample, that correspond to the anomaly and probabilistic manipulation methods, respectively, as well as a hybrid method (see Fig. 1). We compare the translation methods with the climatological base period for each watershed.

1) CLIMATOLOGY

Past climate observations are a common baseline for users to examine historical variability. In particular, the 30-yr climatology (Guttman 1989) is often used. For each watershed, we calculate the observed average temperatures for each year over the watershed [i.e., from 1981 to 2010 (=30 yr)].

2) DELTA

In the delta approach, the average anomaly from the hindcast is simply added as a delta to each year of the 30-yr climatology.

3) WEIGHTED RESAMPLE

Although probabilistic outlooks can be difficult for users to integrate with their decision-making (Nicholls 1999), advances have been made in the application of seasonal forecasts (Goddard et al. 2001). Here, we use a simple weighted resampling technique that is based on

---

**Fig. 2.** (a) Anomaly correlation between initialized hindcast ensemble mean and observations (years 1–5). Also shown are zoomed-in views over the (b) Colorado and (c) Ottawa watersheds. In (a), hatched areas show significance at the 99% level. The stars correspond to the Colorado and Ottawa watersheds. The analysis is for 30 five-year periods initialized annually from 1980 to 2009.
bootstrapping (Efron and Tibshirani 1993) with replacement. As noted previously [i.e., section 4b(2)], each tercile category (below normal, normal, and above normal) contains 10 years of observations. Observation years are resampled with replacement consistent with the probabilities for each tercile category from the hindcast. Here, we choose a resample length of 100 and provide an example as an illustration: if the hindcast probability of being in the below-normal category was 19%, normal was 25%, and above-normal was 56%, we would resample observations from the below-normal tercile category 19 times, the normal category 25 times, and the above-normal category 56 times. Since there are only 10 observations in each tercile category, resampling is done with replacement. This has been successfully employed in weather-generation techniques (Yates et al. 2003; Clark et al. 2004; Appipattanavis et al. 2007) and in seasonal-forecasting applications (Towler et al. 2010, 2013).

4) HYBRID

Given that decadal predictions bridge the gap between seasonal forecasts and centennial projections, it is fitting to explore an approach that is a hybrid between the delta and weighted resample. Here, we perform the weighted resample (referred to as “Resample”), as in section 4c(3), but then we 1) subtract the difference between the Resample average and the climatology average and 2) add the delta from section 4c(2):

$$\text{Hybrid} = \text{Resample} - (\text{Resample}_{\text{Average}} - \text{climatology}_{\text{Average}}) + \Delta.$$

Performing these operations recenters the resample distribution around the climatological average and then shifts the entire distribution by the delta.

5. Results

a. Evaluation

Using the hindcasts initialized from 1980 to 2009 (Fig. 2), we calculate the ACC and cross-validated MSSS over most of North America (Fig. 2a), as well as for the two watersheds (Figs. 2b,c). Figure 2a shows the spatial distribution of ACC for annual-mean temperature for the CCSM4 decadal hindcasts for years 1–5. There is substantial skill (positive ACC) in the CCSM4 decadal hindcast in predicting temperature over much of the United States. Positive ACC is also statistically significant at the 99% level over much of the United States, as indicated by the hatched areas. There are a few notable locations of zero or negative correlation, including the Pacific Northwest and across western Canada, dipping down into the central United States, some of the western U.S. coast, Florida, and Mexico. The ACC is also positive over most of the Colorado watershed (Fig. 2b; except for one of the six grid boxes that overlaps the watershed) with an ACC value of 0.366, and the Ottawa watershed shows higher skill (Fig. 2c) with an ACC value of 0.654 in comparison with the Colorado watershed.

Figure 3a shows the cross-validated MSSS for temperature, where nonwhite areas show places where the hindcasts have skill over climatology. The CCSM4 decadal hindcasts show more skill in the northeastern United States and Canada, but also some skill across the western United States. The skill is statistically significant at the 99% level over only a few areas, such as the northeastern United States and eastern Canada. The skill across the West is noisy, partially because we perform the analysis at the grid scale and do not perform spatial smoothing. We see this over the Colorado watershed (Fig. 3b), where only some of the grid boxes show skill and the watershed average does not show positive skill over climatology (MSSS = −0.038). Over the Ottawa watershed (Fig. 3c), the MSSS is positive for all of the grid boxes; MSSS = 0.246 and is significant.

b. Data manipulation

1) DETERMINISTIC

For the deterministic approach, Fig. 4 shows the temperature anomalies for 2011–15 with respect to the 1981–2010 average for the observations and hindcast. For the observations (Fig. 4, top panel) we see a West–East contrast: there is up to a degree of warming in the West, while the East is normal to slightly cooler. The northeastern United States and Canada also show warm anomalies. For the hindcast (Fig. 4, bottom panel), the warming is more uniform across the United States. The hindcast shows some underestimation of warming in the West and overestimation of warming in the East.

