Performance of Alternative “Normals” for Tracking Climate Changes, Using Homogenized and Nonhomogenized Seasonal U.S. Surface Temperatures

DANIEL S. WILKS

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

ROBERT E. LIVEZEY*

Bethesda, Maryland

(Manuscript received 10 January 2013, in final form 1 April 2013)

ABSTRACT

Eleven alternatives to the annually updated 30-yr average for specifying climate “normals” are considered for the purpose of projecting nonstationarity in the mean U.S. temperature climate during 2006–12. Comparisons are made for homogenized U.S. Historical Climatology Network station data, corresponding nonhomogenized station data, and spatially aggregated (“megadivision”) data. The use of homogenized station data shows clear improvement over nonhomogenized station data and spatially aggregated data in terms of mean-squared specification errors on independent data. The best single method overall was the most recent 15-yr average as implemented by the Climate Prediction Center (CPC15), consistent with previous work using nonhomogenized and spatially aggregated data, although “hinge” functions with the change point fixed at 1975 performed well for the spring and summer seasons. A hybrid normals-specification method, using one of these piecewise continuous functions when the regressions are sufficiently strong and the CPC15 otherwise, exhibits a favorable trade-off between squared error and bias that may make it an optimal choice for some users.

1. Introduction

Access to accurate “normal” temperatures is enormously important to a wide range of decision makers and planners from both the private and public sectors (Arguez et al. 2012, 2013). Normals are estimates of the statistically expected values of climatic variables and are used both as reference values against which variations can be assessed and as simple predictors for upcoming years (e.g., Trewin 2007). Normals are most commonly computed for surface temperatures and precipitation, and the U.S. official standard for their determination is the 30-yr average updated every 10 yr, as recommended by the World Meteorological Organization (WMO). This standard presumes that the climate is stationary (its statistics are not changing) or is changing very slowly. Regardless of the causes, it is clear that this presumption is no longer true, as both the U.S. and global climates have been consistently warming at moderate to rapid rates for three to four decades (Solomon et al. 2007; Livezey et al. 2007, hereinafter L07). As a consequence, 30-yr surface temperature averages as estimates of current or near-future expected temperatures (i.e., as normals) will generally have a cold bias—in many instances, severely so. Good alternatives that are clear improvements to the conventional 30-yr averages for long-range planning and near-term decision making are available.

In this context, L07 examined the expected squared error of four different approaches to tracking a steadily (linearly) changing climate: 1) 30-yr averages, 2) shorter-window averages tailored to the climate record to minimize squared error, called optimal climate normals (OCNs), 3) long-term linear trends, and 4) least squares fits to a simple statistical model of observed large-scale climate change called the hinge. The hinge model is piecewise-continuous and -linear with a zero slope from
1940 to 1975 (no climatic change) and a linear trend thereafter. L07 concluded that OCNs should replace 30-yr normals for records with small to moderate rates of change and that hinge fits should be used for records with large post-1975 slopes in a hybrid approach. They did not quantitatively address the conditions for choosing one or the other.

L07 examined cases and comprehensively computed hinge fits only through 2005 for the set of U.S. megadivisions. The megadivisions are 102 regions in the conterminous United States that have been derived from the 344 National Climatic Data Center (NCDC) divisions (Guttman and Quayle 1996), by aggregating some of the smaller NCDC divisions that are located mainly in the eastern United States to yield approximately equal areas. Wilks (2013, hereinafter W13) recognized the opportunity to test the performance of the L07 hinge and eight other alternatives on a 6-yr (2006–11) independent sample from the same dataset. Two of these alternatives were fixed-window (for all seasons and megadivisions) versions of the OCN that Huang et al. (1996, hereinafter H96) concluded were best overall for seasonal hindcasts for the following year: most recent 10-yr averages for temperature as implemented by the Climate Prediction Center (CPC10) and 15-yr averages (CPC15) for precipitation. Because H96 used data through 1993, W13 also tested all of the methods over 1994–2011. The other six alternatives included two other OCN methods, a linear trend fit to the full record, and three other hingelike approaches to test the validity of L07’s model choices of a fixed change point at 1975 and zero slope before 1975. The other three variants of the hinge allow either 1) the change point to vary from record to record or 2) the pre-change-point slope to be nonzero, or they allow 3) both. Examples of the hinge, linear trend, CPC15, and 30-yr averages updated annually (WMO30) are presented in Fig. 1. The differences among these methods, two non-fixed-period OCNs, and two additional OCN formulations introduced for this study will be discussed further in the next section.

W13’s temperature results (which will be our focus here) surprisingly did not support the recommendations of either H96 or L07. CPC15 (rather than CPC10) was the best alternative to WMO30 (updated annually) overall for both test periods and in 11 of 12 region–season combinations for the 1994–2011 independent sample and 4 of 12 for 2006–11. The hinge and one of its variants (nonzero slope up to 1975) were only competitive with the worst-performing OCN, and the other two hinge variants (estimated change points) were substantially inferior performers over both test periods. In addition, W13 was unable to identify a slope threshold for which L07’s hybrid recommendation (replace a lagging average with a hinge fit for slopes greater than the threshold) did not increase overall error. W13 thus concluded that the use of hinge normals “is in general not yet justified for divisional U.S. seasonal temperature.”

The results in W13 decisively supported L07’s choice of 1975 as a fixed change point, however, together with the model of no climate change from 1940 to 1975, using fully independent test data. Estimating change points considerably degraded the accuracy of the specified climate normals, and allowing nonzero pre-1975 slopes had an inconsequential effect on the performance results. Second, the hinge had close to zero bias over the 1994–2011 test period, whereas all OCN variants, WMO30, and linear trends exhibited cold biases, which were especially large for the latter two methods.

Two dataset characteristics could potentially have affected W13’s conclusions: station records versus areally averaged megadivision records and “homogenized” versus nonhomogenized records. The U.S. megadivision surface temperature data used in W13 are serially complete records that are aggregates of station records
over defined areas. These station records have undergone some modest degree of quality control and have been bias adjusted for time-of-observation changes (NCDC 1994; Guttman and Quayle 1996). Nevertheless, a large number of so-called inhomogeneities related to station moves, modifications to station environments, changes in instruments and measurement practices, instrument drifts and errors, and so on remain in the records. None of these problems are related to climatic change, but many of them can introduce prominent and spurious change points, discontinuities, and trends, which in turn may degrade the predictive utility of normals that are computed from them (Trewin 2007).

Advanced algorithms that have proven effective in correcting these inhomogeneities and in optimally estimating missing data have been developed by NCDC (Menne and Williams 2009; Williams et al. 2012) and have been applied to create a large subset of long and serially complete U.S. surface temperature station records (Menne et al. 2009). The data products resulting from application of these algorithms are referred to as homogenized series.

It is a plausible hypothesis that the use of both station and homogenized temperature series will improve the relative performance of the hinge. Individual station records will be noisier than area averages. They will likely contain some larger warming trends more suitable to representation by the hinge, however, because of averaging of differing station trends across a large mega-division. For homogenization, methods for computing normals that rely on fitting large swaths of station records, particularly hinges and trends but even WMO30, would be especially and negatively affected by uncorrected inhomogeneities. NCDC has long recognized this problem in their normals calculations (Peterson and Easterling 1994; Easterling and Peterson 1995), and all official temperature normals are produced from homogenized records (Arguez et al. 2012).