2) PROBABILISTIC

Figure 5 shows the probabilistic approach. Figures 5a–c show the observed tercile percentages; that is, for each grid box, we counted the number of observed years (from 2011 to 2015) that had annual-average temperatures in the above-normal, near-normal, and below-normal terciles; terciles are calculated on the basis of the 1981–2010 base period. For example, if in a particular grid box two of the five years had annual-average temperatures in the above-normal tercile, then the probability would be 40% (=2/5), and so on. Figures 5d–f show the probabilities derived from the hindcast. Although comparing the frequency of observed years with the frequency of ensemble members in
each tercile is not a direct comparison and is not an indication of the skill over the long term (like the ACC and MSSS are; also see section 6 for additional choices for verification metrics), it gives users a visual sense of how the observations and hindcasts compare for this particular hindcast. For instance, in terms of the above-normal tercile (Figs. 5c,f), the observations (Fig. 5c) show some West-versus-East contrast: most of the West is in the above-normal tercile (i.e., 60%–80%) and the eastern part of the United States had about 40% of the years in the above-normal tercile. The initialized hindcast (Fig. 5f) shows more contrast in the North versus the South: the southern part of the United States, as well as into the Rocky Mountains, has 50%–70% of members in the above-normal tercile; the northern part of the United States is in the 30%–40% range. As seen in the anomaly approach, relative to the observations the hindcast underestimates the probability in the West and overestimates the likelihood in the East. Both show the odds tilting toward the above-normal tercile, as expected in a warming climate. For the below-average tercile (Figs. 5a,d), the observations (Fig. 5a) show that in the West the percentage of years in the below-normal tercile was low (0%–20%), except for an area in south-central Canada and the north-central United States (40%–60%). In the initialized hindcast (Fig. 5d), the majority of locations show <20% of the ensemble members to be in this category.

c. Data translation

For this step, the deterministic anomalies (from Fig. 4) and probabilities (from Fig. 5) from each grid box are averaged over each watershed region; results are shown in Table 1. The values in Table 1 are used for each of the translation methods, which are compared with the climatological base period for each watershed (see Fig. 1).
1) COLORADO WATERSHED

Table 1 shows that for 2011–15 the hindcast anomaly was predicted to be less (0.2°C) than what was observed (0.9°C). For each tercile, both the hindcast and the observation agree in terms of direction relative to climatology. For example, in the above-normal category, both the observed frequency (73%) and hindcast likelihood (56%) were greater than what would be expected under climatology (i.e., 33%).

Figure 6 shows the average temperature for the observed (black) and three translations applied (blue, red, and green) as box plots. It specifically shows the observed climatology (1981–2010), 2011–15 observed, delta translation using the hindcast delta (0.2°C), delta translation using the observed delta (0.9°C), the weighted resample using the hindcast probabilities (19/25/56), the weighted resample using the observed frequencies (0/27/73), the hybrid using the hindcast probabilities/delta, and the hybrid using the observed frequencies/delta. In Fig. 6, there is a marked difference in the distribution from climatology to the observed 2011–15. Using the delta approach, the entire distribution is shifted (blue box plots in Fig. 6), but the distribution shape is the same as climatology. The weighted resample results in a change in shape of the distribution from climatology (red box plots in Fig. 6). We can think of the observed hindcast delta and the observed frequency from Fig. 6 as being the “perfect prediction,” and therefore these results show the best case. The hybrid approach both changes the shape of the distribution and shifts the entire distribution by the delta (green box plots in Fig. 6).

To compare the distributions resulting from the three translation approaches, we calculate the percent errors (PE):

$$PE = \left( \frac{Q_{XP} - Q_{XO}}{Q_{XO}} \right) \times 100,$$

where $Q_{XP}$ is the $X$th percentile of the predicted distribution and $Q_{XO}$ is the $X$th percentile of the observed 2011–15 distribution.