Our objective here is to assess the impact of the use of station and homogenized data records on W13’s conclusions and to modify them if necessary to provide more complete guidance to producers (including NCDC) and users of normals. First, we repeat W13’s error calculations on nonhomogenized station records with time-of-observation bias corrections to compare with the area-averaged results in W13. Next, we compare normals formulations calculated from a matched set of nonmissing station and season combinations from the nonhomogenized data and the corresponding homogenized data values. Third, results from the full set of homogenized station records are compared with the previous matched set. Section 2 contains additional information about the datasets, tested methods, and normals formulations not considered in W13. Section 3a begins with overall results from the three comparisons, section 3b presents disaggregated results for individual region-seasons, section 3c shows new hybrid-normals results, and section 4 contains discussion and conclusions.

2. Data, normals formulations, and performance statistics

a. Data

Two monthly-average surface station temperature datasets extending from January of 1940 through February of 2013 were used here, each of which includes the 1218 stations composing the U.S. Historical Climatology Network (USHCN): 1) quality-controlled records with time-of-observation bias corrections and 2) fully homogenized records. The former, referred to here as nonhomogenized data, have undergone processing comparable to the megadivision data but have missing months that must be accounted for in the analyses. The latter, referred to here as homogenized, are serially complete. Both datasets were obtained from the latest release, version 2.5, of USHCN (ushcn.v2.5.0.20130307; Menne et al. 2009). In addition, the megadivisional data used in W13 were updated through February of 2013, extending the record used there by 12 months. The megadivisional data are spatially aggregated from time-of-observation bias-corrected station data but have not otherwise been homogenized. As in W13, monthly averages from each of the datasets were combined over three consecutive months to form separate series for 12 overlapping “seasons”: January–March (JFM), February–April (FMA), and so on.

b.Normals formulations

Twelve methods for estimating normals were tested. Although several were defined in section 1, we repeat them here for clarity. Equations are only provided for two methods introduced here for the first time. Refer to W13 for equations and further details for the other 10 methods.

Seven of the methods are derived from simple averages of a particular season’s values over some number of the immediately previous years. Three of these use fixed window lengths for all stations and all test years, and four use varying window lengths that depend on the available dependent record. Except for WMO30, all of these are variants of the OCN concept:

1) WMO30 is a 30-yr average, which is the WMO standard that is used for official U.S. normals (Fig. 1), but here it is updated annually rather than every 10 or 30 yr.
2) CPC_{15} is the most recent 15-yr average (Fig. 1), currently used by CPC for 9-month-in-advance seasonal precipitation forecasting but applied here to temperature hindcasts.

3) CPC_{10} is the most recent 10-yr average, currently used by CPC for 9-month-in-advance seasonal temperature forecasting.

4) OCN is the classic approach to tailoring the averaging window to a particular location/season/forecast. The averaging period is the one that performs best in tests over all $M = n - 30$ of the most recent training windows, where $n$ is the length of the dependent record. When predicting for 2006, for example, there are $n = 66$ training years (1940–2005), in which period there are $M = 66 - 30 = 36$ thirty-year training windows (from 1940–69 through 1975–2004) available to optimize the OCN over the $M$ years 1970–2005.

5) OCN$_M$ is the $M$-fold OCN introduced in W13. It is the lag-zero intercept of a weighted regression of conventional OCN estimates, computed over all $M = n - 30$ of the most recent training windows. It is designed to capture accelerating (or decelerating) climate changes.

6) OCN$_{1P}$ is the one-phase parametric hinge. This method yields the theoretically best averaging period under linear climate change at the post-1975 rate implied by the (one phase; i.e., assuming a stationary climate until 1975) hinge, assuming negligible autocorrelation of Gaussian residual variability. Suggested by Eq. (7) in L07, OCN$_{1P}$ optimizes the averaging period $N$ by striking a balance between random averaging error $\eta_a$, which decreases with increasing averaging-window length, and systematic trend bias $\eta_b$, which increases with both averaging-window length and with the strength of the (post 1975) linear trend. To be specific, the OCN$_{1P}$ averaging period is the integer $N$, $1 \leq N \leq 30$, that minimizes the sum of these two errors,

$$N: \left\{ \min_{1 \leq N \leq 30} (\eta_a + \eta_b) = \min_{1 \leq N \leq 30} \left[ \frac{1}{N} + \frac{\beta^2}{4} (N + 1)^2 \right] \right\},$$

(1)

where

$$\beta = b_2 / s_e$$

(2)

is the slope of the post-1975 annual change in the climate mean $b_2$, standardized by the residual standard deviation $s_e$ in the post-1975 period.

7) OCN$_{2P}$ is the two-phase parametric hinge. This normals formulation is computed as described above for OCN$_{1P}$, except that $\beta$ is computed from the post-1975 slope of the two-phase 1975 hinge (pre-1975 slope not constrained to be zero; see below and W13), assuming negligible autocorrelation of the Gaussian residual variability.

The remaining five methods all involve least squares regression fitting of a particular statistical model of piecewise-continuous and piecewise-linear climate change to the dependent record, and a 1-yr extrapolation from the model:

8) Linear trend is the conventional ordinary least squares linear trend, fit to the training record from 1940 through the year immediately preceding the target year (see Fig. 1).

9) The 1975 hinge indicates no change (zero slope) until the change point at 1975, with linear trend thereafter (Fig. 1; see L07 also).

10) The estimated hinge is the same as the 1975 hinge except that the change point is not fixed, but rather is an additional parameter to be estimated.

11) The 1975 two-phase hinge is the same as the 1975 hinge except that the 1940–75 slope is not constrained to be zero, but rather is an additional parameter to be estimated.

12) The estimated two-phase hinge provides for the change point and both the pre- and post-change-point slopes to be determined by the best least squares fit.

c. Performance statistics

For all methods, 9-month-lead hindcasts are made for all seasons from JFM 2006 through DJF 2012 (the year applies to the first month in the target), when possible (section 2d describes the treatment of missing stations/seasons in the unhomogenized data). This period was chosen to provide fully independent verification data, since these years were not available to L07. Thus, 12 more hindcasts are potentially possible for each location than in W13. All corresponding antecedent seasons are used for each season’s hindcasts, and so all methods are updated using an additional year’s data for each subsequent hindcast. For example, JJA temperatures from 1940 to 2007 compose the training dataset for hindcasts for JJA 2008, JJA temperatures are used from 1940 to 2008 for 2009, and so forth. Hindcasts were also made for all seasons during the full period from JFM 1994 onward, but those results do not alter any of our conclusions, and so they will not be discussed.

Four sets of hindcasts are produced: The sets are for the nonhomogenized station data, for homogenized-station-data cases matched to the nonhomogenized cases (i.e., deleting homogenized-data cases corresponding to missing
unhomogenized station–season combinations), for the full set of homogenized station data cases, and for the megadivisional data used in W13 and updated through DJF 2012.

Two error statistics are computed for each station–season hindcast sequence: the bias (average difference between hindcasts and observations, or mean error) and the mean squared error (MSE). The latter is expressed as the reduction in variance (RV) relative to WMO$_{30}$:

$$RV = \frac{\text{MSE}}{\text{MSE}_{\text{WMO}_{30}}}.$$  

Values smaller than RV = 1 indicate that the alternative method yielded more accurate hindcasts overall than WMO$_{30}$. Of course, bias contributes to squared error, and so the two performance measures are not independent. We report both because some users may value small bias as much as or more than small MSE.

The performance measures are also computed for subsets of all of the cases to investigate seasonal and geographic differences. Seasonal windows are defined as in W13: winter combines DJF, JFM, and FMA; spring combines MAM, AMJ, and MJJ; and so on. Geographic regions are defined almost exactly as in W13. The “west” region is composed of the U.S. states including and west of Montana, Wyoming, Colorado, and New Mexico and the “east” region includes the area east of the Mississippi River (east of Wisconsin, Illinois, Kentucky, Tennessee, and Mississippi, inclusive), with the “central” region between the two.