Table 2 shows the PEs for the average and select percentiles for 2011–15 observed from climatology as well as for the observed best cases: delta translation using the observed delta (0.9°C), the weighted resample using the observed frequencies (0/27/73), and the hybrid using both. For the average and all percentiles considered, using climatology always results in higher

### Table 1

<table>
<thead>
<tr>
<th>Color</th>
<th>Obs</th>
<th>Hindcast</th>
<th>Ottawa</th>
<th>Obs</th>
<th>Hindcast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly (°C)</td>
<td>0.9</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Probability (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below normal</td>
<td>0</td>
<td>19</td>
<td>27</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>27</td>
<td>25</td>
<td>30</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>Above normal</td>
<td>73</td>
<td>56</td>
<td>43</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>
percent errors for this watershed (PE ≤ −9%). By definition, both the delta and hybrid methods predict the mean (PE = 0%). In terms of the select percentiles, the three methods perform similarly as measured by PE, sometimes doing better and sometimes doing worse depending on the percentile, but all of them outperform climatology.

2) Ottawa Watershed

Table 1 shows that for 2011–15 the hindcast anomaly was predicted to be slightly higher (0.3°C) than what was observed (0.2°C). For the above-normal category, both the observed frequency (43%) and hindcast likelihood (43%) were slightly greater than what would be expected under climatology (i.e., 33%). Overall, the hindcast probabilities (6/51/43) and especially the observed frequencies (27/30/43) were very similar to climatology (33/33/33).

Figure 7 shows the average temperature for the observed (black) and three translations applied (blue, red, and green) as box plots. Figure 7 shows the observed climatology (1981–2010), 2011–15 observed, delta translation using the hindcast (0.3°C), delta translation using the observed delta (0.2°C), the weighted resample using the hindcast probabilities (6/51/43), and the weighted resample using the observed frequencies (27/30/43). The main thing to notice is that most of these distributions are fairly similar. For instance, the range for climatology is 2.8°C–6.4°C, and the observed 2011–15 range is 3.0°C–6.5°C. This is reflected in the distributions resulting from the hindcast and observed delta approach; this makes sense given the relatively low delta (as compared with the Colorado observed delta). Although the weighted resample results in a change in shape of the distribution from climatology, it is subtle since the hindcast probabilities and observed frequencies are similar to climatology. From Table 2, we see that all three of the methods capture the average (PE = 0%), although climatology also performs reasonably well (PE = −5%). Here, we see that most of the methods, including climatology, perform well in terms of PE,
across all percentiles. The highest error comes from the median prediction from climatology (PE = −12%).

We note that the observed 2011–15 distributions for Ottawa, as well as Colorado, show aspects of strong skewness (i.e., asymmetric both-side tails and/or large difference between mean and median). This skewness is missing from the climatological distribution as well as the translated distributions (Figs. 6 and 7). Although other methods could be developed to better capture the observed skewness, we aimed to make the translation approaches in this framework be familiar to practitioners using climate information at other time scales, such as seasonal forecasts and projections of climate change. Further, we show in terms of PE that for the average and select percentiles of the distribution all three translation methods typically do better than climatology (Table 2). For Colorado this is true for the average and all of the select percentiles; for Ottawa this is true for the average, Q25, Q50, and Q95. For the few cases in which climatology does better, there is a translation method that performs similarly; for Ottawa this is seen for Q5 and Q75.

6. Discussion and conclusions

The purpose of this paper is to introduce decadal prediction to practitioners of applied climatology and to provide a framework for its potential use by them. The framework is purposely built upon approaches and presentation styles that would be familiar to practitioners using climate information at other time scales, such as seasonal forecasts and projections of climate change. The decadal predictions are still exploratory, however. As such, this paper provides a demonstration, and our analysis is confined to the decadal temperature hindcasts from a single model, CCSM4. Although the goal of the paper is not to evaluate skill per se, the CCSM4 skill scores (ACC in Fig. 1 and MSSS in Fig. 2) were included to give users a sense of the skill. We point out that there are many other verification measures that could be examined; for instance, probabilistic metrics such as reliability diagrams (e.g., Wilks 2011; Christidis and Stott 2014) and the continuous rank probability skill score (Goddard et al. 2012b) are additional ways in which the hindcasts could be evaluated. Establishing credibility is key to the adoption of scientific information (Cash et al. 2003), although articulating the level of reliability required is challenging for users (Rayner et al. 2005). In practice, multimodel ensembles have been found to outperform a single model (Meehl et al. 2014), and, as noted in section 2, multimodel skill evaluations have been performed elsewhere (e.g., Meehl et al. 2014; Kirtman et al. 2013; Kim et al. 2012): the general conclusion is that, although the CMIP5 multimodel decadal temperature hindcasts do show some skill, the uninitialized CMIP3 and CMIP5 multimodel climate change projections already showed skill, with the initialization only adding skill in a few regions. As such, we note that the steps developed here are flexible and could also be applied to the uninitialized climate change projections. Decadal predictions offer additional potential for skill improvement, however, as has been shown in several historical case studies that were driven by natural variability, such as the recent global-warming hiatus (Kosaka and Xie 2013; Trenberth and Fasullo 2013; Meehl and Teng 2012), which is why we choose them for the demonstration of this framework.