Bootstrap confidence intervals are estimated for all performance measures. Spatial cross correlations are preserved in these calculations.

d. Treatment of missing data for tests on nonhomogenized data

Seasonal averages of nonhomogenized station data are considered to be missing if one or more of the constituent months is missing. Hindcasts for the nonhomogenized data must accommodate missing seasons in such a way that performance measures can still be accurately estimated from an already relatively short independent period yet still permit fair comparisons of these measures between methods. A test case cannot be included if either the target season is missing or the hindcast method cannot be satisfactorily computed for that target season.

The most important constraint imposed by the missing data is in the computation of the various $k$-year averages (thus affecting seven methods), whether for making the actual hindcast or for determining the best $k^*$ (affecting OCN and OCN$_M$ only) over the dependent training period. We require a minimum of 10 training years in the dependent dataset for determination of $k^*$ for OCN or OCN$_M$. If 5 or fewer years are missing from a $k$-year computational window, the window is extended back in time up to 5 yr until a $k$-year average can be computed. The most frequently affected method is WMO$_{30}$ (for which $k =$ 30). In fact, it is rare that a case cannot be included in a RV calculation other than when an incomputable WMO$_{30}$ precludes it because the corresponding factor in the denominator of Eq. (3) cannot be calculated. There are only a few instances in which WMO$_{30}$ can be estimated that OCN and OCN$_M$ cannot, and there are none for any other method.

Approximately 20% of station–season cases end up being excluded from the overall average RV for the nonhomogenized and matched homogenized station data. For bias calculations, all nonmissing cases are included, regardless of whether that station–season case had valid estimates for other methods, and so the sample sizes are different for the different methods.

3. Results

a. Overall results and comparisons

1) REDUCTION OF VARIANCE

RVs for 11 alternative normals used for hindcasts of 2006–12 seasonal temperatures are presented in Fig. 2 for the four datasets. A comparison of the megadivisional (shaded bars in Fig. 2) and nonhomogenized station data (middle bars in Fig. 2) illustrates that the spatial aggregation mainly (9 of 11 methods) increases RV in the annual and national aggregation—very seriously so in the case of all four hinge variants. Despite the large improvements associated with the station data, for the two estimated change-point hinges, they and OCN$_M$ (which changes little) remain uncompetitive with the benchmark WMO$_{30}$. An increase in RV for station versus megadivisional data is seen for OCN, likely because this method is particularly sensitive to increased noise in the station data. CPC$_{15}$ still clearly performs best, consistent with the conclusion in W13, as it is the only method with RV < 1 for both datasets. The other methods with RV < 1 for the nonhomogenized HCN stations, CPC$_{10}$ and the newly introduced OCN$_{1P}$ and OCN$_{2P}$ (both based on hinge-fit slopes), are barely so.

A comparison of the nonhomogenized station (middle bars) and matched homogenized station (left bars in Fig. 2) results shows small improvements in 8 of 11 methods using the homogenized data, with those involving estimated change-point hinges being notable exceptions. For the homogenized station data, CPC$_{15}$ is again best
overall among several methods with RV < 1 for these nationally and annually aggregated results. Neither OCN nor the two estimated-change-point hinge methods are viable alternatives to WMO30 for any of the datasets considered here and therefore will not be analyzed further.

If one considers the full homogenized HCN dataset (dashed bars in Fig. 2), the RVs are universally smaller but are very similar to those for only matched samples of homogenized data.

2) BIAS

Biases for WMO30 and its 11 candidate alternatives used for hindcasts of 2006–12 seasonal temperatures are presented in Fig. 3 for the same four datasets as in Fig. 2. The biases are generally similar across the four datasets for each of the 12 normals formulations, relative to the 90% bootstrap confidence intervals. The exceptionally warm 2011/12 winter period (‘x’ in Fig. 1a) in the independent test sample contributes to the cold biases exhibited by many of the methods. The coolest biases are exhibited by WMO30 and linear trends, both of which are very slow in responding to the gradual warming over the past 30 or so years. CPC10 shows nearly zero bias for these annual and national aggregations, and the 1975 hinge has the second smallest bias. All hinge variants exhibit warm biases, which result mainly from autumn and winter temperatures in much of the 2006–12 period being near or below the post-1975 trends, as indicated in Fig. 1a, notwithstanding the warm 2011/12 and 2012/13 winters.

b. Best normals by region and season for homogenized USHCN temperatures

In section 3a the overall conclusions of W13 were confirmed for nonhomogenized station temperatures. Therefore, we will only examine seasonal and
geographic variations in hindcast performance for the full homogenized USHCN stations over 2006–12. In addition, the presently uncompetitive methods, OCN and the estimated-change-point hinges, are not included in the discussion below for simplicity.

1) REDUCTION OF VARIANCE

RVs for eight alternative normals used for hindcasts of 2006–12 seasonal temperatures are arranged season by season in the rows of Fig. 4 for three regions: west
Stars denote the smallest RV for each of the 12 region–seasons. No method outperformed WMO30 in winter in the west, and only one (CPC15) did so in the central region. The west region had the most severe cold seasons over the test period, followed by the central region. The 1975 hinges only performed well in the west in the summer. In the central and east regions, CPC15 was dominantly the best performer in both autumn and winter, and the 1975 hinges were dominantly best in spring and summer. In all three instances in which the two-phase 1975 hinge has the smallest RV, the 1975 hinge RVs are comparable and the next smallest. Winter in the central and west regions and secondarily the west in spring (3 of 12 region–seasons) account for virtually all of the overall advantage of CPC15 over the 1975 hinge. This in turn is largely accounted for by unusually cold seasons (in the context of the last several decades; Fig. 1a) during 2006–12.

2) BIAS

Biases for WMO30 and eight alternatives used for hindcasts of 2006–12 seasonal temperatures are arranged in Fig. 5 similarly to Fig. 4. Stars denote the smallest absolute bias for each of the 12 region–seasons. The 1975 hinges have the smallest absolute biases among the alternatives shown in 6 of 12 region–seasons and overall (Fig. 3) for this dataset, including all seasons in the east.

Biases are dominantly cold in the east, mostly cold just for spring and summer in the central region, and cold only (with a few minor exceptions) for summer in the
west, consistent with the pattern of anomalously cold seasons across the United States during 2006–12. Winter seasons were so unusual relative to recent decades in the west that they largely negated previous warming, leading to WMO30 having the smallest absolute bias in both winter and spring. Reinforcing this point, the confidence intervals for WMO30 and the 1975 hinges do not overlap for west–winter.

c. Hybrid 1975-hinge–CPC15 performance

L07’s recommendation to use a “hybrid” approach for tracking normals, which extrapolates a hinge function when the warming trend is large and an OCN variant when it is not, is retested here using CPC15 and W13’s \( t^2 \) cutoff criterion. The \( t^2 \) is the square of the conventional \( t \) statistic characterizing the strength of the post-1975 hinge slope and is related to the standardized slope \( \beta \) [Eq. (2)] according to

\[
\frac{t^2}{(t^2 + 1)(t^2 + 2)}
\]

where \( \tau \) is the number of training-data years from 1975 onward. On the basis of positive results for the 1975 hinge in section 3b, the approach is applied here to the full homogenized USHCN temperatures during the 2006–12 period that was not available to L07. The 1975 hinge is used to make the hindcast if \( t^2 \) for its slope exceeds a specified threshold, and CPC15 is used otherwise. All-case RMSEs and biases for 2006–12 are plotted for a range of thresholds in Fig. 6 along with the RMSE results reproduced from Fig. 8 of W13.

Our expectation from section 3b was that the hybrid method would perform better than in W13, where no \( t^2 \) threshold could be identified that did not degrade the results using only CPC15. This is reflected in the gray line in Fig. 6 (reproducing W13 results) curving monotonically downward to the right to the \( t^2 = 20 \) that excludes any use of the hinge. A relative minimum in this RMSE (i.e., some value of \( t^2 \) that yields an overall RMSE less than those for either the 1975 hinge or CPC15 alone) would be diagnostic for successful application of the hybrid concept.