In addressing the need to provide potential users with skill evaluations (Taylor et al. 2015), we emphasize that skill evaluations will be more salient if they are tailored
to a particular user application. To illustrate this, we evaluate water years, rather than calendar years, but note that particular seasons or months may be of most interest for an application. This suggests that skill evaluations may need to be interactive so that users can examine particular time periods and prediction periods relevant to their application. Further, in addition to the skill metrics shown here, other verification efforts, such as the evaluation of how the model captures particular user-relevant patterns (e.g., IPO, AMO, etc.) and/or certain time series evolutions (e.g., predicted differences between year 1, year 2, and so forth), should also be examined to explore the utility of the decadal hindcasts for practitioners.

The manipulation and associated translation approaches have different advantages and disadvantages. Like seasonal forecasts, the tercile-based approach is appealing because it is probabilistic, reflecting the uncertain nature of climate. This provides an advantage in that the probabilities can be used as weights to resample historical data for use in water management (Table 1; Figs. 6 and 7, red box plots). We point out that, in our demonstration of the framework, we calculate the tercile probabilities using the annual predictions from 2011 to 2015 (i.e., 50 members); this has the disadvantage of not being a uniform sample because as the forecast goes on from year 1 to year 5 the variance increases. It does, however, have the advantage of seamlessly translating into an annual distribution of temperatures, which is useful for water-management applications. Note that the framework is general, making it flexible to incorporate other ways of manipulating the predictions that may be more suitable for other applications (e.g., see the online supplemental material). We also point out that currently the probabilistic approach is limited by the sample size of the decadal predictions; in the ideal case, we would have a larger ensemble, which would allow for a better characterization of the prediction accuracy of the annual values for a specific year.

In addition, we calculate the probabilities for each grid box and then present the prediction as three maps for each tercile (Figs. 5d–f); this is relatively complicated, however, and may not be the best way to communicate information to the end user. We note that seasonal forecasts issued by agencies such as CPC and IRI combine the probabilistic forecasts into a single map that is more user friendly. In comparison, the deterministic-anomaly/delta approach is simpler. The anomaly/delta can be presented as a single map that is straightforward to read (Fig. 4, bottom panel), and, in terms of translation, adding the delta to the historical distribution is simple (Table 1; Figs. 6 and 7, blue box plots). The trade-off is that the entire distribution is simply shifted and does not reflect changes to the shape of the distribution. Last, the hybrid approach captures the change in shape from the resample as well as the shift in average from the delta but requires some additional manipulation and adds complexity for the verification. For the Colorado watershed, where there was a noticeable shift toward higher temperatures, the delta, weighted resample, and hybrid translations all did a better job of capturing the 2011–15 prediction than did using the 1981–2010 climatology (Table 2). For the Ottawa watershed, the differences among the translation methods were less noticeable since the 2011–15 prediction period was only subtly different than climatology.

Taylor et al. (2015) found that users prefer familiar formats, which both the probabilistic and anomaly approach are, but whether to use the associated weighted resample or delta approach will likely depend on the context of the application. In terms of using the probabilistic predictions as weights in a resampling scheme, users will only resample historical data that they have already seen. This fits with conservative planning practices that rely on historical data, yet it cannot inform the potential for new values. On the other hand, the delta approach can result in new values, and if the delta is positive then it would result in temperatures higher than have been experienced. The delta approach can therefore help managers to understand how close they are to their monitoring signposts and action triggers (Raucher et al. 2015). The hybrid method offers an alternative that users can consider if they are interested in seeing how their system responds to both the reshuffling and shifting of their historical record to reflect the probabilistic and deterministic aspects of the prediction.

In closing, we have presented three steps toward applying decadal temperature predictions but recognize that additional steps may need to be taken before they can be readily incorporated in applications. Many applications rely on the use of impact models, such as hydrologic models, that often require more refined time steps and are driven with precipitation in addition to temperature. This would require an additional step to downscale to finer time scales. Precipitation prediction is also more challenging than temperature, yet many users have found ways to consider seasonal and long-term climate information in their planning (Raucher et al. 2015). Future work is under way to illustrate the next steps that would be needed, in addition to those presented here, to incorporate the decadal predictions into a specific impact-modeling application.

Acknowledgments. This work was partially funded by NSF Grant AGS-1419563 as part of the Understanding Decision–Climate Interactions on Decadal Scales
REFERENCES