A minimum does appear in the RMSE function in Fig. 6 (solid line) for the hybrid method applied to homogenized station data, but only a slight one, and for use of the hinge in only a tiny portion of total cases. There are two other differences between these results and those in W13 that are encouraging, however. First, the slope is clearly less for the solid curve (for homogenized station data) than for the gray curve (for nonhomogenized megadivision data), and the former is essentially flat (around 0.2% change) for \( t^2 \geq 10 \). Second, a larger proportion of cases use the 1975 hinge instead of CPC15 at each threshold than in W13 (e.g., 23% vs 18% for \( t^2 = 3 \)); that is, the \( t^2 \) values are larger overall for the homogenized USHCN data than for the divisional data used in W13. Thus, there is almost no degradation in overall RMSE (and therefore also RV) performance for a \( t^2 = 10 \) threshold (hinge replacing CPC15 in 3% of all cases) and less than 1% for \( t^2 = 5 \) (12% of all cases). On the other hand, using the 1975 hinge in the 12% of cases exhibiting \( t^2 > 5 \) decreases the overall cold bias by about 30%.

Thus, implementation of the hybrid approach may be viable at present when using the homogenized USHCN, by users for whom a very small penalty in squared error or RV is worth the corresponding reduction in cold bias. The results of section 3b suggest that this hybrid method may well be viable more generally if a decision maker’s interest is either in the east or central regions, since the top-performing methods in all four seasons in the two regions are one or the other of the two hybrid components (Fig. 4).

4. Conclusions and recommendations

Our objective has been to extend the work of W13 to address the impact of the data used (U.S. megadivision
surface temperatures) on the conclusions about alternatives to WMO\textsubscript{30} for tracking changing climate means. To meet this objective, W13’s analysis was repeated on both nonhomogenized and homogenized station data (USHCN) for the challenging independent test period 2006–12. From these tests of 11 (2 more than W13) alternative normals, four implications of W13 remain unchanged and, in fact, are reinforced here.

First, the time to seriously consider operational alternatives to WMO\textsubscript{30}, whether updated annually or not, for U.S. surface mean temperatures is here, if not overdue. The results presented here also indicate that alternative normals should be investigated for quantities such as monthly-mean and seasonal-mean maximum and minimum temperatures. We recommend that NCDC should continue to move in this direction (Arguez and Vose 2011; Arguez et al. 2013). Also, we recommend that the WMO Commission on Climatology should form an expert team to consider formulation of new recommendations for official normals, for which there is adequate time before the end of the next WMO-required 30-yr period (1991–2020). Second, L07’s model choices for hingelike representations of recent climate changes, specifically a stationary climate from 1940 up to a change point fixed at 1975, are revalidated. Third, CPC\textsubscript{15} is overwhelmingly the best alternative normal to WMO\textsubscript{30} for nonhomogenized station and megadivision surface temperature data and is the best overall for homogenized station (i.e., USHCN) data. In particular, CPC\textsubscript{15} outperformed CPC\textsubscript{10} in predicting U.S. surface temperatures during 2006–12, as was also found by W13 for 1994–2011 and by Wilks (1996) for the 1961–93 period. Fourth, OCN\textsubscript{M} and both estimated change-point hinges are not viable alternatives to WMO\textsubscript{30} at this time for any of the surface temperature datasets studied. Of course, the specific results depend on the particular verification period used, and 2006–12 has been chosen here to provide test data that are fully independent of the analysis and conclusions of L07. It should be worthwhile to revisit periodically these comparisons as ongoing climate changes unfold and more data accumulate.

Two outcomes of this analysis using homogenized station data were notably different from those in W13. First, the 1975 hinges collectively emerged as the second most accurate overall alternatives (after CPC\textsubscript{15}) to WMO\textsubscript{30}, exhibiting RVs smaller than CPC\textsubscript{15} for summer in all three regions and for spring in the east and central regions. Also, the 1975 hinges were better alternatives than CPC\textsubscript{15} with respect to bias in 8 of 12 region–seasons. Further, in the cases in which the 1975 hinges perform well, an additional advantage of their use is their stability, since the 1975 hinges will change less than the viable alternatives when updated annually.

Second, the hybrid approach (use of the 1975 hinge instead of CPC\textsubscript{15} when the post-1975 slopes exceed a defined $r^2$ threshold) can now be considered to be a potentially advantageous alternative if a user’s general interest spans any time of year in the east or central regions. The hybrid method may also be viable if a user is willing to pay a small penalty in RV to gain a reduction in bias. We know of two applications in which this has been the case: removing the nonstationary climate-mean signal from the record to quantify impacts of modes of climate variability for risk estimates (e.g., CPC and El Niño/La Niña teleconnections; L07),\textsuperscript{1} and specifying winter normals to estimate natural gas demand for heating for rate-setting purposes (Arguez et al. 2013). In the second case, the large cold biases of such methods as WMO\textsubscript{30} and linear trends produce errors that favor customers because expected demand will be erroneously high, leading regulatory bodies to set prices that are too low. The opposite is true in the case of large warm biases, for which producers are favored.\textsuperscript{2}

The short independent test periods [2006–11 or 2006–12 (used here)] were particularly challenging for trend-based methods, and especially for the hinge methods, because of their exceptional year-to-year stability. In particular, the cold half of the year in the United States was comparatively cold relative to recent decades during this period, and hinge functions will be notably slow to react to such anomalies. For example, in Fig. 1a, 4 of the 6 December–April periods rank among the coldest half for the post-1975 period. This was far from the case for the June–October period (Fig. 1b). All regions of the United States experienced relatively cold winters during the test period, but especially the western two-thirds. Thus, because warming is expected to resume (it may already have, based on December–April 2011–12; note the x in Fig. 1a), comparable performance of the hinge or its variants to the CPC\textsubscript{15} method during this most recent test period should be considered promising. Note, however, that for W13’s megadivision data assessment (see his Fig. 6 and note the error bars) CPC\textsubscript{15} is unambiguously a better performer than the hinge overall.

Of course, seasonally averaged temperature is not the only climate statistic that has undergone progressive change in recent decades. Adaptive adjustment of climate normals for other aspects of the changing climate is

\textsuperscript{1}Using, for example, CPC\textsubscript{15} and offsetting it to the midpoint of the window is an unsatisfactory solution, because fictitious data must be generated to extend removal of the nonstationary climate-mean signal to the beginning and end of the record.

\textsuperscript{2}The second author has participated as an expert witness in eight natural gas rate-setting cases before state public service or utility commissions.
an equally valid exercise and should also yield improved specification and projection of the evolving climate. For example, Krakauer (2012) considers increases in the extreme-value statistic of annual minimum temperature, modeling its evolution over the past century using a nonparametric approach. The result (his Fig. 4) is qualitatively similar to the 1975 hinge, exhibiting quasi stationarity before about 1975 and a nearly linear upward trend since then.

Our expectation was that homogenization would generally improve the performance of most alternative normals, but especially the two 1975 hinge formulations, because homogenization eliminates many of the nonclimatological features in records that may mask warming signals, which a vast body of corroborating evidence suggests should be prominently present (e.g., Solomon et al. 2007; L07). The results here confirmed these expectations. Thus, we agree with NCDC’s practice and recommendation that climate normals should be produced from homogenized records whenever possible. We consequently urge NCDC to make homogenized divisional data publicly available as soon as possible.

Acknowledgments. We thank the anonymous reviewers for constructive comments that lead to improvements in this paper. This research was supported by the National Science Foundation under Grant AGS-1112200.

REFERENCES
